Probabilistic inference of the genetic architecture of functional enrichment of complex traits

Marion Patxot[1,†], Daniel Trejo Banos[1,†], Athanasios Kousathanas[1,†], Etienne J. Orliac[2], Sven E. Ojavee[1], Gerhard Moser[3], Julia Sidorenko[4], Zoltan Kutalik[5,6], Reedik Mägi[7], Peter M. Visscher[4], Lars Rönnegård[8,9], Matthew R. Robinson[10,*]

1 Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.
2 Scientific Computing and Research Support Unit, University of Lausanne, Lausanne, Switzerland.
3 Australian Agricultural Company Limited, Brisbane, QLD, Australia
4 Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia.
5 University Center for Primary Care and Public Health, Lausanne, Switzerland
6 Swiss Institute of Bioinformatics, Lausanne, Switzerland
7 Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
8 School of Technology and Business Studies, Dalarna University, Falun, Sweden
9 Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden
10 Institute of Science and Technology Austria, Klosterneuburg, Austria.
† These authors contributed equally to this work.
*corresponding author: matthew.robinson@ist.ac.at

Due to the complexity of linkage disequilibrium (LD) and gene regulation, understanding the genetic basis of common complex traits remains a major challenge. We develop a Bayesian model (BayesRR-RC) implemented in a hybrid-parallel algorithm that scales to whole-genome sequence data on many hundreds of thousands of individuals, taking 22 seconds per iteration to estimate the inclusion probabilities and effect sizes of 8.4 million markers and 78 SNP-heritability parameters in the UK Biobank. Unlike naive penalized regression or mixed-linear model approaches, BayesRR-RC accurately estimates annotation-specific genetic architecture, determines the underlying joint effect size distribution and provides a probabilistic determination of association within marker groups in a single step. Of the genetic variation captured for height, body mass index, cardiovascular disease, and type-2 diabetes in the UK Biobank, only $\leq$ 10% is attributable to proximal regulatory regions within 10kb upstream of genes, while 12-25% is attributed to coding regions, up to 40% to intronic regions, and 22-28% to distal 10-500kb upstream regions. $\geq$60% of the variance contributed by these exonic, intronic and distal 10-500kb regions is underlain by many thousands of common variants, each with larger average effect sizes compared to the rest of the genome. We also find differences in the relationship between effect size and heterozygosity across annotation groups and across traits. Up to 24% of all cis and coding regions of each chromosome are associated with each trait, with over 3,100 independent exonic and intronic regions and over 5,400 independent regulatory regions having $\geq$95% probability of contributing $\geq$0.001% to the genetic variance for just these four traits. In the Estonian Biobank, we show improved prediction accuracy over other approaches and generate a posterior predictive distribution for each individual.

As whole-genomes are collected for hundreds of thousands of individuals, we require regression methods that are not only computationally efficient, but which also provide improved inference. Methods should fully utilize the data, rather than relying on subsets of the SNPs, exploiting computational power to facilitate discovery of additional genomic regions, improve understanding the genomic architecture of common disease, and provide more informative genomic prediction.

Recent studies [1–4] highlight the importance of accounting for minor allele frequency (MAF) and LD structure of the genomic data when estimating the proportion of phenotypic variance attributable to different categories of genetic markers (the SNP-heritability, h^2_{SNP}). Assessment of the relative contribution of different genomic regions is currently made assuming that markers within a category all contribute to the variance, with enrichment of the category defined as the estimated share of the variance explained divided by its expected share [5,6]. However ideally, the estimated distribution of marker effects for each category would be directly obtained, accounting for MAF and LD structure and allowing for some of the marker effects to be

zero, to improve our understanding of the genetic architecture underlying the relative contributions of different genomic regions. Furthermore, if approaches could enable a probabilistic understanding of the number of associated genes and genomic regions, than it would yield a better understanding of the polygenicity of genomic effects.

Current mixed-linear association models such as those implemented in the software fastGWA [7], boltLMM [8], and REGENIE [9], use a two-step approach, first estimating the variance contributed by the SNP markers generally without the use MAF-LD-annotation information, and then using the point estimates when estimating the marker effect sizes in a second step, essentially assuming effects in the model come from a single distribution [7, 8, 10]. Summary statistic approaches such as LDSC [11] and SumHer [6], then use these baseline estimated effects coupled with independent LD reference panels to then alter the weightings of the marker effects allowing for annotation differences, showing improved genomic prediction. However, currently no model directly provides joint estimates of the marker effects, testing for association accounting for effect size differences across MAF, LD, or annotation groups.

Here, we outline the fastest Bayesian penalised regression model to date, with a hybrid-parallel algorithm for analysing large-scale genomic data that: (i) provides unbiased MAF-LD annotation-specific genetic effect size estimates and h^2_{SNP} of different annotations in a single step, allowing for a contrasting of the genetic architectures of complex traits under a flexible prior formulation, (ii) yields the probability that each marker, genomic region, annotation, or gene-coding region, is associated with a phenotype, alongside the proportion of phenotypic variation contributed by each describing the *gene* architecture of complex traits, (iii) conducts fine-mapping automatically, and (iv) gives a posterior predictive distribution for each individual at each genomic region.

A Bayesian model for large-scale genomic data

The model we derive is based on grouped effects with mixture priors, improving on the formulations of [12, 13] and [14]. Like these former methods, we consider a spike probability at zero (Dirac delta function), and a scale mixture of Gaussian distributions as a slab probability density; unlike these models, we have genetic markers grouped into MAF-LD-annotation specific sets, with independent hyper-parameters for the phenotypic variance attributable to each group. Assuming N individuals and p genetic markers, our model of an observed phenotype vector $\mathbf{y}$ is:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{\varphi=1}^{\Phi} \mathbf{X}_\varphi \boldsymbol{\beta}_\varphi + \boldsymbol{\epsilon} \tag{1}$$

where there is a single intercept term $\mathbf{1}\mu$ and a single error term, a vector $(N \times 1)$ of residuals $\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}|\sigma^2_\epsilon \sim \mathcal{N}\left(0, \mathbf{I}\sigma^2_\epsilon\right)$. An N by p matrix of single nucleotide polymorphism (SNP) genetic markers, centered and scaled to unit variance, which we denote as $\mathbf{X}_\varphi$. The effects are allocated into groups $(1, \ldots, \Phi)$. Each group has a set of model parameters $\Theta_\varphi = \left\{\boldsymbol{\beta}_\varphi, \pi_\varphi, \sigma^2_{G\varphi}\right\}$, with $\boldsymbol{\beta}_\varphi$ as a $p_\varphi \times 1$ vector of partial regression coefficients, where β_{φ_j} is the effect of a 1 SD change in the j^{th} covariate within the φ^{th} group. The spike and slab prior, contains what is called a Dirac spike [15, 16] for $\boldsymbol{\beta}_\varphi$, which induces sparsity in the model through a Dirac-delta at zero, excluding variables from the model by setting their coefficients to zero. A finite scale mixture of normal distributions centered at zero constitute the slab component. The slab shrinks the non-zero coefficients towards zero according to the slab's width, by having a scale mixture of Gaussians, the distribution has heavier tails and can accommodate big and small effects [17]. Therefore, each β_{φ_j} is distributed according to:

$$\beta_{\varphi_j} \sim \pi_{0\varphi}\delta_0 + \pi_{1\varphi}\mathcal{N}\left(0, \sigma^2_{1\varphi}\right) + \pi_{2\varphi}\mathcal{N}\left(0, \sigma^2_{2\varphi}\right) + \ldots + \pi_{L_\varphi\varphi}\mathcal{N}\left(0, \sigma^2_{L_\varphi\varphi}\right) \tag{2}$$

where for each SNP marker group $\left\{\pi_{0\varphi}, \pi_{1\varphi}, \ldots, \pi_{L_\varphi\varphi}\right\}$ are the mixture proportions and $\left\{\sigma^2_{1\varphi}, \sigma^2_{2\varphi}, \ldots, \sigma^2_{L_\varphi\varphi}\right\}$ are the mixture-specific variances proportional to

$$\begin{bmatrix} \sigma^2_{1\varphi} \\ \vdots \\ \sigma^2_{L_\varphi\varphi} \end{bmatrix} = \sigma^2_{G\varphi} \begin{bmatrix} C_{1\varphi} \\ \vdots \\ C_{L_\varphi\varphi} \end{bmatrix}$$

with $\sigma^2_{G\varphi}$ the phenotypic variance associated with the SNPs in group φ, which, like all the other parameters, is estimated directly from the data. Thus, related approaches of BayesRC and BayesRS that are heavily

utilized in animal and plant breeding [18, 19] are extended as the mixture proportions, the variance explained by the SNP markers, and the mixture constants are all unique and independent across SNP marker groups. This enables estimation of the amount of phenotypic variance attributable to the group-specific effects, and differences in the underlying distribution of the $\boldsymbol{\beta}_{\varphi}$ effects among MAF-LD-annotation groups, with different degrees of sparsity.

Inference from our grouped effects mixture priors model with LD

While comparisons of different approaches have been made under different simulation scenarios, we have limited understanding of why approaches differ. We show in theory (see Methods Eq.23) and in simulation (see Methods, Figure 1, Figure S1) the importance of our model formulation for accurate estimation of h^2_{SNP} and the SNP regression coefficients. We find that when highly correlated common variants (under multicollinearity, Figure 1, Figure S1, and Methods Eq.23) contribute more to the phenotypic variance than low-LD markers, penalized regression or mixed linear model approaches will inaccurately estimate their effects. This occurs due to the assumption made by these models that effects come from a single Gaussian distribution, and thus that there is a single regularisation parameter appropriate to all markers. To demonstrate this in simulation study, we used real genomic data where 50 replicate phenotypes were generated by either (i) allocating 5000 LD independent causal variants to high LD SNPs (Figure 1a y-axis panel: high LD), or (ii) randomly allocating SNPs as causal variants (Figure 1a y-axis panel: random). Within (i) and (ii) we then either randomly allocated effect sizes to those SNPs (Figure 1a x-axis: random), or allocated effect sizes proportional to their LD and inversely proportional to the MAF (Figure 1a x-axis: MAF-LD, see Methods). In these simulation settings, overestimation of the SNP heritability occurs for mixed-linear model association methods (MLMA) [7] and Bayesian dirac spike and slab models with a single global hyperparamter (BayesR), when high-LD SNPs are allocated as causal variants, replicating previous results [1–4].

Our theory in Eq.23 gives the expectation that this overestimation should occur specifically at common, high LD variants, and we show this empirically using the scenario where causal variants are allocated to high-LD SNPs. While the 5000 causal variants are LD-independent, they are each correlated with other SNPs of simulated effect size 0. So, for each of the 5000 independent high-LD causal variants, we calculated the sum of the squared estimated regression coefficients for the causal variant and all markers in LD ≥ 0.05. From this, we subtract the true simulated value, which is simply the square of the effect size allocated to the causal variant. We then divided by the SD of the simulated genetic effects, to give a z-score for each causal variant, plotted in Figure 1b. MLMA and BayesR overestimate the effects of variants that are in LD with a high-LD causal variant (Figure 1b), and with MLMA models this overestimation is severe. Both h^2_{SNP} and SNP marker regression coefficient estimation accuracy improves when using MAF-LD specific shrinkage (Figure 1a BayesRR), because estimated common variant effects in high LD are shrunk to a greater degree, alleviating the influence of multicollinearity (Figure 1, Figure S1).

We then present a further empirical example, where 50 pairs of SNP markers with LD = 0.9 were simulated for each of 50 simulation replicates, where only one marker of each pair has an effect (0,0.1 SD), giving the sum of the squared regression coefficients as 0.5 for each simulation (Figure 1c: dotted red line). In order to compare formulations of different statistical approaches, we define lambda as the shrinkage parameter, which is the ratio of the residual (error) variance and the variance attributable to the SNP markers. This hyperparameter is used for MLMA, ridge regression (Ridge) [21] and the BayesR model in the estimation of the effects (see Methods). We show that under multicollinearity (Figure 1c: collinear), unless lambda is large, meaning that the shrinkage of marker effect sizes is large, SNP marker effects are consistently overestimated. This is of fundamental importance to accurate estimation of SNP marker effect sizes in either penalized regression or mixed linear association models (Figure 1, Figure S1). We show that when specifying enrichment using prior knowledge (Figure 1d: multiple group enrichment), the genetic architecture is accurately inferred by BayesRR-RC. In comparison with other recent approaches providing annotation-specific variance component estimates in individual-level data, BayesRR-RC performs as boltREML [22] and RHEmc [4] when estimating the overall variance explained by each annotation group, with RHEmc estimates showing higher variability (Figure 1d). We then additionally show how BayesRR-RC takes this a step further to integrate prior biological knowledge to improve power to infer the genetic architecture of complex traits, both in simulation (Figure S2) and empirically in our UK Biobank analysis described below.

A hybrid parallel Gibbs sampling scheme for large-scale genomic data

We then overcome a major-hurdle limiting the application of penalized regression approaches to large-scale biobank data, by deriving a Bulk Synchronous hybrid-parallel (BSP) sampling scheme for Eq.(1) that allows

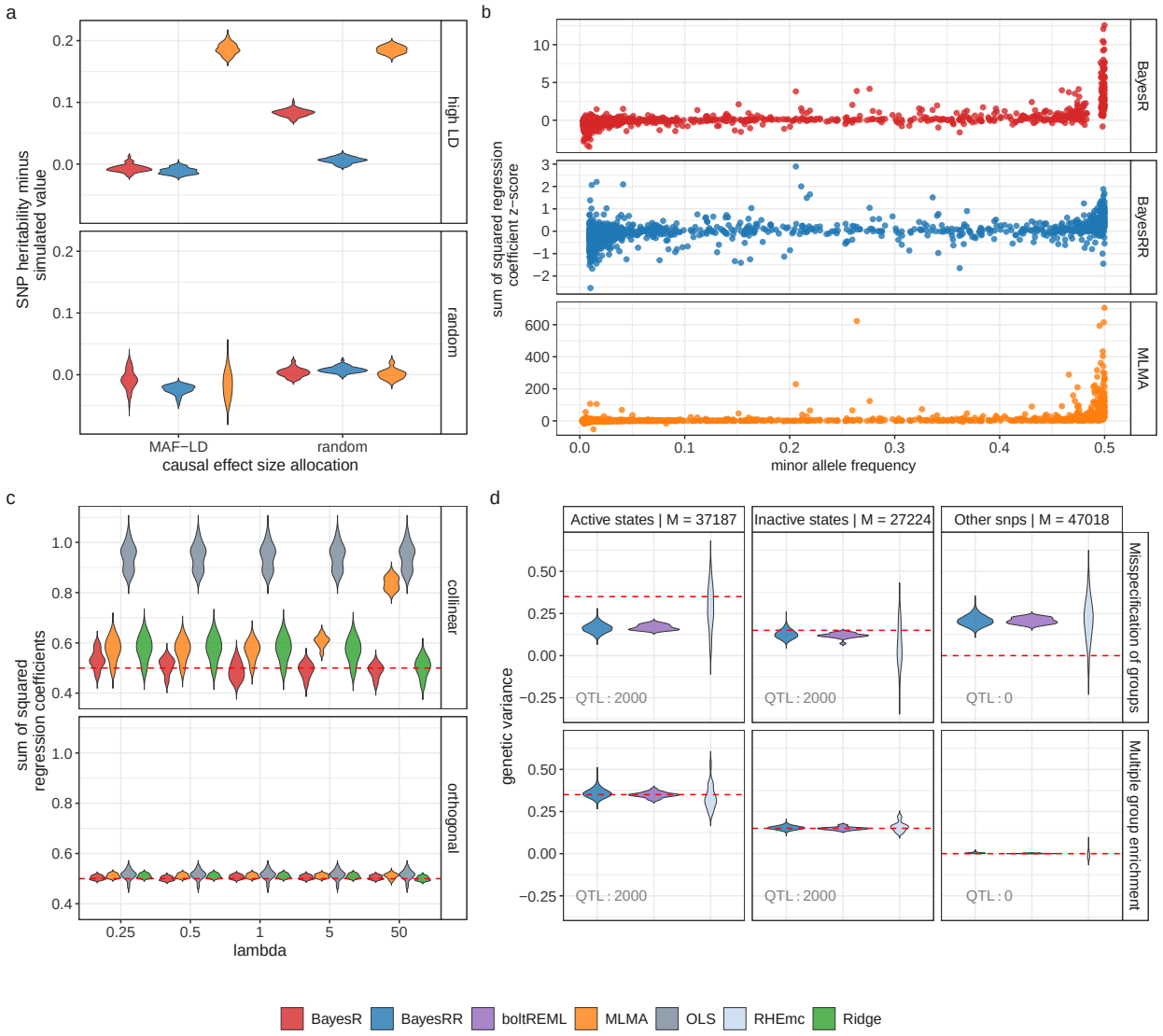

Figure 1. Theory and simulation study for genetic penalized regression models under multicollinearity. (a) Simulation study using real genomic data where 50 replicate phenotypes were generated by either allocating 5000 LD-independent causal variants to high LD SNPs (y-axis panel: high LD), or randomly allocating 5000 SNPs as causal variants (y-axis panel: random), and then either randomly allocating effect sizes to those SNPs (x-axis: random), or allocating effect sizes proportional to their LD and MAF (x-axis: MAF-LD, see Methods). SNP heritability estimation error is plotted as the difference of the estimate and the true simulated value across the 50 replicates. (b) We then investigated this further for the scenario where causal variants are allocated to high-LD SNPs. While the 5000 causal variants are LD-independent, they are each correlated with a large number of SNPs of simulated effect size 0. For each causal variant, we took all the markers in LD ≥ 0.05 and summed the squared estimated regression coefficients of these markers. The true simulated value is simply the square of the effect size allocated to the causal variant, and we subtracted this from the sum of the squared regression coefficients divided by the SD of the simulated genetic effects, to give a z-score for each causal variant and this is plotted on the y-axis for MLMA, BayesRR, and BayesRR. (c) Our theory outlines how this overestimation is the result of the effect of multicollinearity (see Methods) and an example is shown here, where 50 pairs of SNP markers with LD = 0.9 were simulated for each of 50 simulation replicates, where only one marker of each pair has an effect (0,0.1 SD), giving the sum of the squared regression coefficients as 0.5 for each simulation (dotted red line). lambda is the shrinkage parameter, the ratio of the error variance and the variance attributable to the SNP markers, used for MLMA, ridge regression (Ridge) and the BayesR model to estimate the effects. (d) Simulation of a genetic architecture (dotted red line) using real annotations from the Epigenome Roadmap Project [20] (active states, inactive states, other snps). We compared BayesRR to other recent approaches providing annotation-specific variance component estimates in individual-level data when SNPs are randomly assigned to an annotation (labelled: misspecification of groups), or when specifying enrichment using prior knowledge (labelled: multiple group enrichment)

both the data and the compute tasks to be split within and across compute nodes in a series of message-passing interface (MPI) tasks. This BSP Gibbs sampling scheme, implemented based on a hybrid MPI + OpenMP model with residual updating and message interfacing, allows the MCMC Gibbs sampling simulations to retain accuracy of the estimation of the partial regression coefficients of each SNP marker $\boldsymbol{\beta}_{\varphi}$ (the joint effect

of each marker, conditional on all other markers), whilst allowing the marker effects to be updated in parallel (see Methods and simulation study of Figure S3).

Our Gibbs sampling algorithm enables all sampling steps to utilize genetic data stored in mixed binary/sparse-index representation, reducing computational complexity of a single Gibbs step from $\mathcal{O}(n)$ to $\mathcal{O}(n_z)$, with n_z the number of non-zero genotypes. This provides a highly vectorizable mixed representation of genomic marker data as a series of indices (Figure S4) and this facilitates highly vectorized and highly parallel SNP-phenotype covariance estimation (dot product calculation) in a series of look-up tables which greatly extends previous sparse residual updating schemes.

We provide publicly available open source software (see Code Availability) with capacity to easily extend to a wider range of models than that demonstrated here (see Methods). Our software requires as little as 22 seconds per MCMC sample to estimate 78 group-specific h^2_{SNP} parameters, and the inclusion probabilities and effect sizes of 8,433,421 markers in 382,466 individuals on standard Intel Xeon CPU processors (Figure S4, see Code Availability for hardware specifications).

The genetic architecture of enrichment in the UK Biobank

This sampling scheme enabled us to apply the model to cardiovascular disease outcomes (CAD), type-2 diabetes (T2D), body mass index (BMI) and height measured for 382,466 unrelated individuals from the UK Biobank data genotyped at 8,433,421 imputed SNP markers. These markers were selected as they overlap with the Estonian Genome Centre data (see Methods) and have minor allele frequency >0.0002. Although the model can account for relatedness and data structure automatically [14, 23] (Figure S5), we wished to contrast the genetic architecture of different phenotypes and estimate the phenotypic variance contributed by MAF-LD-annotation groups from markers that enter the model only due to LD with underlying causal variants (as closely as we can in a correlational study). Thus, we also adjust each phenotype for age, sex, year of birth, genotype batch effects, UK Biobank assessment centre, and the leading 20 principal components of the SNP data. We provide evidence through theory and in simulation that by better correcting for multicolinearity, the BayesRR-RC model also better controls for underlying populations structure as compared to a mixed-linear association model with the leading PCs of the genomic data included (Figure S5).

We partition SNP markers into 7 location annotations preferentially assigning SNPs to coding (exonic) regions first, then in the remaining SNPs we preferentially assigned them to intronic regions, then to 1kb upstream regions, then to 1-10kb regions, then to 10-500kb regions, then to 500-1Mb regions. Remaining SNPs were grouped in a category labelled "others" and also included in the model so that variance is partitioned relative to these also. Thus, we assigned SNPs to their closest upstream region, for example if a SNP is 1kb upstream of gene X, but also 10-500kb upstream of gene Y and 5kb downstream for gene Z, then it was assigned to be a 1kb region SNP. This means that SNPs 10-500kb and 500kb-1Mb upstream are distal from any known nearby genes. We further partition upstream regions to experimentally validated promoters, transcription factor binding sites (tfbs) and enhancers (enh) using the HACER, snp2tfbs databases (see Code Availability). All SNP markers assigned to 1kb regions map to promoters; 1-10kb SNPs, 10-500kb SNPs, 500kb-1Mb SNPs are then split into enh, tfbs and others (unmapped SNPs) extending the model to 13 annotation groups. Within each of these annotations, we have three minor allele frequency groups (MAF<0.01, 0.01>MAF>0.05, and MAF>0.05), and then each MAF group is further split into 2 based on median LD score. This gives 78 non-overlapping groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1). We conducted a series of convergence diagnostic analyses of the posterior distributions to ensure we obtained estimates from a converged set of four Gibbs chains, each run for 6,000 iterations with a thin of 5 for each trait (Figure S6, S7, S8, S9).

We find that 32-44% of the h^2_{SNP} is attributable to intronic regions, 12-25% is attributable to exonic regions, 22-28% is attributable to markers 10-500kb upstream of genes, with proximal (within 10kb) promotors, enhancers and transcription factor binding sites cumulatively contributing <10% (Figure 2b, Figure S10, with estimates summed across MAF and LD groups Table 1, and full results in Table S1). The large contribution of exonic and intronic annotations to variation is in-line with the fact that these annotations account for $\sim 40\%$ of the total genome length. All four traits show the same pattern of group-specific variation, with the exception of height, where the proportion of h^2_{SNP} attributable to exons is almost twice as large as the other phenotypes (Figure 2b, Figure S10, Table 1, and Table S1). For all annotation groups in exons, introns, and within 500kb of genes across all traits, $\geq 60\%$ of the h^2_{SNP} attributable to these groups is contributed by many thousands of common variants, each of small effect (Figure 2b, Figures S10 and S11. We find

differences in the underlying effect size distribution across annotation groups. For each group, we modelled the SNP effects as coming from a series of five Gaussian mixtures, and we find that at least 45% of the h^2_{SNP} attributable to both introns and 500kb upstream regions is underlain by many thousands of SNPs that on average each contribute 0.001% (estimates summed across MAF and LD groups in Figure 2b, Figures S10 and S11). In contrast, the variance is spread more evenly across the mixtures for the other groups, implying that 10-500kb upstream regions and introns are more polygenic than other groups. This is especially so for BMI where 35% of the h^2_{SNP} is attributable to many thousands of intronic variants (Figure 2 and Figure S10). Therefore, the polygenicity of the genetic effects varies across different genomic regions, with remarkably consistent patterns across traits in the partitioning of h^2_{SNP} across the genome.

Figure 2. Genetic architecture of enrichment for height (HT), body mass index (BMI), cardiovascular disease (CAD) and type-2 diabetes (T2D) for 382,466 unrelated European ancestry UK Biobank individuals genotyped at 8,430,446 SNP markers.(a) We partition SNP markers into 7 location annotations (coding regions, intronic regions, and windows 1kb, 1-10kb, 10-500kb and 500kb-1Mb upstream of genes, with other SNPs grouped in a category labelled "others"). Windows 1-10kb, 10-500kb and 500kb-1Mb upstream of genes are further split into SNPs mapped to enhancers (enh), transcription factor binding sites (tfbs) and others. Within each of the 13 annotations, we have three minor allele frequency groups (MAF<0.01, 0.01>MAF>0.05, and MAF>0.05), and then each MAF group is further split into 2 based on median LD score. This gives 78 groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1). (b) Posterior distribution of the proportion of the total phenotypic variance attributable to the SNP markers that is contributed by each of the four non-zero mixtures within each MAF-annotation group for HT, BMI, CAD and T2D. Within these, are boxplots of the posterior mean and 95% credible intervals. Values are summed over LD groups. (c) Bar plots with error bars giving the 95% credible intervals for the average effect size of markers in the model for each MAF-annotation group, split by mixture.

group	trait	mean %	q(0.025) %	q(0.975) %
variance attributable to SNP markers genome-wide	HT	57.66	56.09	59.14
	BMI	28.74	27.62	30.00
	CAD	5.94	5.30	6.67
	T2D	8.45	7.83	9.18
proportion of genetic variance attributable to exonic regions of genes	HT	24.75	23.39	26.071
	BMI	12.98	10.98	14.84
	CAD	13.23	8.40	18.84
	T2D	14.49	10.74	18.54
proportion of genetic variance attributable to intronic regions of genes	HT	41.54	39.91	43.39
	BMI	44.17	41.36	47.25
	CAD	32.05	24.98	39.51
	T2D	37.28	32.22	42.57
proportion of genetic variance attributable to snps 500kb upstream of genes	HT	22.13	21.00	23.40
	BMI	28.58	26.41	31.01
	CAD	28.02	21.24	35.04
	T2D	27.42	22.68	32.36
proportion of genetic variance attributable to exonic regions that is explained by common variants	HT	72.09	69.77	74.14
	BMI	69.41	62.60	76.42
	CAD	64.97	43.08	83.16
	T2D	68.57	56.00	79.82
proportion of genetic variance attributable to intronic regions that is explained by common variants	HT	81.19	79.30	83.02
	BMI	85.05	78.28	91.49
	CAD	84.68	65.64	65.64
	T2D	65.64	65.64	65.64
proportion of genetic variance attributable to snps 500kb upstream of genes that is explained by common variants	HT	81.59	78.91	83.96
	BMI	86.78	80.56	91.60
	CAD	66.49	49.11	81.79
	T2D	72.35	58.71	83.75

Table 1. Proportion of genotypic variance genome-wide and predominantly explained by common SNPs located 10-500kb upstream of genes and coding regions for height (HT), body mass index (BMI), type-2 diabetes (T2D) and cardiovascular disease (CAD).

We then directly assessed the magnitude of the effect sizes within each group, calculating the average effect size of markers in the model, for each mixture, within each group, at each iteration of the model. Across traits, effect sizes scale to their differences in h^2_{SNP}, and we find that exonic and intronic region effect sizes were higher than the rest of the genome, across all mixture groups, followed by 10-500kb upstream regions (Figure 2c). We find little evidence that SNPs located in proximal promotors, enhancers, and transcription factor binding sites within 10kb of genes showed average effect sizes that were higher than SNPs located 1MB away from genes, or those that were not mapped to a specific category, with perhaps the exception of high MAF variants (Figure 2c). Generally, all phenotypes simply appear to be predominantly underlain by very many common variants, with SNPs within distal regulatory regions, coding and intronic regions each contributing more to the phenotypic variance and having higher allele substitution effects. As these results are for the effect sizes of standardized markers, it represents the square root of the average contribution of a marker to the total variance. Thus, we also re-scaled the marker effects by the standard deviation of each marker, to give effect sizes on the allele substitution effect size scale. Again, average effect sizes scaled to the h^2_{SNP} of the traits and we find that rare variants have higher average allele substitution effects than common variants for exonic, intronic, promotors and enhancers (Figure S11b). An exception to these patterns were BMI-associated intronic and 10-500kb group SNPs, where we find no evidence that the allele substitution effect size differs across frequency groups (Figure S11b). We also did not find evidence that the allele substitution effect size differed across frequency groups for transcription factor binding sites, distal SNPs 1MB upstream of genes, or those not mapping to an annotation group (Figure S11b). These results highlight that assuming an equal contribution of each marker within each annotation group may give misleading results when determining SNP enrichment. Evolutionary theory predicts that selection should result in higher effect sizes for rare variants and our results imply that selection pressures vary both across traits, but also across genomic regions with exons, promotors, and enhancers showing the strongest differentiation of effect sizes

across frequency groups as compared to the rest of the genome.

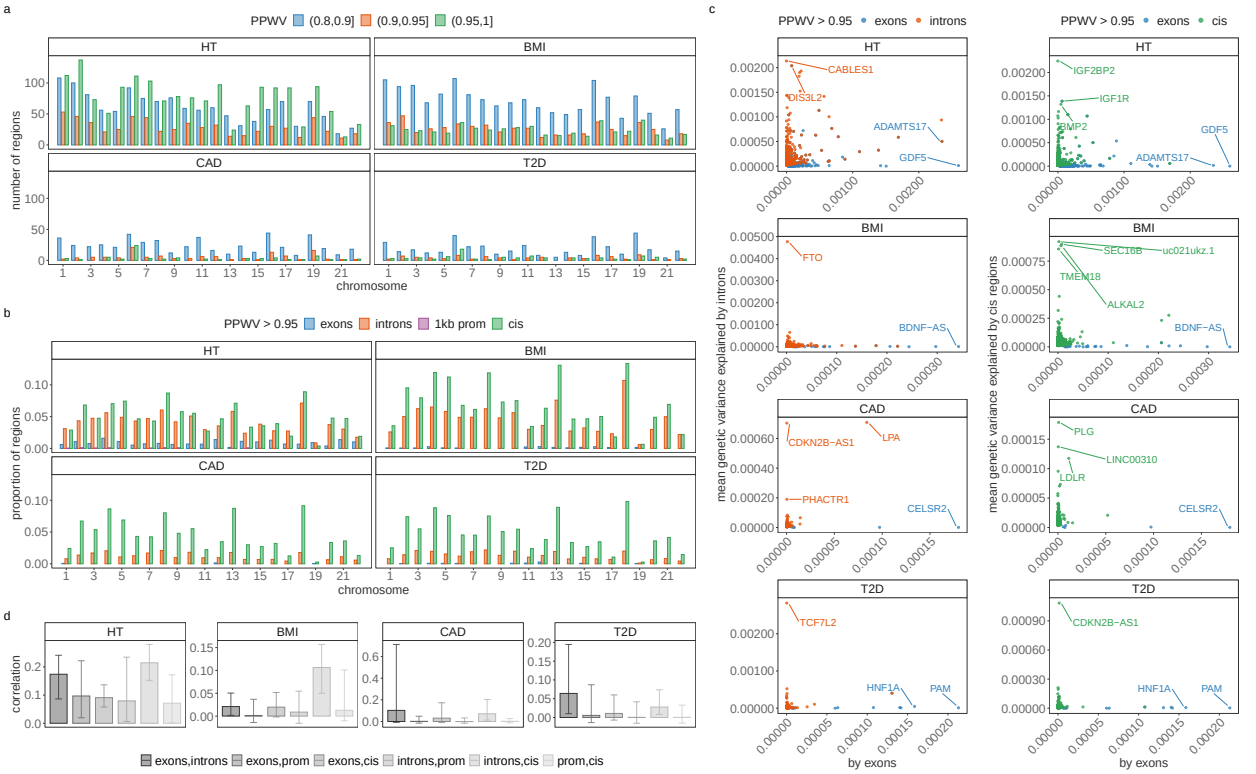

Figure 3. Contribution of genes and 50kb regions to height (HT), body-mass-index (BMI), cardiovascular disease (CAD) and type-2-diabetes (T2D). (a) We grouped SNPs in 50kb-regions genome-wide and estimated the sum of the squared regression coefficient estimates for each 50kb-region. We then select the number of 50kb regions that explain at least 0.001 percent of the variance attributed to all SNP markers in 80%, 90% and 95% of the iterations. This gives a measure called the posterior probability that the window variance (PPWV) [24] exceeds 1/10,000 of the phenotypic variation attributed to SNP markers. (b) We mapped SNPs to the closest gene +/- 50kb from the SNP position and labelled them as located in a coding region, an intron, 1kb upstream of a gene using our functional annotations (Figure **??**). Remaining snps are labelled as located in a cis-region (up to +/- 50kb from a gene). We then select the number of regions where PPWV is higher than 95% and explains at least 0.001 percent of the phenotypic variance attributed to all SNP markers. We then calculate the number of significant coding regions, introns, 1kb regions and cis regions as a proportion of the total number of genes for each chromosome. Genic associations that explain at least 0.001% of the phenotypic variance attributed to all SNP markers are again spread across chromosomes according to the chromosome length. (c) Shows the mean of the phenotypic variance attributed to intron and cis regions (y-axis) and coding regions (x-axis) that explain at least 0.001 % of the phenotypic variance attributable to SNP markers in $\geq$ 95% of the iterations (PPWV>0.95). These results provide joint estimates of the proportions of variance contributed by different gene bodies and automatic fine-mapping of gene bodies and their cis-regulatory regions. For example, introns and cis-regulatory regions of FTO respectively contribute 0.48% (95% CI 0.29, 1.12) and 0.01% (95% CI 0, 0.01) to the phenotypic variance of BMI.

The gene architecture of enrichment for common complex traits

We then partitioned the variance attributed to SNP markers across 50kb regions of the genome, then across SNPs annotated to genes, and then to SNPs themselves. We determined the posterior inclusion probability that each region and each gene contributes at least 0.001% to the h^2_{SNP}, providing a probabilistic approach to assess the contribution of different genomic regions to trait variation (termed PPWV, see Methods and simulation study of Figure S12). We first divide the genome into 50kb blocks and find 1660 50kb regions for height with $\geq$ 95% posterior probability of explaining 0.001% of the h^2_{SNP}, 520 regions for BMI, 70 regions for CAD and 87 regions for T2D (Figure 3a).

We then map SNPs to their closest gene (+/- 50kb from SNP position) and we use our annotations to label them (see Methods). We find 243 independent coding regions for height with $\geq$ 95% posterior probability of explaining at least 0.001% of the h^2_{SNP}, 29 independent coding regions for BMI, 5 for CAD and 13 for T2D. We find many more associations in the cis region of genes with 1254 independent cis-regions for height with $\geq$ 95% posterior probability of explaining 0.001% of the h^2_{SNP}, 1765 independent cis-regions for BMI, 1166 for CAD and 1221 for T2D. We additionally find 9 independent promoter regions and 1072

independent introns for height with $\geq 95\%$ posterior probability of explaining at least 0.001% of the h^2_{SNP}, 1162 independent intronic gene regions for BMI, 307 for CAD and 347 for T2D. With many thousands of SNP markers entering the model for each trait, summarising the posterior distribution obtained over gene annotations provides an understanding of the gene architecture of common complex traits.

When we calculate the number of exons, introns, promotors and cis regions with $\geq 95\%$ posterior probability of explaining 0.001% of the h^2_{SNP}, as a proportion of the total number within each chromosome, we find that up to 24% of the genes on each chromosome are associated with each of the four traits (Figure 3b). Generally, we find that only 1% or less of the available exons and promotor regions of genes per chromosome show an association with each of the phenotypes, but up to 14% of the available intronic regions and up to 10% of the cis-regions surrounding genes contribute to the phenotypic variance with $\geq 95\%$ probability (Figure 3b). The variance contributed by each exonic, intronic, promotor, or cis region is typically only a small fraction of a percent, with largest effect sizes being the exonic region of GDF5 contributing 0.26% (95% CI 0.21, 0.32) to the phenotypic variance of height, the intronic region of FTO contributing 0.48% (95% CI 0.29, 1.12) to BMI, both the exonic- and intronic-region of LPA contributing a combined 0.08% (95% CI 0.04, 0.13) to the risk of CAD, and the intronic region of TCF7L2 contributing 0.28% (95% CI 0.23, 0.35) to the risk of T2D (Figure 3c, full results in Table S2 to S5). Taken together, these results support an infinitesimal contribution of many thousands of genes to common complex trait variation and give joint estimates of the proportions of variance contributed by each gene and their probability of association.

For each gene, we also calculated the phenotypic variance contributed by exonic, intronic, promotor region, and cis SNPs and then calculated the correlation among the variances explained by the groups across genes. Across traits, we find small positive correlations of the variance attributable to exonic and intronic regions of 0.17 (0.09, 0.24 95% CI) for height, 0.02 (0.001, 0.05 95% CI) for BMI, 0.103 (-0.007, 0.71 95% CI) for CAD, and 0.064 (0.01, 0.19 95% CI) for T2D. Similarly, we find small positive correlations between introns and cis regions(Figure 3d). With the exception of height, the variance attributable to the following groups were independent: (i) SNPs in the exons of each gene and SNPs +/- 50kb outside of the exon and promotor regions; (ii) SNPs in the exons of each gene and SNPs in proximal promotors; and (iii) intronic SNPs and SNPs in promotor regions (Figure 3d). This implies that trait associated SNPs in proximal and distal regulatory regions are largely independent of the effects of SNPs in their closest exon, as they do not align in terms of the variance they explain (Figure 3d). For height, small weakly positive correlations across all gene regions in their contribution to variance, implies a degree of alignment across genes in regulatory variants and the closest exon (Figure 3d). These results suggest a regulatory link between introns and distal cis regions outside of the promotor, or that introns may be correlated with structural variation. They also imply that the variance contributed by regulatory regions and those in the closest coding regions are not strongly coupled for these common complex traits.

Finally, our approach provides automatic fine-mapping of SNP loci, and of these region- and gene-level associations, 360 SNPs for height, 20 for BMI, 2 for CAD and 9 for T2D could be mapped to a single SNP with greater than 95% inclusion probability across all 4 chains (Supplementary Table S6, Figure S13). Of these fine-mapped SNPs, only 53.45% are top loci with a p-value $< 5x10^{-8}$ from the fastGWAS UK Biobank summary statistic data for standing height, BMI, angina / heart attack and type-2 diabetes (fastGWA, see Code Availability). This indicates that selecting on the top SNP markers identified by standard association studies would give a different set of variants than those obtained from a BayesRR-RC model.

Out-of-sample prediction into another European healthcare system

Finally, we then generated a full posterior predictive distribution for each trait in each of 32,500 individuals from the Estonian Genome Centre data, which allows the transmission of uncertainty in the marker effect estimates from the UK Biobank to the genomic predictors created in Estonia. First, despite this study having almost half the sample size, we show improved genomic prediction as compared to recently proposed summary statistic approaches [25], when taking the mean of the predictor across iterations and correlating this with the phenotype with correlation of 0.62 for height, 0.34 for BMI, 0.16 for T2D, and 0.07 for CAD (Figure 4a). The area under the receiver operator curve (AUC) for T2D was 0.67 and 0.57 for CAD. We then estimated the distribution of the partial correlations between the trait and genomic predictors created from our different annotation groups and find that exonic, intronic, and 10-500kb upstream regions contribute proportionally more to the prediction accuracy than other genomic groups, replicating our results from the UK Biobank (Figure 4b).

Our approach enables a posterior predictive distributions to be generated for each individual. As an alternative measure of prediction accuracy, for height and BMI we determined the proportion of the posterior predictive distribution for each individual that was within +/-1 SD of their true phenotypic value. On average

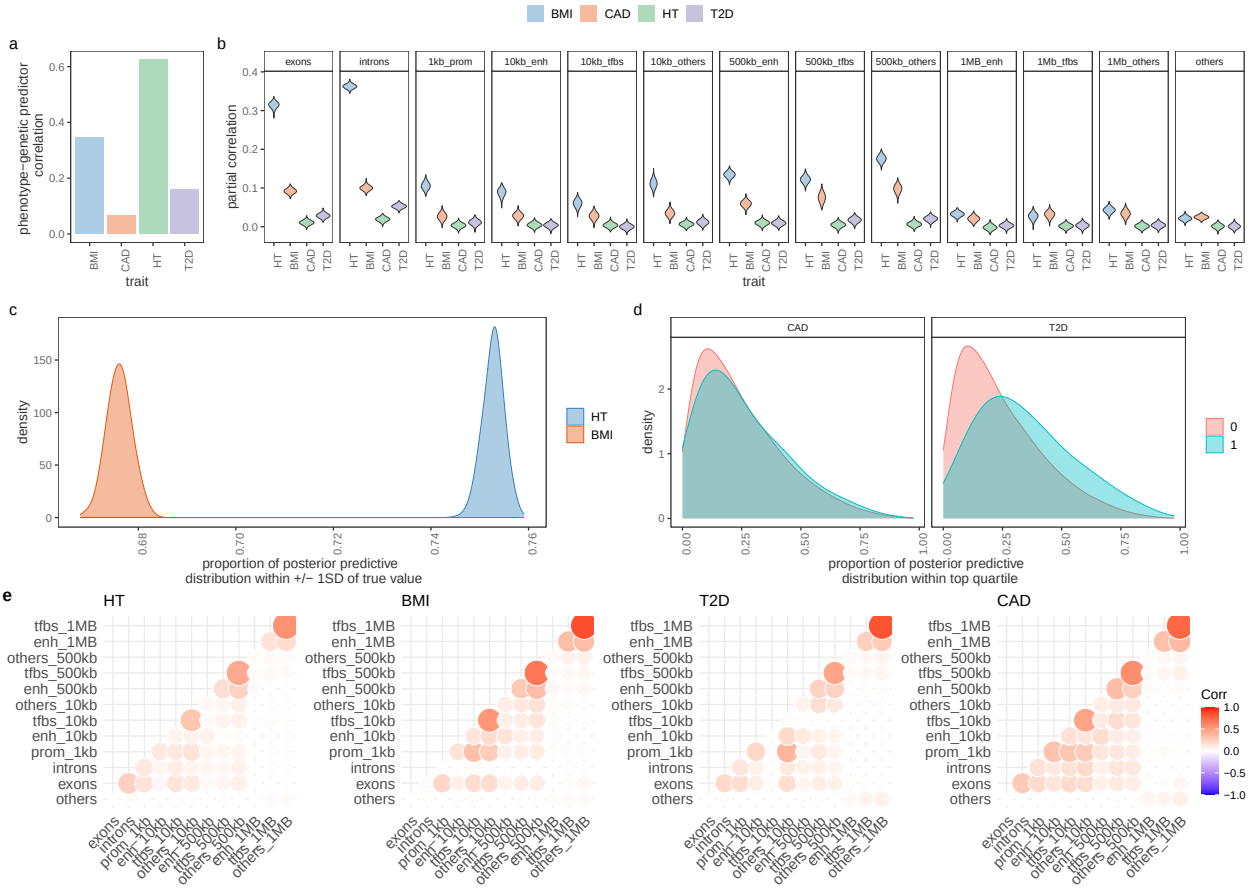

Figure 4. Cross-cohort prediction accuracy and the posterior predictive distribution. (a) Correlation of the posterior mean predictor and height (HT), body mass index (BMI), type-2 diabetes (T2D), and cardiovascular disease (CAD). (b) the partial correlations of the phenotype and genomic predictors specific to different genomic annotations. (c) For height and BMI, we calculate the probability that the distribution of genomic predictors obtained for each individual is within 1 SD of the true phenotypic value. The density of these probabilities is shown. (d) Correlation of genetic predictors obtained across annotation groups.

67.5% of an individuals posterior predictive distribution is within +/-1 SD of their true phenotype for BMI and 75% for height, with similar prediction accuracy across individuals (Figure 4c). For T2D and CAD, we extended the PCF metric, typically defined as the proportion of cases with larger estimated risk the then top p^{th} percentile of the distribution of genetic risk in the general population. For each individual, we calculated the proportion of their posterior predictive distribution that falls above the top 25% of the distribution of genetic risk in the general population. The distribution of these probabilities is shown for confirmed cases and those without diagnosis in the Estonian Biobank (Figure 4d). We find 25 individuals for T2D and 15 individuals for CAD where $\geq 90\%$ of their posterior predictive distribution is within the high risk group of which 40% and 18% are currently defined as cases for T2D and CAD respectively based on recent medical records. This is compared to 1% and 2% case rate for those with $\leq 10\%$ probability of being in the high risk group for T2D and CAD respectively, giving an odds ratio of 20 and 18 between the $\geq 90\%$ and $\leq 10\%$ groups. However, our results clearly show that the individual-level sensitivity and specificity of genomic prediction for these common complex diseases is very poor, as 75% of T2D cases and 92% of CAD cases have $\leq 50\%$ of their distribution within the high-risk category. Thus, genomic prediction for personalized medicine with patient-specific predictions will remain limited for these diseases without vastly increased study power.

We find evidence for zero/low correlations of genomic predictors created from different annotation groups, which supports our results from the UK Biobank (Figure 4e). This suggests that individuals have a different portfolio of risk variants, with different genomic regions contributing for different individuals to their overall genetic value, as expected under a highly polygenic model . These results highlight the variation contained within a posterior predictive distribution that is typically ignored in human genomic prediction.

Discussion

Here, we have shown that a grouped Dirac spike-and-slab model (termed BayesRR-RC), explicitly quantifies the uncertainty on estimation and prediction under minimal assumptions irrespective of the underlying genetic architecture of the trait, or the structure of the genomic data. The posterior distributions obtained allow for direct fine-mapping of individual SNP effects, give a probabilistic understanding of the relative contributions of different genes and genomic regions, provide a distribution of polygenic risk scores for each individual that are specific to different regions of the DNA, and facilitate comparisons across traits of the underlying genetic architecture of different genomic groups. We develop a range of computational and statistical approaches which allow this, or any similar Gibbs sampling algorithm, to scale to whole genome sequence data on many hundreds of thousands of individuals.

There has been debate on how to best estimate SNP heritability [1, 3, 4] and here we validate the need to split SNP markers by LD to obtain unbiased genetic effect size estimates, demonstrating through theory and simulation why penalized regression models inaccurately estimate effects under multicollinearity and how differential shrinking of SNPs corrects this bias. Our results show the same pattern of total variance partitioning for height, BMI, CAD and T2D in-line with recent results from SumHer [6]. However, we observe that all phenotypes simply appear to be predominantly underlain by very many common variants, with SNPs within distal regulatory regions, coding and intronic regions each contributing more to the phenotypic variance and having higher allele substitution effects.

Recent studies have also attempted to quantify the gene architecture of complex traits, in terms of the number and contribution to phenotypic variance of markers either in coding regions, or directly involved in the expression of genes [26, 27]. Our results suggest that the proportion of genomic variation attributable to mutations in regulatory regions and mutations in the closest genic regions are largely independent. Additionally our model tests association within groups in a probabilistic way and we find 290 independent coding, 2,888 independent intronic, and 5,406 independent cis regions with $\geq 95\%$ probability of of contributing at least 0.001% of the SNP heritability. A challenge is to now better understand how these coding, intronic and proximal and distal regulatory regions combine to contribute to phenotypic variance and our results suggest a predominant role for introns and for distal, and thus likely more global enhancers, rather than locally dominant proximal expression QTL. The recent "omnigenic" model [28], suggests that trait-associated variants in regulatory regions influence a local gene which is not directly causal to the disease, and also co-regulate other disease causal genes (or "core" gene). Our findings of little correlation of exonic and proximal regulatory variance and a large number of trait-associated intronic and cis regions do not rule this out, but suggest a more complex infinitesimal picture with differences occurring among traits, potentially due to their evolutionary history.

There are important caveats and limitations to consider. In this work, we do not extend past a limited number of functional annotations and thus we do not provide a model capable of further partitioning the variation into specific regulatory functions (eQTL, mQTL, pQTL etc.) or directly modelling the relationships among components. Doing this requires the use of more information in the prior, allowing more groups, potentially allowing markers to swap groups with a prior probability of function, and allowing for correlations in marker effects across groups. While our future work is in this direction, a first requirement is an improvement in annotations as MAF-LD multicollinearity biases have to be removed from studies of eQTL, mQTL, pQTL etc. before these annotations can be reliably used, as otherwise marker function will likely be biased by the data structure (e.g. common, high LD variants may be more likely to be allocated as eQTL). LDSC functional methods take the approach that SNPs can be assigned to different categories (e.g. both coding and conserved), with the categories competing against each other to explain the signal, with the downside that enrichment is relative and that the total variance is not partitioned. Here, the total variance is partitioned but this is based on preferential allocation of SNPs to coding regions, introns, and then to their nearest upstream gene position. Coding regions, introns and 10-500kb distal regions could contribute the most variance as these SNPs are most likely to be allocated accurately, with 1kb and 1-10kb groups being more ambiguous in high gene density regions and likely mislabelled. However, if this was the case then variance would still be partitioned to these mislabelled groups and it would just be evenly split across them, with experimentally validated promotor, enhancer and tfbs regions assisting to some degree in alleviating this. This was not the case, and here we see a clear pattern of increasing variance contributed, increasing average effect size, and an increasing pattern of higher rare allele substitution effects by individual markers as distance from the nearest gene increases. 10-500kb distal regions may contribute more variance as marker density and marker coverage is higher in these regions, with missing variation within 10kb upstream as causal variants are poorly correlated with SNPs. The posterior distributions for the variance explained by 1kb, 1-10kb regions, and 10-500kb regions are negatively correlated (Figure S9, meaning that these groups are competing with each

other, as if variance goes to one then it is being taken away from the other (as they are in LD), and thus there is the risk that the model cannot separate these effectively. However, this is true of any enrichment analysis conducted to date and we can only make inference in the data that we have currently available. Resolving this requires the application of this model to whole genome sequence data where the total variance can be partitioned across upstream regions without marker coverage concerns. Irrespective of exactly which upstream region variance is allocated to, our inference that genic regions are uncorrelated in their contribution to variance with the promotor and upstream regions still holds as does our probabilistic inference on the associations of each gene and their contribution to the phenotypic variation.

Other approaches may also provide continuous SNP shrinkage, regularising each SNP differently, such as a Finnish horseshoe model [29], and we are working to place a grouped version of this model within our computational framework to explore this possibility. Recent work has shown that tree sequence algorithms can also be used to massively increase the scalability of methods for genomic data, making it possible to infer trees for millions of samples [30] and to conduct regression models using tools such as TreeLD [31] or inferred ancestral recombination graphs [32]. We expect that our current algorithm combining sparse dot products and highly vectorized look-up tables to outperform these methods in terms of performance as there are costs to tree-traversal and tree-calculation. However, a tree-approach would provide benefits in terms of memory usage and future work to computationally engineer the tree-structure data may be beneficial. Finally, our focus is limited to two common complex diseases with case proportions 11.6% for CAD and 7.2% for T2D within the UK Biobank. Less prevalent complex diseases, likely require additional model extensions to the prevent effect size bias as reported elsewhere [10] and this will also be a focus of future work.

Summary

Our results provide evidence for an infinitesimal contribution of many thousands of common genomic regions to common complex trait variation and for a predominant role of intronic, exonic, and distal regulatory regions. This highlights the immense challenge of understanding the molecular underpinning of each association and the difficulties in improving the estimation of many tens of thousands of small-effect associations that are required to improve genomic prediction. This work represents a first step toward maximising the probabilistic inference that can be obtained from large-scale Biobank studies.

Methods

Model Specification

We begin by outlining the basic model bayesR, before then presenting our extensions. Consider p single nucleotide polymorphism (SNP) markers. If we gather samples for $i = 1, ...N$ subjects in an $N \times p$ matrix, $\mathbf{G}$, in which the elements are coded as 0 for homozygous individuals at the major allele, 1 for heterozygous individuals and 2 for minor allele homozygotes. Now, we wish to model their linear association with the phenotype $\mathbf{y} = (y_i)$ of subjects $i = 1, \ldots, N$ in a standard linear regression model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

We assume that the genotypes are standardized so that $\mathbf{X}_j = \frac{(\mathbf{G_j} - \mu_j \mathbf{1})}{\sigma_j}$ is the vector of genotypes for the j^{th} marker ($j = 1, p$) with zero mean and unit variance, i.e. the centered and scaled j^{th} column of $\mathbf{G}$. The column's mean $\mu_j \approx 2f_j$ and the column's standard deviation $\sigma_j \approx \sqrt{2f_j\,(1 - f_j)}$ being f_j the minor allele frequency(MAF) of the SNP. We define $\boldsymbol{\beta}$ as a $p \times 1$ vector of partial regression coefficients with β_j the effect of a 1 SD change in the j^{th} covariate, and $\boldsymbol{\epsilon}$ is a vector (N x 1) of residuals.

We estimate the model's parameters using Bayesian inference, assuming that the error term $\boldsymbol{\epsilon}|\sigma^2_\epsilon \sim \mathcal{N}\left(0, \mathbf{I}\sigma^2_\epsilon\right)$. The log-likelihood of this model can be written as

$$l\left(\mu, \boldsymbol{\beta}, \sigma^2_\epsilon\right) = -\frac{N}{2}\log\left(2\pi\sigma^2_\epsilon\right) - \frac{1}{2\sigma^2_\epsilon}\left(N\left(\hat{y} - \mu\right)^2 + \left(\mathbf{y}_c - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}_c - \mathbf{X}\boldsymbol{\beta}\right)\right) \tag{4}$$

with $\mathbf{y}_c = \frac{(\mathbf{y} - \mathbf{1}\mu)}{\sigma_y}$ a vector of centred and scaled responses(SD 1).

As we adopt a Bayesian approach, we place priors over the model parameters. For the covariate effects, $\boldsymbol{\beta}$, we use a mixture prior with Dirac spike and slab components, which have been extensively used for variable selection [15, 16]. The prior induces sparsity in the model through a Dirac-delta at zero, excluding variables from the model by setting their coefficients to zero. A slab component is centered at zero and shrinks the

non-zero coefficients towards zero according to the slab's width. In our approach, the slab component is a scale mixtures of normals and thus each $\beta_j \in \boldsymbol{\beta}$ is distributed according to:

$$\beta_j \sim \pi_0\delta_0 + \pi_1\mathcal{N}\left(0, \sigma_1^2\right) + \ldots + \pi_L\mathcal{N}\left(0, \sigma_L^2\right)$$

where $\pi_\beta = (\pi_0, \pi_1, \ldots, \pi_L)$ are the mixture proportions, $\left\{\sigma_1^2, \ldots, \sigma_L^2\right\}$ are the mixture-specific variances, and δ_0 is a discrete probability mass at zero. We further constrain the prior by assuming a single parameter representing the total variance explained by the effects σ_G^2, with the component-specific variances proportional to σ_G^2 multiplied by a constant $\{C_1, \ldots, C_L\}$ so that

$$\begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_L^2 \end{bmatrix} = \sigma_G^2 \begin{bmatrix} C_1 \\ \vdots \\ C_L \end{bmatrix}$$

The remaining prior structure for the model is then

$$\begin{aligned} \pi &\sim \text{Dirichlet}\left(\mathbf{1}\right) \\ \sigma_G^2 &\sim \text{Inv}-\text{Scaled}\chi^2\left(\text{v}_0, \text{s}_0^2\right) \\ \sigma_\epsilon^2 &\sim \text{Inv}-\text{Scaled}\chi^2\left(\text{v}_0, \text{s}_0^2\right) \end{aligned} \tag{5}$$

with weakly informative parameters for hyperparameters $\text{v}_0 = \text{s}_0^2 = 0.001$.

For notational convenience, we will refer to the mixture membership labels as $(l_0, l_1, \ldots, l_L)$ and we define a latent indicator of each SNP, j, $\gamma = (\gamma_j, \ldots, \gamma_p)^T$ with $\gamma_{j,l} = 0$ or 1, indicating whether or not the effect of SNP j falls into the zeroth mixture $\gamma_{j,l} = 0$, or follows a normal distribution with variance σ_l^2. We define the "active set of coefficients" as those β_j such that $\beta_j \neq 0$ denoted as $\boldsymbol{\beta}_{\gamma\neq 0}$ with cardinality $||\gamma_\varphi||_0$. Thus the objective of our inference scheme is to compute an estimate of the posterior distribution $f\left(\boldsymbol{\beta}_{\gamma\neq 0}, \sigma_\epsilon^2, \sigma_G^2, \mu|\mathbf{y_c}\right)$. This model has been termed BayesR [12, 13] and an effective proposed Gibbs sampling scheme [13] follows the following steps:

(i) sample μ from $\mathcal{N}\left(\frac{\sum_{i=1}^{N}(\mathbf{y_c}_i - \mathbf{X}_j\boldsymbol{\beta}_{\gamma\neq 0})}{N}, \frac{\sigma_\epsilon^2}{N}\right)$

(ii) sample $\boldsymbol{\beta}_{\gamma\neq 0}$ from its conditional as described below

(iii) sample σ_G^2 from $\text{Inv}-\text{Scaled}\chi^2\left(||\gamma_\varphi||_0 + \text{v}_0, \frac{||\gamma_\varphi||_0\|\boldsymbol{\beta}_{\gamma\neq 0}\|^2 + \text{v}_0 S_0^2}{\text{v}_0 + ||\gamma_\varphi||_0}\right)$

(iv) sample σ_ϵ^2 from $\text{Inv}-\text{Scaled}\chi^2\left(\text{v}_0 + N, \frac{\|\mathbf{y_c} - \mu - \mathbf{X}\boldsymbol{\beta}_{\gamma\neq 0}\|^2 + \text{v}_0 S_0^2}{\text{v}_0 + N}\right)$

From the former algorithm, steps (i), and (iv) are straight-forward applications of conjugacy and are common to many Gibbs sampling algorithms for linear regression. Step (iii) follows from conjugacy and the assumption that the individual mixtures represent fractions of the total variance explained by the coefficients. Step (ii) is the biggest bottleneck in any linear regression problem, and in the next section we will proceed to detail the derivations of the sampling scheme for this step.

While it is not uncommon to use non-proper priors for the residual's variance σ_ϵ^2, in our case we chose to keep a proper prior for algorithmic and modeling reasons as: a) conjugacy is amenable to Gibbs sampling b) we assume σ_ϵ^2 and σ_G^2 are not nuisance parameters, and in some cases we possess prior information on its distribution. It is also common to specify the distribution of β_j having a variance depending on the residual's variance σ_ϵ^2, which would make the estimates transformation-invariant. Recent results suggest the estimates for σ_ϵ^2 in this latter transformation-invariant formulation are biased [33]. Another concern may be that the prior's hyperparameters induce biased estimates for small variances [34], we acknowledge that may be an issue, and allow parameters v_0, s_0^2 to be adjusted if deemed necessary. The scale mixture of Gaussians, allows the prior distribution to have heavier tails than a single Gaussian, which allows big effects to be shrunk to a lesser degree than small effects [17]. Finally, the original formulation of [12, 13] assumes $\sigma_G^2 = r^2\sigma_y$ which for centered and scaled phenotypes and genotypes, with heritability h^2 equal to reliability $r^2 = \frac{\text{Var}[\mathbf{X}\boldsymbol{\beta}_{\gamma\neq 0}]}{\text{Var}[y]}$, would mean $\sigma_G^2 = h^2 = r^2 = \text{Var}\left[\mathbf{X}\boldsymbol{\beta}_{\gamma\neq 0}\right] = \sum_{\gamma\neq 0}\boldsymbol{\beta}_{\gamma\neq 0}^2$, but there is no constraint in the model ensuring $\sigma_G^2 + \sigma_\epsilon^2 = \sigma_y^2$. As we will see, further assumptions are necessary for having unbiased estimates of σ_G^2 and h^2

under varying LD and MAF. These estimates will achieve the equivalence $\sigma^2_G = r^2 = h^2$ without relying in either using a point estimate of r^2 [12], informative priors on σ^2_G, or normalising the posterior variances by $h^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_\epsilon}$ [14].

Sampling the effects

For sampling $\boldsymbol{\beta}$, the challenge is two-fold: (a) determining if the effect β_j is part of $\boldsymbol{\beta}_{\gamma \neq 0}$, and if so, to which component it belongs; and then (b) sampling the vector $\boldsymbol{\beta}_{\gamma \neq 0}$ from a multivariate Gaussian with covariance matrix $\Sigma = \mathbf{X}^T_{l \neq 0}\mathbf{X}_{l \neq 0} + \Lambda$ where Λ is the diagonal matrix with entries $\lambda_{l,j} = \frac{\sigma^2_\epsilon}{\sigma^2_{j,l}}$, with $\sigma^2_{j,l}$ the variance of the mixture component to which marker β_j was assigned. For (a), marginalization of each effect individually is required to compute the membership probability, which requires solving a determinant of the size of $||\gamma_\varphi||_0 - 1$ [16]. For (b), either a system of size $||\gamma_\varphi||_0$ must be solved through LU decomposition, or Cholesky decomposition of size $||\gamma_\varphi||_0$, and both operations are resource intensive when the size of $||\gamma_\varphi||_0$ is large. Instead, we determine the inclusion of a marker in the active set, along with its mixture membership and its partial regression coefficient β_j, in single-site updates. Single-site Gibbs sampling is also known as stochastic relaxation [35] has a long history given its equivalence to iterative Gauss Siedel methods to solve matrix equations [36]. Although we choose to use the BayesR model, many alternative models can easily be placed within the iterative solving and computational framework we outline here.

In this scheme, we sample each element, j, of $\boldsymbol{\beta}$ from the full conditional posterior $f\left(\beta_j | \boldsymbol{\beta}_{\setminus j}, \mathbf{y}\right) \propto f\left(\beta_j, \boldsymbol{\beta}_{\setminus j}, \mathbf{y}\right)$ which can be written as $f\left(\beta_j, \boldsymbol{\beta}_{\setminus j}, \mathbf{y}\right) = f\left(\mathbf{y}|\boldsymbol{\beta}\right) f\left(\beta_j\right) f\left(\boldsymbol{\beta}_{\setminus j}\right)$ where $f\left(\mathbf{y}|\boldsymbol{\beta}\right)$ is the density function of the conditional distribution of $\mathbf{y}|\boldsymbol{\beta}$ and $f\left(\beta_j\right)$ and $f\left(\boldsymbol{\beta}_{\setminus j}\right)$ are the densities of the prior distributions of β_j and $\boldsymbol{\beta}_{\setminus j}$ respectively, with notation $\setminus j$ representing all other covariates except j. The kernel of the full conditional posterior for β_j is proportional to the product of the likelihood, the prior distribution for β_j and the prior distributions of the variances, and thus ignoring factors that are constant with respect to β_j gives

$$f\left(\beta_j \mid l_j, \boldsymbol{\theta}_{\setminus j}, \mathbf{y}\right) \propto \exp\left[-\frac{\left(\mathbf{y_c} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y_c} - \mathbf{X}\boldsymbol{\beta}\right)}{2\sigma^2_\epsilon}\right] \exp\left[-\frac{\beta^2_j}{2\sigma^2_{j,l}}\right] \tag{6}$$

where l_j represents the mixture β_j is assigned, $\boldsymbol{\theta}_{\setminus j} = \{\beta_{\setminus j}, \sigma^2_\epsilon, \sigma^2_G, \pi_\beta, \mu\}$ and $\sigma^2_{j,l}$ the corresponding mixture variance. We can reduce the expanded form and drop terms that are free from β_j as

$$\begin{aligned}
f\left(\beta_j \mid l_j, \boldsymbol{\theta}_{\setminus j}, \mathbf{y}\right) &\propto \exp\left[-\frac{1}{2\sigma^2_\epsilon}\left(\mathbf{y_c} - \mathbf{X}_j\beta_j - \mathbf{X}_{\setminus j}\boldsymbol{\beta}_{\setminus j}\right)^T\left(\mathbf{y_c} - \mathbf{X}_j\beta_j - \mathbf{X}_{\setminus j}\boldsymbol{\beta}_{\setminus j}\right) + \frac{\beta^2_j\sigma^2_\epsilon}{2\sigma^2_{j,l}}\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2_\epsilon}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - 2\mathbf{X}^T_j\tilde{\mathbf{y}}\beta_j + \mathbf{X}^T_j\mathbf{X}_j\beta^2_j + \frac{\beta^2_j\sigma^2_\epsilon}{2\sigma^2_{j,l}}\right)\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2_\epsilon}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - 2\mathbf{X}^T_j\tilde{\mathbf{y}}\beta_j + \beta^2_j\Sigma_{j,l}\right)\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2_\epsilon}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - 2\hat{\beta}_j\Sigma_{j,l}\beta_j + \beta^2_j\Sigma_{j,l} + \hat{\beta}^2_j\Sigma_{j,l} - \hat{\beta}^2_j\Sigma_{j,l}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\frac{(\beta_j - \hat{\beta}_j)^2}{\frac{\sigma^2_\epsilon}{\Sigma_{j,l}}}\right]
\end{aligned} \tag{7}$$

with $\tilde{\mathbf{y}} = \mathbf{y}_c - \mathbf{X}_{\setminus j}\boldsymbol{\beta}_{\setminus j}$, $\Sigma_{j,l} = \mathbf{X}^T_j\mathbf{X}_j + \lambda_{j,l}$ and $\hat{\beta}_{j,l} = \frac{\mathbf{X}^T_j\tilde{\mathbf{y}}}{\Sigma_{j,l}}$. This gives the Gibbs sampling update for β_j as

$$\beta_j \sim \mathcal{N}(\Sigma^{-1}_{j,l}\mathbf{X}^T_j\tilde{\mathbf{y}}, \sigma^2_\epsilon\Sigma^{-1}_{j,l}) \tag{8}$$

To avoid reducibility of the Markov chain, prior to drawing the effect β_j, we first need to select the mixture K for each covariate j, and as above we can condition on the individual coordinates and to obtain the probability that a coefficient j belongs to a given mixture.

$$\mathbb{P}\left(l_j = K \mid \boldsymbol{\theta}_{\setminus j}, \mathbf{y}\right) = \frac{f\left(\tilde{\mathbf{y}} \mid l_j = K, \boldsymbol{\theta}, \mathbf{y}\right)\mathbb{P}\left(l_j = K\right)}{\sum^L_{k=1} f\left(\tilde{\mathbf{y}} \mid l_j = k, \boldsymbol{\theta}, \mathbf{y}\right)\mathbb{P}\left(l_j = k\right)} \tag{9}$$

We integrate out the β_j coordinate following the equations above with

$$f\left(\tilde{\mathbf{y}} \mid l_j, \boldsymbol{\theta}, \mathbf{y}\right) = \int f\left(\tilde{\mathbf{y}} \mid \beta_j, \sigma_\epsilon^2\right) f\left(\beta_j \mid l_j, \sigma_{j,l}^2\right) d\beta_j$$
$$= \int (2\pi\sigma_\epsilon^2)^{-n/2} \exp\left[-\frac{(\tilde{\mathbf{y}} - \mathbf{X}_j\beta_j)^T(\tilde{\mathbf{y}} - \mathbf{X}_j\beta_j)}{2\sigma_\epsilon^2}\right] (2\pi\sigma_{j,l}^2)^{-q/2} \exp\left[-\frac{\beta_j^2}{2\sigma_{j,l}^2}\right] d\beta_j$$

where $q = 2$. We then expand this equation using the relationship $\Sigma_{j,l}\hat{\beta}_j = \mathbf{X}_j^T\tilde{\mathbf{y}}$ from Eq. 8 and complete the squares

$$f\left(\tilde{\mathbf{y}} \mid l_j, \boldsymbol{\theta}, \mathbf{y}\right) = \int (2\pi\sigma_{j,l}^2)^{-q/2}(2\pi\sigma_\epsilon^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - 2\hat{\beta}_{j,l}\Sigma_{j,l}\beta_j + \beta_j^2\Sigma_{j,l} + \hat{\beta}_{j,l}^2\Sigma_{j,l} - \hat{\beta}_{j,l}^2\Sigma_{j,l}\right)\right] d\beta_j$$
$$= (2\pi|\sigma_\epsilon^2\Sigma_{j,l}^{-1}|)^{1/2}(2\pi\sigma_{j,l}^2)^{-q/2}(2\pi\sigma_\epsilon^2) \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - \hat{\beta}_{j,l}^2\Sigma_{j,l}\right)\right] \times$$
$$\int (2\pi|\sigma_\epsilon^2\Sigma_{j,l}^{-1}|)^{-1/2} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\beta_j - \hat{\beta}_{j,l})^2\Sigma_{j,l}\right)\right] d\beta_j$$
$$= \left(|\lambda_{l,j}\Sigma_{j,l}^{-1}|\right)^{\frac{1}{2}} \left(2\pi\sigma_\epsilon^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - \hat{\beta}_{j,l}^2\Sigma_{j,l}\right)\right] \tag{10}$$

where the final reduction in Eq. 10 occurs as the integral component is now a normal distribution that integrates to 1 and then terms are removed that do not contain, nor depend upon $\Sigma_{j,l}$ nor $\hat{\beta}_{j,l}$. The probability for inclusion in the model in the first mixture, as compared to the spike, then depends upon the ratio

$$\frac{f\left(\tilde{\mathbf{y}} \mid l_j = 0, \boldsymbol{\theta}, \mathbf{y}\right)}{f\left(\tilde{\mathbf{y}} \mid l_j = 1, \boldsymbol{\theta}, \mathbf{y}\right)} = \frac{\left(2\pi\sigma_\epsilon^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}})\right]}{\left(|\lambda_{l,j}\Sigma_{j,2}^{-1}|\right)^{\frac{1}{2}} \left(2\pi\sigma_\epsilon^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} - \hat{\beta}_{j,l}^2\Sigma_{j,2}\right)\right]}$$
$$= \left(|\lambda_{l,j}\Sigma_{j,2}^{-1}|\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}) + \frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}) - \frac{1}{2\sigma_\epsilon^2}(\hat{\beta}_{j,l}^2\Sigma_{j,2})\right]$$
$$= \left(|\lambda_{l,j}\Sigma_{j,2}^{-1}|\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}(\hat{\beta}_{j,l}^2\Sigma_{j,2})\right] \tag{11}$$

Analogous to equation 11, any comparison between mixtures has the same form and allows us to omit the $\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}$ term. Thus placing Eq.11 into Eq.9 and re-arranging to a numerically more stable version [12] gives

$$\mathbb{P}\left(l_j = K|\boldsymbol{\theta}_{\setminus j}, \mathbf{y}\right) = \frac{1}{1 + \sum_{k=0}^{L} \exp\left[\log(LK_K) - \log(LK_k)\right]} \tag{12}$$

with $\log(LK_0) = \log(\pi_0)$ and $\log(LK_l) = -\frac{1}{2}\left[-\log\left(|\lambda_{l,j}\Sigma_{j,l}^{-1}|\right) - \left(\frac{\hat{\beta}_{j,l}^2\Sigma_{j,l}}{\sigma_\epsilon^2}\right)\right] + \log(\pi_l)$ for l in $(1...L)$.

Having derived the regression coefficients and their inclusion probabilities, fully specifying the BayesR model, we now proceed in the following sections to: (1) derive the properties of the model parameters when applied to highly correlated genomic data (under multicollinearity) and compare these to estimates made by other approaches in the field; (2) extend the model to account for genomic annotations, minor allele frequency (MAF) and linkage disequilibrium (LD) among markers; and finally (3) derive a computational implementation that facilitate the application of the model to biobank sized data.

Comparison to other approaches under collinearity

Genome-wide association studies have predominantly been conducted using single marker regression via ordinary least squares (OLS). Recently, it has been proposed that if aggregation due to familial or molecular similarity (e.g. population stratification) exists in the data, a better estimation approach is generalized least squares (GLS), as it poses a more general covariance structure than OLS. GLS estimates can be obtained within mixed-linear association models, which first declare all marker effects as random variables, for example, assuming that $u_j \sim N(0, \sigma_u^2)$, or from a mixture of distributions, with all markers in the set taken as

independently and identically distributed random variables. Second, when the markers are evaluated for association, they are then treated as a fixed effect. The resulting model can be written as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_1 u_1 + \mathbf{X}_{\setminus 1}\mathbf{u}_{\setminus 1} + \boldsymbol{\epsilon} \tag{13}$$

where a focal genetic marker, here $\mathbf{X}_1$ is fitted twice, first as a fixed effect to estimate the regression coefficient β_1, and also as part of all of the other markers with their effects, u, estimated as random (note here $\setminus 1$ indicates all markers other than marker 1). Under this model the phenotypic covariance structure is

$$\mathbf{V} = \mathbf{X}_1\mathbf{X}_1^T\sigma^2_G + \mathbf{X}_{\setminus 1}\mathbf{X}_{\setminus 1}^T\sigma^2_G + \mathbf{I}\sigma^2_\epsilon \tag{14}$$

With orthogonal covariates, the estimated variance components that compose $\mathbf{V}$ can remain constant when testing each marker in turn. However, with collinearity among markers the situation becomes more complex. Below, we first describe the impact of multicollinearity on ridge regression estimates. We then outline the equivalence of a ridge regression and a mixed linear model, before then demonstrating increased variance of the estimates obtained from Eq. (13) under multicollinearity. Finally, we then go on to show that estimates from BayesR are less subject to inflated variance, except under extensive multicollinearity, before then describing how extending the model to provide minor allele frequency and LD specific hyperparameters provides estimates with improved properties across a range of underlying generative data models.

In Eq. (13) if markers were all simply estimated as random, following a single distribution, then a ridge regression estimator of Hoerl and Kennard 1970 [21] would be obtained, which was proposed to replace $\mathbf{X}^T\mathbf{X}$ in the OLS solutions by $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$, with $\lambda \in [0, \infty]$ a tuning or penalty parameter. This gives the ridge regression estimator

$$\hat{\boldsymbol{\beta}}(\lambda) = [\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{Y} \tag{15}$$

where λ is strictly positive and the solution or regularization path of the ridge estimate $\hat{\boldsymbol{\beta}}(\lambda) : \lambda \in [0, \infty]$ is the set of ridge estimates across the values of λ. The expectation of the ridge estimator

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathbb{E}[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}\end{aligned} \tag{16}$$

with $\hat{\boldsymbol{\beta}}$ the maximum likelihood OLS estimator. If we consider an orthonormal design matrix $\mathbf{X}$, with $\mathbf{X}^T\mathbf{X} = \mathbf{I} = (\mathbf{X}^T\mathbf{X})^{-1}$ then we can express the relationship between $\hat{\boldsymbol{\beta}}$, and the ridge estimator, $\hat{\boldsymbol{\beta}}(\lambda)$, as

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{I} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}\mathbf{I}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1 + \lambda\mathbf{I})^{-1}\hat{\boldsymbol{\beta}}\end{aligned} \tag{17}$$

If we define $\mathbf{W}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})$ then the ridge estimator $\hat{\boldsymbol{\beta}}(\lambda)$ can be expressed as $\mathbf{W}_\lambda\hat{\beta}$ for

$$\begin{aligned}\mathbf{W}_\lambda\hat{\boldsymbol{\beta}} &= \mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= [(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \hat{\boldsymbol{\beta}}(\lambda)\end{aligned} \tag{18}$$

The variance of the ridge estimator is then

$$\begin{aligned}\mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathrm{Var}[\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}] \\ &= \mathbf{W}_\lambda\mathrm{Var}[\hat{\boldsymbol{\beta}}]\mathbf{W}_\lambda^T \\ &= \sigma^2_\epsilon\mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{W}_\lambda^T \\ &= \sigma^2_\epsilon(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}]^T\end{aligned} \tag{19}$$

and the mean square error of $\hat{\boldsymbol{\beta}}(\lambda)$ is

$$\begin{aligned}
\mathbf{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathbb{E}[(\mathbf{W}_\lambda\hat{\boldsymbol{\beta}})^T(\mathbf{W}_\lambda\hat{\boldsymbol{\beta}})] \\
&= \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\boldsymbol{\beta}^T\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^T\boldsymbol{\beta}) \\
&= \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta}) \\
&- \mathbb{E}(\beta^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}) + \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta}) \\
&- \mathbb{E}(\boldsymbol{\beta}^T\mathbf{W}_\lambda\boldsymbol{\beta}) - \mathbb{E}(\hat{\boldsymbol{\beta}}^T\mathbf{W}_\lambda^T\boldsymbol{\beta}) - \mathbb{E}(\boldsymbol{\beta}^T\boldsymbol{\beta}) \\
&= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&- \boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{W}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{W}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{W}_\lambda^T\mathbf{W}_\lambda(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] + \boldsymbol{\beta}^T(\mathbf{W}_\lambda - \mathbf{I})^T(\mathbf{W}_\lambda - \mathbf{I})\boldsymbol{\beta} \\
&= \sigma_\epsilon^2\mathrm{tr}[\mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{W}_\lambda^T] + \boldsymbol{\beta}^T(\mathbf{W}_\lambda - \mathbf{I})^T(\mathbf{W}_\lambda - \mathbf{I})\boldsymbol{\beta}
\end{aligned} \tag{20}$$

The first summand is the sum of the variances of the ridge estimator, while the second summand is the squared bias of the ridge estimator. With an orthonormal design matrix, $\mathbf{X}$, Theorem 2 of Theobald 1974 [37] shows:

$$\mathbf{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] = \frac{p\sigma_\epsilon^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{21}$$

which achieves a minimum at $\lambda = p\sigma_\epsilon^2/\boldsymbol{\beta}^T\boldsymbol{\beta} = \sigma_\epsilon^2/\sigma_\beta^2$, with σ_β^2 the variance of the β coefficients. This has been stated in the genetics literature as the optimal shrinkage parameter [38] for a ridge regression. However, this is derived under the assumption of uncorrelated covariates within the design matrix $\mathbf{X}$.

To explore the effects of correlated covariates we use the ridge loss function, defined as

$$\mathcal{L}_{ridge}(\boldsymbol{\beta};\lambda) = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{22}$$

which is the sums-of-squares with a penalty, $\lambda\sum_{j=1}^{p}\beta_j^2$, referred to as the ridge penalty, which shrinks the regression coefficients towards zero. The radius of the ridge constraint, the squared Euclidean norm of β, $||\beta||_2^2$, depends upon λ, $\mathbf{X}$ and $\mathbf{Y}$, and taking its expectation

$$\begin{aligned}
\mathbb{E}[||\hat{\beta}(\lambda)||_2^2] &= \mathbb{E}\left[[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}]^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}\right] \\
&= \mathbb{E}[\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^T\mathbf{Y}] \\
&= \sigma_\epsilon^2\mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^T] + \boldsymbol{\beta}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}
\end{aligned} \tag{23}$$

provides a measure that can be evaluated given different properties of the design matrix $\mathbf{X}$. With the same λ and the same $\boldsymbol{\beta}$, Eq. (23) shows that the degree of collinearity among the covariates alters the variance of the estimated effects. Thus, in a ridge regression penalization does not remove collinearity but simply reduces it's effects on the variance of the ridge estimator provided that the λ value is sufficiently large (and thus the σ_β^2 is small). We explore Eq. (23) in a simulation study described below and presented in Figure 1. This theory is an extension of previous work [39] which showed that the inflation of the SNP heritability is proportional to a ratio of the average LD among causal variants and the markers and the average LD among all the markers, with inflation expected when causal variants are in higher LD with the markers than on average. Eq. (23) is a function of X'X, with the LD values the off-diagonal elements in X'X, but it suggests that inflation would be irrespective of the average LD across the genome, simply being expected if high-LD markers had strong effects and showing that inflation would occur only for the estimates of markers that are in LD with those causal variants. Thus, if SNP heritability is allocated across SNPs at random then estimation will on average be correct, irrespective of the LD among SNPs. If the effects of SNPs vary according to the MAF or LD of the SNP, and assumptions are made that all SNP effects are sampled from the same distribution, then this will lead to bias as the estimates at high-LD markers in strong LD with underlying causal variants will be inflated and this inflation will be sufficiently large and occur at a sufficient number of genomic locations so as to impact upon the global estimate of SNP heritability.

This issue has been detected, and demonstrated in simulation, in a number of recent papers [1–4]. However, to date it has remained little understood from a theoretical perspective. The LD-MAF corrections proposed in the literature all serve to alter the lambda value for SNPs, or sets of SNPs, so that it becomes proportional to the LD and MAF of the marker, in essence reducing the σ_G^2, or making it more specific to the markers

in question, and increasing the λ value for common, highly correlated covariates. The equivalence of ridge regression and mixed-linear models has been shown many times, using well-established results from prediction of random variables dating back to Henderson [40]. The model $\mathbf{Y} = \mathbf{g} + \boldsymbol{\epsilon}$, with $\mathbf{g}$ the genetic value of the individuals, and the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{g} \sim N(0, \mathbf{X}\mathbf{X}^T\sigma_G^2)$ with marker effects thus $\boldsymbol{\beta} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{g}$, are equivalent. Following Henderson [40], assuming σ_ϵ^2 and σ_G^2 are known, with no fixed effect component, the log-likelihood can be shown to be proportional to:

$$\sigma_\epsilon^{-2}\|\mathbf{Y} - \mathbf{g}\|_2^2 + \mathbf{g}^T\mathbf{I}\sigma_G^2\mathbf{g} \tag{24}$$

equating the partial derivatives of this mixed model loss function with respect to $\mathbf{g}$ to zero, yields the estimating equations known as Henderson's mixed model equations. Returning to the mixed linear association model described in Eq.(13), using $\mathbf{u}$ to denote the marker effects estimated as random, β for the focal marker effect estimated as fixed, and assuming independent marker effects, Henderson's mixed model equations (MME) take the form:

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{I}\lambda & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_{\setminus 1}^T\mathbf{X}_1 & \mathbf{X}_{\setminus 1}^T\mathbf{X}_1 & \mathbf{X}_{\setminus 1}^T\mathbf{X}_{\setminus 1} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ u_1 \\ \mathbf{u}_{\setminus 1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_{\setminus 1}^T\mathbf{y} \end{bmatrix} \tag{25}$$

where $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$. Subtracting the u_1 from the β equations gives $u_1 = 0$ and thus the MME reduce to:

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_{\setminus 1} \\ \mathbf{X}_1^T\mathbf{X}_{\setminus 1} & \mathbf{X}_{\setminus 1}^T\mathbf{X}_{\setminus 1} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ \mathbf{u}_{\setminus 1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_{\setminus 1}^T\mathbf{y} \end{bmatrix} \tag{26}$$

This has been derived previously [41], however there is an explicit assumption that the any estimation error of the random marker effect estimates go into the residual and does not influence the fixed estimate of the marker. For the random effect component, the equivalence with the ridge regression estimator of Eq.(15) is evident, as is the equivalence of Eq. (24) with Eq. (22) above. Thus an MLMi model returns "ridge regression" estimate of the marker effects, and as we show above ridge regression estimates are inflated when effect sizes are higher for high LD markers. It then follows that mixed model effect size estimates could be biased when effect sizes are higher for high LD markers.

Seen in this light, we can now explore the influence of multicollinearity on the BayesR dirac spike and slab model described above and compare it to that of a ridge regression. If we denote a measure of fit, such as the ridge loss function described above, being composed of $l(\beta)$ and a penalty function $pen_\lambda(\beta)$, then from a Bayesian perspective these correspond to the negative logarithms of the likelihood and the prior distribution, respectively. We can parameterize the BayesR dirac spike and slab model described above using the latent indicator of each SNP, j, $\gamma = (\gamma_j, \ldots, \gamma_p)^T$ with $\gamma_{j,l} = 0$ or 1, indicating whether or not the effect of SNP j follows a normal distribution with variance $\sigma_l^2 (l = 1, 2, 3, 4)$. Then $p(\gamma_{j,l} = 1|\pi_l) = \pi_l$ and the prior distribution of each SNP effect β_j conditional on the indicator $\gamma_{j,l}$ is

$$f\left(\beta_j|\gamma_{j,l}\right) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_l^2}}\exp(-\frac{\beta_j^2}{2\sigma_l^2}), & \text{if} \gamma_{j,l} = 1 \quad (l = 2, 3, 4) \\ \delta_0(\beta_j), & \text{if} \gamma_{j,l} = 0 \end{cases} \tag{27}$$

The joint distribution $p(\beta_j, \gamma_j)$ conditional on π_β is

$$\begin{aligned} f\left(\beta_j, \gamma_j|\pi_\beta, \sigma_\beta^2\right) &= \textstyle\prod_{l=1}^4 f\left(\beta_j|\gamma_{j,l}\right) f\left(\gamma_{j,l} = 1|\pi_l\right) \\ &= \left(\delta_0(\beta_j)\pi_1\right)^{\gamma_{j,1}} \textstyle\prod_{l=2}^4 \left(\frac{1}{\sqrt{2\pi\sigma_l^2}}\exp(-\frac{\beta_j^2}{2\sigma_l^2})\pi_l\right)^{\gamma_{j,l}} \end{aligned} \tag{28}$$

to simplify the following, we assume only a single normal distribution with $\pi_1 + \pi_2 = 1$ and we redefine the regression coefficient as $\beta_j = \gamma_j\alpha_j$ with $\alpha_j|\sigma_\beta^2 \sim N(0, \sigma_\beta^2)$. then:

$$\begin{aligned} f\left(\alpha_j, \gamma_j|\pi_\beta, \sigma_\beta^2\right) &= \left(\delta_0(\alpha_j)\pi_1\right)^{\gamma_{j,1}} \left(\frac{1}{\sqrt{2\pi\sigma_\beta^2}}\exp(-\frac{\alpha_j^2}{2\sigma_\beta^2})\pi_l\right)^{\gamma_{j,2}} \\ &= \pi_1^{\gamma_{j,1}}(1 - \pi_1)^{\gamma_{j,2}} \frac{1}{\sqrt{2\pi\sigma_\beta^2}}\exp(-\frac{\alpha_j^2}{2\sigma_\beta^2}) \end{aligned} \tag{29}$$

Now as above, if we define an active set of markers, $\mathbf{X}_{\gamma\neq 0}$, as those columns of $\mathbf{X}$ where $\beta_{\gamma\neq 0}$, with an active set of γ, and $||\gamma||_0 = \sum_{j=1}^{p}\gamma_j$ be its cardinality. The joint prior on the vector γ, α then factorizes across all the markers as

$$\begin{aligned} f\left(\alpha,\gamma|\pi_\beta,\sigma^2_\beta\right) &= \prod_{j=1}^{p} f\left(\alpha_j,\gamma_j|\pi_\beta,\sigma^2_\beta\right) \\ &= \pi_1^{||\gamma||_0}(1-\pi_1)^{p-||\gamma||_0}(2\pi\sigma^2_\beta)^{-\frac{p}{2}}\exp\left\{-\tfrac{1}{2\sigma^2_\beta}\sum_{j=1}^{p}\alpha_j^2\right\} \end{aligned} \tag{30}$$

as above we can express the likelihood in terms of γ, α as

$$f\left(y|\gamma,\alpha,\pi_\beta,\sigma_\epsilon\right) = (2\pi\sigma^2_\epsilon)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2_\epsilon}||y-\mathbf{X}_{\gamma\neq 0}\alpha_{\gamma\neq 0}||_2^2\right\} \tag{31}$$

and then under this reparamterisation the posterior is given as

$$\begin{aligned} f\left(\alpha,\gamma|\pi_\beta,\sigma^2_\beta,\sigma^2_\epsilon,y\right) &\propto f\left(\alpha,\gamma|\pi_\beta,\sigma^2_\beta\right) f\left(y|\gamma,\alpha,\pi_\beta,\sigma_\epsilon\right) \\ &\propto \exp\left\{\frac{1}{2\sigma^2_\epsilon}||y-\mathbf{X}_{\gamma\neq 0}\alpha_{\gamma\neq 0}||_2^2 - \frac{1}{2\sigma^2_\beta}||\alpha||_2^2 - log\left(\frac{1-\pi_1}{\pi_1}\right)||\gamma||_0\right\} \end{aligned} \tag{32}$$

The regularized maximum a posterior estimator is equivalent to minimising over γ, α the least squares objective function as

$$\min_{\gamma,\alpha}\|y-\mathbf{X}_{\gamma\neq 0}\alpha_{\gamma\neq 0}\|_2^2 + \lambda\|\alpha\|_2^2 + 2\sigma^2_\epsilon\log\left(\frac{1-\pi_1}{\pi_1}\right)\|\gamma\|_0 \tag{33}$$

In comparison to the ridge loss function described above, the first two terms are very similar and the third term imposes a sparsity constraint on the model. The term $\lambda\|\alpha\|_2^2$ has the same expectation as in Eq. (23) but with $\mathbf{X}$ replaced with $\mathbf{X}_{\gamma\neq 0}$. To give some insight into the influence of collinearity on $\mathbb{E}[\|\gamma\|]_0$ and on the active set, we explore a two SNP scenario.

In a single site updating scheme, the probability that the first marker enters the model is given by Eq. 12. We seek to derive the probability that the second marker enters the model conditional on the first marker being in the model. We consider a scenario where we observe our standardised outcome $\tilde{\mathbf{y_c}}$ and two correlated predictors $\mathbf{X}_1$ and $\mathbf{X}_2$. We assume that $\tilde{\mathbf{y_c}}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are scaled with zero mean and unit variance. We can then derive the partial least squares regression for $\tilde{\mathbf{y_c}}$ regressed on $\mathbf{X}_2$, adjusting for $\mathbf{X}_1$. If $\beta_{x_1,\tilde{y}} = \frac{\mathbf{X}_1^T\tilde{\mathbf{y}}}{\Sigma_{1,1}}$, with $\Sigma_{1,1} = \mathbf{X}_1^T\mathbf{X}_1 + \lambda_1\mathbf{I}$, then a residual vector $\boldsymbol{\epsilon}_{y_c,X_1} = \mathbf{y}_c - \beta_{x_1,\tilde{y_c}}\mathbf{X}_1$ is the vector left after backfitting $\beta_{x_1,\tilde{y_c}}$ and we define $\boldsymbol{\epsilon}_{X_1,X_2} = \mathbf{X}_2 - \rho_{X_1,X_2}\mathbf{X}_1$ as the additional information in X_2 left to fit $\beta_{x_2,\epsilon_{y_c,X_1}}$, with ρ_{X_1,X_2} the correlation of X_1 and X_2. The correlation between the two residuals $\boldsymbol{\epsilon}_{y_c,X_1}$ and $\boldsymbol{\epsilon}_{X_1,X_2}$ can be used to estimate $\beta_{x_2,\epsilon_{y_c,X_1}}$, since $\beta_{x_2,\epsilon_{y_c}} = \frac{N}{\Sigma_{1,l}}\rho_{\epsilon_{y_c,X_2}}$. The correlation is a ratio between a covariance and a variance as

$$\begin{aligned} Cov_{\boldsymbol{\epsilon}_{y_c,X_1},\boldsymbol{\epsilon}_{X_1,X_2}} &= \frac{1}{N}\sum\left(\mathbf{y}_c - \beta_{x_1,\tilde{y_c}}\mathbf{X}_1\right)\left(\mathbf{X}_2 - \rho_{X_1,X_2}\mathbf{X}_1\right) \\ &= \frac{1}{N}\sum\left(\mathbf{y}_c X_2 - \rho_{x_1,x_2}X_1\tilde{y_c} - \beta_{x_1,\tilde{y_c}}X_1X_2 + N\beta_{x_1,\tilde{y_c}}\rho_{X_1,X_2}\right) \\ &= \rho_{\epsilon_{y_c,X_2}} - \rho_{X_1,X_2}\beta_{x_1,\tilde{y_c}}\frac{\Sigma_{1,l}}{N} - \beta_{x_1,\tilde{y_c}}\rho_{X_1,X_2} + \beta_{x_1,\tilde{y_c}}\rho_{X_1,X_2} \\ &= \rho_{\epsilon_{y_c,X_2}} - \rho_{X_1,X_2}\beta_{x_1,\tilde{y_c}}\frac{\Sigma_{1,l}}{N} \\ &= \rho_{\epsilon_{y_c,X_2}} - \rho_{X_1,X_2}\frac{1}{N}X_1\tilde{y_c} \end{aligned} \tag{34}$$

The variance in the correlation denominator is $S^2_{\boldsymbol{\epsilon}_{X_1,X_2}} = 1-\rho^2_{X_1,X_2}$ which gives

$$\beta_{y_c,X_2|X_1} = \frac{N}{\Sigma_{2,l}} \times \frac{\rho_{\epsilon_{y_c,X_2}} - \rho_{X_1,X_2}\frac{1}{N}X_1\tilde{y_c}}{1-\rho^2_{X_1,X_2}} \tag{35}$$

Eq. 35 can then be used in Eq. 11 and Eq. 12 to determine the posterior inclusion probability of the second covariate conditional on the first covariate being in the model. From this, the expectation, $\mathbb{E}[\|\gamma\|_0]$ for

a two SNP scenario is then

$$\begin{aligned}\mathbb{E}[\|\gamma\|_0] &= p(l_1 = 1|\boldsymbol{\theta}, \mathbf{y}) + p(l_2 = 1|\boldsymbol{\theta}, \mathbf{y}) \\ &= \frac{1}{1 + \exp\left[log(\pi_0) - (-\frac{1}{2}[-\log(|\lambda\Sigma_{1,1}^{-1}|) - (\frac{\beta^2_{y_c,X_1}\Sigma_{1,1}}{\sigma^2_\beta})])\right]} \\ &+ \frac{1}{1 + \exp\left[log(\pi_0) - (-\frac{1}{2}[-\log(|\lambda\Sigma_{2,1}^{-1}|) - (\frac{\beta_{y_c,X_2|X_1}\Sigma_{2,1}}{\sigma^2_\beta})])\right]}\end{aligned} \tag{36}$$

With the dirac spike and slab and ridge regression estimators minimizing the same sum-of-squares, the key difference with the constrained estimation formulation of ridge regression is not in the explicit form of λ but in what is bounded the domain of acceptable values for α. For the BayesR estimator the domain is specified by a bound on the $\ell_0 norm$ of the regression parameter, while for its ridge counterpart the bound is applied to the squared $\ell_2 norm$ of β. Multicollinearity will reduce the likelihood of the second covariate entering the model as it's inclusion is dependent upon ρ_{X_1,X_2} the correlation among covariates and $\rho_{\epsilon_{y_c,X_2}}$ the correlation of the second marker and the residual vector after backfitting the first marker. This will limit the range of possible estimates to be lower than those obtained from ridge regression, reducing inflation of $\lambda\|\alpha\|_2^2$ under high collinearity, but not entirely removing it. Due to the sampling of markers from a series of normal distributions collinearity will still inflate $\lambda\|\alpha\|_2^2$, however, the degree to which this occurs will depend upon the number of correlated markers, the degree of correlation among them and the strength of the effects. Therefore, our aim here is not to derive a general solution predictive of all situations, merely it is to highlight that in order to make some inference as to the underlying distribution of genetic effects, it is required to extend the model as outlined in the following section.

Extending the model to account for collinearity and genomic annotation

We extend the BayesR model to a BayesRR-RC model as follows

$$\mathbf{y} = \mathbf{1}\mu + \sum_{\varphi=1}^{\Phi} \mathbf{X}_\varphi \boldsymbol{\beta}_\varphi + \epsilon \tag{37}$$

where there is a single intercept term $\mathbf{1}\mu$ and a single error term ϵ but now SNPs are allocated into groups $(\varphi_1, \ldots, \varphi_\Phi)$, each of which having it's own set of model parameters $\Theta_\varphi = \left\{\beta_\varphi, \pi_{\beta_\varphi}, \sigma^2_{G_\varphi}\right\}$. As such, each β_{φ_j} is distributed according to:

$$\beta_{\varphi_j} \sim \pi_{0_\varphi}\delta_0 + \pi_{1_\varphi}\mathcal{N}\left(0, \sigma^2_{1_\varphi}\right) + \pi_{2_\varphi}\mathcal{N}\left(0, \sigma^2_{2_\varphi}\right) + \ldots + \pi_{L_\varphi}\mathcal{N}\left(0, \sigma^2_{L_\varphi}\right) \tag{38}$$

where for each SNP marker group $\left\{\pi_{0_\varphi}, \pi_{1_\varphi}, \ldots, \pi_{L_\varphi}\right\}$ are the mixture proportions and $\left\{\sigma^2_{1_\varphi}, \sigma^{2_\varphi}_2, \ldots, \sigma^2_{L_\varphi}\right\}$ are the mixture-specific variances proportional to

$$\begin{bmatrix} \sigma^2_{1_\varphi} \\ \vdots \\ \sigma^2_{L_\varphi} \end{bmatrix} = \sigma^2_{\beta_\varphi} \begin{bmatrix} C_{1_\varphi} \\ \vdots \\ C_{L_\varphi} \end{bmatrix}$$

Thus the mixture proportions, variance explained by the SNP markers, and mixture constants are all unique and independent across SNP marker groups. This extends previous models (known as BayesRC [18] and BayesRS [19]), which have used additional mixtures for different SNP groups, but kept a single global variance component. Importantly, a single variance component with more mixtures serves only to change the amount of mass allocated at different sizes of the distribution, but does not alter the sizes of the effects themselves as there is still a single distribution. In contrast, the formulation presented here of having an independent variance parameter $\sigma^2_{\beta_\varphi}$ per group of markers, and independent mixture variance components, enables estimation of the amount of phenotypic variance attributable to the group-specific effects and enables differences in the distribution of effects among groups.

We can sketch the difference in the models by looking at the respective conditional posteriors, again, assuming a single component for simplification purposes. We have a BayesRC or BayesRS estimator by

assuming different groups of effects in eq. 32, which yields:

$$f\left(\alpha,\gamma|\pi_{\beta_\varphi},\sigma^2_\beta,\sigma^2_\epsilon,y\right)\propto\exp\left\{\frac{1}{2\sigma^2_\epsilon}||y-\mathbf{X}_{\gamma\neq0}\alpha_{\gamma\neq0}||^2_2-\frac{1}{2\sigma^2_\beta}||\alpha||^2_2-log\left(\frac{1-\pi_{1_\varphi}}{\pi_{1_\varphi}}\right)||\gamma_\varphi||_0\right\}\tag{39}$$

where π_{β_φ} are the group-specific mixture proportions and $||\gamma_\varphi||_0$ is the cardinality of the group. The corresponding MAP estimate would amount to adding extra penalisation on sparsity through the π_φ terms, while keeping the same level of shrinkage as the baseline BayesR.

In our model the conditional posterior is:

$$f\left(\alpha,\gamma|\pi_{\beta_\varphi},\sigma^2_{\beta_\varphi},\sigma^2_\epsilon,y\right)\propto\exp\left\{\frac{1}{2\sigma^2_\epsilon}||y-\mathbf{X}_{\gamma\neq0}\alpha_{\gamma\neq0}||^2_2-\frac{1}{2\sigma^2_{\beta_\varphi}}||\alpha||^2_2-log\left(\frac{1-\pi_{1_\varphi}}{\pi_{1_\varphi}}\right)||\gamma_\varphi||_0\right\}\tag{40}$$

now each marker has a group-specific shrinkage $\sigma^2_{\beta_\varphi}$, which translates to a specific λ_φ per group in the MAP estimate. This amounts to markers being shrunk according to the scale of the effects of their group, instead of the scale of all other markers. So instead of solving a single model selection and regularisation problem we are solving Φ model selection and regularisation problems, with shared information only through the residuals. If we subset by MAF and LD bins, the resulting groups of columns will have a correlation pattern similar to an exponential decay (LD decays with distance). If we take the whole genotype matrix, the pattern would be closer to a block diagonal matrix of correlations, in [17, 42] it is showed that the former case requires weaker conditions in order to recover the true vector β consistently than the latter. Although the sampling scheme was different, we have shown that a similar model with only two groups: genetic markers and epigenetic markers, is successful in identifying BMI and smoking epigenetic signatures [14].

A Gibbs sampling scheme for biobank size data

For "$p >> n$" regimes, such as in genomics, where the number of covariates is greater than the number of individuals, hierarchical models controlling assumptions over the sparsity of the model are typically proposed, with examples of sparsity-inducing priors like the "spike and slab" prior [15, 23], the Bayesian LASSO [43] and the Horseshoe [44] prior. There are efficient tools to perform Bayesian regression analysis "out-of-the-box" using MCMC and variational inference [45–47], but these methods are limited to problems with explanatory variables in the low thousands of observations. Recent results show that Gibbs samplers for the Horseshoe prior [29], or for the Bayesian LASSO [48], offer a competitive advantage when combined with approximation schemes for problems of high dimensionality (over 100,000 covariates). These latter methods exchange the inversion of the coefficient matrix, for a matrix multiplication, thus reducing complexity from cubic to almost quadratic on the number of variables. However, despite these good properties, scaling these approaches up to a factor of millions of variables remains prohibitive.

We now describe an effective algorithmic implementation of our BayesRR model that scales to millions of individuals, each genotyped at millions of genetic markers. We outline a Gibbs sampling algorithm that enables all sampling steps to utilize genetic data stored in mixed binary/sparse-index representation, reducing computational complexity of a single Gibbs step from $\mathcal{O}(n)$ to $\mathcal{O}(n_z)$, with n_z the number of non-zero genotypes. We then outline a Bulk Synchronous Parallel Gibbs sampling scheme implemented based on a hybrid MPI + OpenMP model, distributing data across MPI tasks over as many compute nodes as required to hold all the data in memory. Uniquely, this enables large-scale genomic data to be split up into smaller manageable segments, whilst still conducting the analysis in the same way, estimated the marker effects jointly.

Algorithm 1 provides a full overview of the sampling scheme of the model as it has been previously implemented. For each marker j, we must compute $\hat{\beta}_{j,l}$ to determine which mixture a marker belongs to, before then sampling $\hat{\beta}_{j,l}$ given the mixture group assigned. This quantity depends on the dot product $\mathbf{X}_j^T\mathbf{y}_c$, with $\mathbf{y}_c$ the centred phenotype. If we keep in memory the vector of residuals $\epsilon=\mathbf{y_c}-\mathbf{X}\boldsymbol{\beta}_{\gamma\neq\mathbf{0}}$, then we can compute efficiently $\mathbf{y_c}-\mathbf{X}_{\setminus\mathbf{j}}\boldsymbol{\beta}_{\gamma\neq\mathbf{0}_{\setminus\mathbf{j}}}$ by the update $\mathbf{y_c}-\mathbf{X}_{\setminus\mathbf{j}}\boldsymbol{\beta}_{\gamma\neq\mathbf{0}_{\setminus\mathbf{j}}}=\tilde{\epsilon}+\mathbf{X_j}\beta_\mathbf{j}$, thus sampling from the joint distribution with a complexity $\mathcal{O}(p)$. The most expensive operation in Algorithm 1 is computing the numerator in step 9: $\mathbf{X}_j^T\left(\tilde{\epsilon}+\mathbf{X}_j\beta_j^{old}\right)$. As the column vector $\mathbf{X}_j$ contains the centered and scaled genotypes, step 9 involves one sum of two dense vectors and a dot product of two dense vectors. However, if we store in memory the mean, μ_j, and standard deviation σ_j of each column of the genotype matrix, we can express the numerator in step 9 with these quantities and the j-th column of the original genotype matrix $\mathbf{G}$ as (with

Algorithm 1: Serial Algorithm for sampling over the posterior distribution $p(\mu, \beta, \epsilon, \sigma_\epsilon, \theta)$. $\mathbf{X}_{marker_j}$ represents column of $\mathbf{X}$ corresponding to the column j of the vector $marker$. Given that $marker$ is shuffled before sampling the effects, this is equivalent to permuting the order of the effects to be sampled.

Data: Coefficient matrix $\mathbf{X}$, measurement vector $\mathbf{y}$, prior hyperparameters ν_0, s_0^2, iterations I
Result: mean μ, effects vector $\boldsymbol{\beta}$, residual vector ϵ, residual variance σ_ϵ^2 and variance contributed by the marker effects, σ_G^2

Initialize $\boldsymbol{\beta}, \mu, \sigma_\epsilon^2, \sigma_G^2, \pi_\phi$;
$effects = 1, \ldots, p$;
$\epsilon = \mathbf{y} - \mu$;
for $i \leftarrow 1$ **to** I **do**
  Sample μ;
  Shuffle $(effects)$;
  for $j \leftarrow 1$ **to** p **do**
    $\beta_j^{old} = \beta_j$;
    $\hat{\beta}_{j,l} = \frac{\mathbf{X}_j^T(\tilde{\epsilon} + \mathbf{X}_j \beta_j^{old})}{\Sigma_{j,l}}$;
    Determine mixture component and sample the new value β_j;
    $\epsilon^{new} = \epsilon + (\beta_j^{old} - \beta_j)\mathbf{X}_j$;
  Sample σ_ϵ^2;
  Sample σ_G^2;

$\sigma_j^2 = (\mathbf{G}_j - \mu_j \mathbf{1})^T (\mathbf{G}_j - \mu_j \mathbf{1}) / (n-1)$ by definition):

$$\begin{aligned} num &= \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j}\left(\epsilon + \beta_j^{old}\frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j}\right) \\ &= \frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j}\epsilon + \beta_j^{old}\frac{(\mathbf{G}_j - \mu_j \mathbf{1})^T}{\sigma_j}\frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j} \\ &= \frac{\mathbf{G}_j^T}{\sigma_j}\epsilon - \frac{\mu_j}{\sigma_j}\sum_{i=1}^{n}\epsilon + \beta_j^{old}(n-1) \end{aligned} \tag{41}$$

and we can do the same for the ϵ update:

$$\epsilon_{new} = \epsilon + (\beta_j^{old} - \beta_j)\frac{(\mathbf{G}_j - \mu_j \mathbf{1})}{\sigma_j} = \epsilon + \frac{(\beta_j^{old} - \beta_j)}{\sigma_j}(\mathbf{G}_j - \mu_j \mathbf{1}) \tag{42}$$

for which we only have to compute the difference of a sparse vector and a dense vector, and the sum of two dense vectors. Finally, to avoid computing $\sum_{i=1}^{n}\epsilon_{new}$ for each marker, we assign a variable to this quantity and update it after each ϵ update as follows (with $\mu_j = \sum_{i=1}^{n}\mathbf{G}_{i,j}/n$ by definition):

$$\sum_{i=1}^{n}\epsilon_{new} = \sum_{i=1}^{n}\epsilon + \frac{(\beta_j^{old} - \beta_j)}{\sigma_j}\left(\sum_{i=1}^{n}\mathbf{G}_{i,j} - n\mu_j\right) = \sum_{i=1}^{n}\epsilon \tag{43}$$

meaning that the sum of ϵ elements is constant during the algorithm execution (as expected as all involved vectors are zero-mean). Therefore, the only quantity to be computed per run (apart from the ϵ update) is the dot product $\frac{\mathbf{G}_j^T}{\sigma_j}\epsilon$ which can also be reduced, as the elements of $\mathbf{G}_j$ can only be either $\{0, 1, 2\}$ with sequence data or hard-coded genotype. We call $\mathcal{I}_1$ the indicator function such that $\epsilon\mathcal{I}_1 = \begin{cases}\epsilon_j & x_j = 1 \\ 0 & else\end{cases}$ and similarly $\epsilon\mathcal{I}_2 = \begin{cases}\epsilon_j & x_j = 2 \\ 0 & else\end{cases}$ which then gives the dot product as $\frac{\mathbf{G}_j^T}{\sigma_j}\epsilon = \frac{\sum\epsilon\mathcal{I}_1 + 2\sum\epsilon\mathcal{I}_1}{\sigma_j}$ meaning that multiple $\mathcal{O}(n)$ multiplications are now $\mathcal{O}(n_z)$ sums, and also that instead of storing in memory a sparse matrix of elements plus its indexes, we just need to store three ragged arrays of indexes, one for the "1" elements, a second one for the "2" elements, and a third one for the "M"issing elements. Those arrays contain information for all markers processed by a MPI task and are of unsigned integer type (32 bits). They store indices of the 1, 2 and M elements within the marker (i.e. ranging from 0 to $N-1$). It corresponds to the smallest integer type

that allows us to scale to hundreds of thousands or millions individuals. On top of those 3 ragged arrays there are two meta-data arrays for each element type which provide the starts and lengths of the 1, 2 and M elements for each marker in the ragged arrays. They are loaded in memory from reading sparse data files stemming from the conversion of the original Plink .bed file and accessed in parallel by the tasks with MPI I/O.

Even though the sparse representation is optimal in number of operations, performance may vary depending on hardware as a vectorised dot product may be faster than sparse dot product. Spatially, the sparse representation is optimal as long as the columns are sparse. In genotype data, even though the expected number of non-zeros per column is given by the average MAF ($\sim 20\%$ in the UK Biobank data described below), the distribution is long tailed (Figure S4). These columns at the tail of the distribution can dominate the total size of the data structure in memory. Encoding a single column has a constant size of $N \times 2$ bits in plink's .bed file format (referred from now on as binary format), while in sparse representation a column has varying size of $n_z \times 32$ bits. If we encode the columns with less than 6% of non-zeros as sparse and the rest in the original binary format, we can have a total memory occupancy of 60% the size of the original genotype matrix in Plink bed format. In (Figure S4) we represent on panel (b) the distribution of the proportion non-zeros per column of a genotype matrix for $\sim 4 \times 10^5$ individuals and $\sim 1.5 \times 10^7$ SNPs, solid line representing the mean of the distribution and slashed line the median. In panel (c) we show the total size of the data in memory as a function of the threshold used to split between binary and sparse format, in purple we see how the binary representations dominates the total size up until the mean of the distribution, after which, the size of the sparse data structure starts to dominate and ends up being around four times bigger than the original .bed file size(dotted horizontal line). We found the optimal threshold to be around 0.064(6.4%, Figure S4).

Finally, we implement a vectorized dot product for genotype data stored in the raw binary format based on a couple of look-up tables, by writing the dot product as:

$$\begin{aligned} \frac{\left(\mathbf{G}_j - \mu_j \mathbf{1}\right)^T}{\sigma_j}\epsilon &= \sum_i \frac{\psi_{i,j}\epsilon_i}{\sigma_j} \\ &= \frac{1}{\sigma_j}\left(\sum_i a_i\epsilon_i - \mu_j \sum_i b_i\epsilon_i\right) \end{aligned} \tag{44}$$

with coefficients a_i and b_i being 0.0, 1.0 or 2.0 depending on the value of $\mathbf{G}_{i,j}$ and following Table 2.

$\mathbf{G}_{i,j}$	0	1	2	NA
2-bit	11	10	00	01
a_i	0.0	1.0	2.0	0.0
b_i	1.0	1.0	1.0	0.0
$\psi_{i,j}$	$0.0 - 1.0\mu_j$	$1.0 - 1.0\mu_j$	$2.0 - 1.0\mu_j$	$0.0 - 0.0\mu_j$

Table 2. a and b coefficient values used for building up the two look-up tables needed for the vectorization of the dot product computation when processing binary data.

As 1 byte of plink's .bed can contain $4^4 = 256$ different combinations of information for 4 individuals, we can setup two lookup tables with 256×4 entries each that will give for any byte the corresponding 4 a_i and b_i coefficients, hence allowing for vectorisation of Eq. 44 by performing $a_i\epsilon_i$ and $b_i\epsilon_i$ and accumulating them for 4 individuals at once. Additionally, we use OpenMP to parallelize the loop over the marker's bytes. This greatly extends previously proposed sparse residual updating schemes and also facilitates the synchronous, fully parallel bulk-synchronous Gibbs sampling scheme that we describe in the next section below.

Bulk-synchronous parallel Hogwild Gibbs sampling with sparse data

Bulk-synchronous parallel Hogwild Gibbs sampling [49] assigns block of columns from $\mathbf{X}$ to workers that then sample from $f\left(\beta_j|\boldsymbol{\beta}_{\setminus j}, \mathbf{y}\right)$ for each of the columns in their block. Workers can communicate between each other exchanging the current values of the variables they are sampling, or the whole state of variables for workers in particular. If we perform global synchronisation steps the algorithm is called Bulk-synchronous parallel Hogwild (BSP), if on the other hand, workers exchange messages without a global synchronisation, the algorithm is called Asynchronous parallel Hogwild (ASP) [50].

Algorithm 2: Hogwild Gibbs with '$\Delta\epsilon$-exchange'.

components : Define K parallel workers
Define global variables μ, β, ϵ,π, σ_g^2, σ_e^2;
Initialize variables;
for $i \leftarrow 1$ **to** I **do**
    Update μ;
    Update β in parallel using **DEpsX**(K);
    Update hyperparameters $\pi, \sigma_g^2, \sigma_e^2$;

We propose Algorithm 2, which is a modification of a BSP algorithm where we sample the individual coefficients in parallel conditioned on the hyperparameters. We assign workers (MPI tasks) subsets of coefficients to sample, and each worker performs local Gibbs steps until a global synchronisation is triggered. This global synchronisation happens many times in each iteration, during the phase in which we sample the individual coefficients β_j. For this algorithm, we developed a synchronisation scheme called '$\Delta\epsilon$-exchange'as outlined in Algorithm 3. In this scheme each individual worker is assigned a block of columns from $\mathbf{X}$ and is in charge of sampling from $f\left(\beta_j|\boldsymbol{\beta}_{\setminus j}, \mathbf{y}\right)$ for each of the columns in its block. We add an additional parameter for the synchronisation rate Ω. After Ω columns have been sampled in all workers (around 5-10 in practice to avoid divergence occurring), a synchronisation move is executed.

The purpose of the synchronisation move is to update all of the workers' state based on the coefficients sampled from $t = 1$ until $t = \Omega$ in all workers. The sufficient statistic for this state is contained in the residual vector ϵ. Thus from $t = 1$ until $t = \omega$ each worker computes $f\left(\beta_j|\epsilon_{t=1}\right)$ and keeps track of its local change in ϵ which we denote $\Delta\epsilon = \sum_1^{\Omega} \mathbf{X}_\omega \beta_\omega$ for ω in the set of indexes for the current batch of variables in the workers list of variables. For the synchronisation step, we use the MPI_Allreduce collective, meaning that each task will receive the sum of locally accumulated $\Delta\epsilon$ from all tasks to update its $\epsilon_{t=1} = \sum^w \Delta\epsilon_w$ for $w = (1...W)$ workers. With the new $\epsilon_{t=1}$, the worker proceeds to sample the next Ω-sized batch of columns from its set of columns. This synchronisation scheme allows workers to exchange state information in compact form, as the total size of memory occupied in total by the messages is $\mathcal{O}\left(NW\right)$.

Algorithm 3: '$\Delta\epsilon$-exchange'for synchronising changes in backfitted residuals in our BSP Gibbs sampling algorithm.

DEpsX (K)
    components : Set of K workers, each one β_k, Set of K messages, each one $\Delta\epsilon_K$, K sets of $\sim \frac{p}{K}$ columns, each set of columns assigned to a worker.
    foreach *worker* β_k **do**
        $\epsilon_k = \epsilon$;
        $\Delta\epsilon_k = 0$;
        foreach *column* i *in a subset of size* Ω *of the columns assigned to* β_k **do**
            β_j^{old}=β_i;
            draw β_i from $f\left(\beta_i \mid \epsilon, \sigma_\epsilon^2, \sigma_G^2, \pi\right)$;
            $\Delta\epsilon_k$=$\Delta\epsilon_k$-X_i $\left(\beta_i - \beta_j^{old}\right)$;
        Wait until all workers are finished processing their Ω sets;
        $\epsilon = \epsilon + \sum_k \Delta\epsilon_k$;

Previous results point to BSP Gibbs sampling for a multivariate Gaussian converging if the covariance matrix is strictly diagonal-dominant [50] with zero covariance of the markers split across workers. The risk for genomic data, is that two markers in LD get updated at the same time in parallel, double counting their effects, and leading to ϵ being mis-estimated after a synchronization has occurred. Suppose we have one fixed causal marker and two other markers i and j that are assigned to different MPI tasks. Suppose that the Pearson correlation between the causal marker and marker i or j is ρ_i and ρ_j, respectively. Finally, let ρ denote the correlation between the markers i and j. For simplicity in this example suppose that the inclusion probability of the causal marker is q and we make an assumption that the inclusion probability of the marker i is then $P(\beta_i \neq 0) = q\rho_i$ and for marker j it is $P(\beta_j \neq 0) = q\rho_j$, that means that the inclusion probability is proportional to the correlation between causal and other markers. In reality, the effect size estimate is actually proportional to the causal effect: $\hat{\beta}_i = \rho_i \beta_{causal}$ and the function between posterior

inclusion probability and causal effect size $q(\beta_{causal})$ is not linear for $\beta_{causal} \geq 0$ as described in Eq.(12) and thus we cannot assume that $P(\beta_i \neq 0) = q\rho_i$ in practice. In the case of parallelising the markers between two tasks we are interested in the probability that two markers from different tasks will absorb the effect of a same causal variant. Thus, we are interested in the probability $P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U)$, where U is the set of markers that are updated simultaneously in two different tasks. Thus, we can write:

$$P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U) = P(\beta_i \neq 0)P(\beta_j \neq 0) = q^2\rho_i\rho_j.$$

We see that the probability of making a mistake is dependant on the product $\rho_i\rho_j$. The correlation matrix R of the three markers

$$R = \begin{pmatrix} 1 & \rho_i & \rho_j \\ \rho_i & 1 & \rho \\ \rho_j & \rho & 1 \end{pmatrix}$$

has to be positive semi-definite and thus we can examine what are the possible values for the product $\rho_i\rho_j$ given that we know ρ. Note that the value of ρ can be controlled by providing some blocking mechanism that would assign SNPs to the tasks so that the correlation for the markers from different tasks would be limited to ρ and this is what we advocate here, placing contiguous blocks of markers into different tasks, so as to maximise the LD within a block (MPI task), but minimise the LD across blocks. The maximum possible values for the product follow a linear function that depends on ρ as

$$\max_{\rho_i, \rho_j, \rho = \tilde{\rho}} = 0.5 + 0.5\tilde{\rho}.$$

To get better estimates for the constraints for the product $\rho_i\rho_j$ then we need to make further assumptions about the distribution of ρ_i or ρ_j. Therefore, we can say that $P(\beta_i \neq 0, \beta_j \neq 0 | i, j \in U) \leq q^2(0.5 + 0.5\rho)$. This result and inequality only holds per sampled pair (i, j). We then multiply this result with the probability of sampling the pair (i, j) that both have correlations $\rho_i, \rho_j > 0$. Denoting a set of markers that have a positive correlation with one specific causal marker as the causal radius C, The probability of sampling any pair (i, j) is

$$P(i, j \in U) = \frac{1}{T^2},$$

where T is the number of markers per one task. The probability of pair (i, j) belonging to C is $P(i, j \in C) = c(<< 1)$, some reasonable values could be proposed or estimated for this (for example, $c = (\frac{\#(markers-in-LD)}{2T})^2$). Combining the results together we get that the probability of making a mistake at one update of a pair (i, j):

$$P(\beta_i \neq 0, \beta_j \neq 0) = P(\beta_i \neq 0, \beta_j \neq 0 | (i, j) \in U; (i, j) \in C)P((i, j) \in U)P((i, j) \in C) = P(\beta_i \neq 0, \beta_j \neq 0 | (i, j) \in U)\frac{c}{T^2} \leq q^2(0.5 + 0.5\rho)\frac{c}{T^2}.$$

This result goes for one fixed causal marker and it also represents the expected number of mistakes per sampled pair (i, j) for one causal marker. If we want to find the expected number of mistakes per sampled pair, we should sum across the P causal markers:

$$Errors \leq \sum_{i=1}^{P} q_i^2(0.5 + 0.5\rho)\frac{c}{T^2} = (0.5 + 0.5\rho)\frac{c}{T^2}\sum_{i=1}^{P} q_i^2 \leq (0.5 + 0.5\rho)\frac{cP}{T^2}$$

To provide some intuition, we can think of an extreme scenario and assume that there are 100,000 variants in the SNP marker data that would enter the model as they are in LD with underlying causal variants, that each of these variants has posterior inclusion probability of 1, and that for each variant there are two blocks with 30,000 markers in total of which 100 markers have LD = 1 with the causal variant, and that both blocks contain 30,000 markers. Placing these values into what we derive above and sampling over 10,000 iterations leads to probability of an error ~ 0.1 throughout the sampling for this extreme example. Having derived a stable highly parallel Gibbs sampling algorithm for large-scale genomics data, we then performed exhaustive empirical validation of our algorithm in simulation study as described below.

Implementation and processing setup

We implement algorithms 2 and 3 in C++ as a pure CPU MPI + OpenMP hybrid solution. All data structures were properly aligned in memory to assist vectorization and assembly code was examined to ensure that the code was properly vectorized where expected. We utilize the scientific libraries eigen and boost (see Code Availability) and we profiled and benchmarked the code with Intel performance analysis tools such as Advisor and Ampflier. Current implementation requires to be compiled with Intel compiler on an architecture supporting at least AVX2 although support for AVX512 is recommended for performance. The code is freely available from our Github repository and we also provide a statically compiled binary (see Code Availability). All the results were generated on the cluster Helvetios from EPFL (see Code Availability) using 10 compute nodes and setting 8 MPI tasks per node and dedicating 4 (physical) cores to each task. 10 is the minimal number of nodes that was required to hold all the data in memory in its mixed-representation.

Simulation study

Our theory suggests that there will be increased variance of the regression coefficient estimates and, as a result, an inflated estimate of the phenotypic variance attributable to SNP markers under high multicollinearity for both mixed linear model approaches and a Dirac spike and slab mixture model. To demonstrate this visually, we conducted a simulation study where for each of 50 replicates, we simulated 50 independent genomic regions, each containing two SNP markers. In each simulation replicate, we simulated values for 5,000 individuals at each of the 50 SNP marker pairs, by first simulating from a standard multivariate normal distribution with correlation set to either 0 or 0.99. From this, we obtained the integral from $-\inf$ to q of the probability density function, where q is the z-score of the values obtained for each individual from the multivariate normal. From these integrals, we then made two draws from the inverse of the cumulative density function of the binomial distribution to obtain the marker value for each individual, with frequency 0.3. This gave marker values (0, 1, or 2), with the pairs of SNPs having either all LD = 0, or all LD = 0.99. For each of the 50 pairs of SNPs, we assigned effect size 0 to the first marker and 0.1 to the second marker. We then scaled the SNP markers to zero mean and unit variance and multiplied the markers by the effect sizes to obtain the genetic values for the 5,000 individuals, with variance 0.5. We then simulated the environmental component of the phenotype from a normal distribution with zero mean and variance 0.5 and then created a phenotype as the sum of the genetic values and the environmental values, with zero mean and unit variance.

We then analysed these 50 data sets using different methods of single-marker OLS regression (OLS), mixed-linear model association (MLMA), ridge regression (Ridge), and a Dirac spike and slab mixture of regressions model (BayesR), all of which are described above. For the frequentist approaches, we directly solved the estimation equations, scaling the SNP markers to have zero mean and unit variance. For BayesR we sampled the effects for 5000 iterations, with burn-in period of 2000 iterations to obtain the posterior mean effect sizes, again scaling the SNP markers to zero mean and unit variance. We repeated these analyses many times, each time fixing the estimated phenotypic variance attributable to the markers σ^2_G to be a different value. We selected (2, 1, 0.5, 0.1, and 0.01) and fixed the residual variance σ^2_ϵ to be 0.5, to give different lambda values $\lambda = \frac{\sigma^2_\epsilon}{\sigma^2_G}$, giving $\lambda = 0.25, 0.5, 1, 5,$ and 50. Our aim here was to explore the pattern of effect sizes that we obtain under these λ values. So first, we plotted the effect sizes obtained for each of the 50 SNP pairs obtained across the 50 simulation replicates in Figure S1, to show the differences in the variance of the estimates obtained across approaches when the pairs of SNP markers were orthogonal (LD=0), or collinear (LD=0.99), under different lambda values. Second, we then plot the distribution of the sum of the squared regression coefficients in Figure 1d across approaches, when the pairs of SNP markers were orthogonal (LD=0), or collinear (LD=0.99), under different lambda values, where the expectation is 0.5 (sum of the 50 squared 0.1 SD effect sizes). This simulation confirmed, that regression coefficients under all approaches have higher variance under multicollinearity, resulting in inflation of the sum of the squared coefficient estimates for all approaches when the variation attributable to SNP markers is overestimated, resulting in a reduction in the lambda values.

We then further explored the performance of the MLMA and BayesR models under multicollnearity to (i) better understand the interplay between the fixed GLS estimate obtained and the random marker effects, and (ii) to better understand how the prior of the BayesR model changes with lambda and how this constrains the inclusion probabilities of correlated markers. We first examined the influence of varying lambda and varying the collinearity of markers on the variation of the effect size estimates obtained from the Henderson's mixed model equations, where one focal marker is estimated as fixed, and a further five markers are estimated as random, with LD between the markers estimated as fixed and random. To do this, we simulated five markers in the same manner as described above that were either (i) entirely orthogonal with LD = 0, or (ii) had

LD = 0.99 among the first three markers, with the final two markers having LD = 0 with all others. We assigned effect sizes to the five markers as $beta = (0.25, 0, 0, 0.25, 0.25)$, multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values (0.1875) to give a phenotype with zero mean and unit variance. We directly solved the Henderson's mixed model equations, fixing the lambda value at different levels (the appropriate lambda from theory assuming orthogonal covariate would $= (1 - 0.1875)/0.1875 = 4.333$). We find that even with high shrinkage, a lambda value of almost 20 times greater than the theoretical orthogonal expectation is required to produce effect sizes under collinearity, with similar variance to those obtained under orthogonality (Figure S1).

For BayesR, we first explored the density of the posterior distribution by simulating draws from the prior as we change the variance attributable to the SNP markers. Figure S1c shows these densities, revealing how the prior becomes strongly centred on zero and almost exponentially distributed as the variance becomes small. This is in contrast to the almost flat prior observed with high variance, which will do little to constrain effect size estimates toward zero. We then conducted 1000 simulation replicates of paired SNP markers for 10 different scenarios of variance attributable to the SNP markers of 0.01, 0.05, 0.1, 0.2, and 0.5, for pairs of SNPs with correlation of either 0 or 0.99. For each of these 10,000 data sets we simulate a pair of SNPs for 5000 individuals, assuming error variance of 0.5, effect size for the first marker of 0.01 SD and then we simulated a sequence of 1000 different effect sizes from -0.05 to 0.05. Of these 10 million phenotypes and pairs of SNPs obtained, we then determine the posterior inclusion probability of the second marker, given that the first marker is in the model, with the effect size correctly estimated as 0.01, from the BayesR model derivations presented above. The lines presented in Figure S1d go through the mean posterior inclusion probability of the second SNP marker across the 1000 simulation replicates, for each of the 1000 different effect sizes from -0.05 to 0.05 for marker 2, with a different colour for each scenario of the variance attributable to the SNP markers. The plot shows a reduction in the posterior inclusion probability of the second SNP marker as the variance attributable to the SNP markers decreases under multicollinearity. Thus, if the hyperparameter estimates of the variance contributed by markers is kept small, by having different hyperparameters for different groups of markers, then the BayesR model acts to constrain the inclusion of any additional correlated markers in the model.

Having confirmed our theory, we then conducted a further simulation study to replicate these observations using real genomic data. We randomly selected 50,000 individuals from the UK Biobank study (see below) and used the imputed SNP data from chromosome 22 as supplied in the data release. We simulated phenotypes under contrasting generative models:

- We chose markers of high LD with other SNPs to be the causal variants and we assigned effects proportional to the LD score of those markers and their minor allele frequency. To do this, we first grouped the SNPs using the clumping procedure in Plink (see Code Availability) based on 1 - MAF, selecting the highest frequency variants and removing any variants with LD < 0.01, to obtain 4988 independent SNPs. For these 4988 SNPs we calculated the LD score of the markers. We then assigned effect sizes to these selected SNPs, drawing them from a single normal distribution with variance $\sim \text{LD\_score}^{1}\text{MAF}^{-1}$. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

- We then took the same 4988 SNPs but assigned effect sizes to the markers at random from a normal distribution with zero mean and variance 0.5/4988. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

- We then sampled randomly 4988 evenly spaced markers as causal variants, but assigned effect sizes proportional to the LD score and minor allele frequency of the markers as described above. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.6, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

- Finally, we then sampled randomly 4988 evenly spaced markers as causal variants and randomly assigned the effect sizes from a normal distribution with zero mean and variance 0.5/4988. We multiplied these effect sizes by the simulated marker values scaled to zero mean and unit variance to create the genetic values with variance 0.5, and then added an environmental component simulated from a normal distribution with mean zero and variance 1 minus the variance of the genetic values to give a phenotype with zero mean and unit variance.

We analysed 50 simulation replicates of each of the four scenarios with our BayesR software, our BayesRR software with 20 MAF-LD groups (deciles of MAF, each split into two groups based on median LD score within each MAF decile), and a MLMA implemented in software GCTA (see Code Availability). For the Bayesian methods we ran three chains with different starting values for each of the 200 simulation replicates for 3000 iterations, removing the first 1500 iterations as burn-in and taking the posterior mean across the three chains. In Figure 1a we plot the distribution of the posterior mean for BayesR and BayesRR, and the MLMA point estimates, of the proportion of variance attributable to the SNP markers minus the true simulated value obtained across the 50 simulation replicates for each of the four scenarios, showing inflation of the MLMA estimates when selecting high LD variants, and inflation of the BayesR estimates with high LD and random effect size estimates. In contrast, estimates obtained from BayesRR were unbiased across all scenarios. By simulating an effect size MAF relationship $\sim$ LD_score^{1}MAF^{-1}, we are assigning the smallest absolute effect size values to the most common SNPs, which appears to limit the inflation of the estimates for BayesR, when selecting high LD SNPs as causal variants (Figure 1a). We then examined the effect size estimates obtained from these three approaches across the MAF spectrum under the second scenario of high LD causal variant selection, but random effect size allocation, to show using z-scores calculated as the estimated effects minus the simulated effects, divided by the SD of the simulated effects. We find overestimation of common variant effect sizes under BayesR, and dramatic inflation of effect size estimates under MLMA showing poor recovery of the underlying effect size distribution (Figure 1b). Grouping effects by MAF and LD in a BayesRR model resolved this overestimation issue (Figure 1b).

We then explore the ability of the model to recover annotation-specific variation using the same set of 50,000 randomly selected UK Biobank individuals and imputed genotype data for chromosome 22 grouped by chromatin state annotations (15-state ChromHMM model) from the epigenome of primary mononuclear cells from peripheral blood (E062) of the Epigenome Roadmap Project [20]. We simulated the genetic architecture as follows :

- We first mapped SNPs to active and inactive chromatin states from the mnemonic bed files for E062 (see Code availability). 37,187 SNPs mapped to active chromatin states including transcription start site (TSS) and their flanking regions, genic and other enhancers, untranslated transcribed regions (UTR) and actively transcribed regions and zinc finger genes states. 27,224 SNPs mapped to inactive states including heterochromatin, bivalent/poised TSS and their flanking regions, bivalent enhancers and repressed polycomb states. The remaining 47,018 SNPs were grouped and labelled as Other SNPs (Figure 1d).

- To simulate enrichment in both chromatin states, we randomly sampled 2000 SNPs as causal variants from variants mapped to active chromatin states and another 2000 SNPs from variants mapped to inactive chromatin states. We then assigned effect sizes to these 4000 selected SNPs, drawing them from a normal distribution with zero mean and variance 0.35/2000 for active states and 0.15/2000 for inactive states.

- We multiplied annotation-specific effect sizes by the simulated marker values scaled to zero mean and unit variance to create the annotation-specific genetic values with variance 0.35 for active states, 0.15 for inactive states and 0 for other SNPs. We finally added an environmental component simulated from a normal distribution with mean zero and variance 1 minus 0.5 (the sum of the genetic values) to give a phenotype with zero mean and unit variance.

We analyzed 20 simulation replicates with our BayesRR software specfiying annotations (active states, inactive states and other SNPs) with 2 LD groups based on median LD score within each annotation. We compared our software to boltREML [22] and RHEmc [51] both multi-variance component methods that also use individual-level data but provide single heritability estimates per genetic component. For BayesRR we ran three chains with different starting values for each of the 20 simulations replicates for 3000 iterations, removing the first 1000 iterations as burn-in and taking the posterior mean across the three chains. We then

performed the same analysis but randomly assigning SNPs to each annotation resulting in mis-specification of the underlying genetic architecture. In Figure 1d, we plot the estimated sum of the squared regression coefficients that is evenly split across the three annotations when misspecifying the underlying genetic architecture (labelled : Misspecification of groups) and shows enrichment when we properly assign SNPs to annotation (labelled : Multiple group enrichment). We find that BayesRR performs as boltREML and RHEmc, with RHEmc estimates showing higher variability.

We also further examined the ability of BayesRR to recover effect sizes compared to our BayesR software by comparing 10 simulations of 5 chains with different starting values where each simulation has two groups in high LD with an interdigitated structure where one in two SNPs is assigned to group 1 (Figure S2). We then simulated phenotypes as previously described, randomly selecting 1000 causal variants in group 1 only, using 20,000 randomly selected UK Biobank individuals and imputed genotype data for chromosome 2 (with MAF > 0.05). In Figure S2, we compare the proportion of markers entering the model in group 1 and group 2 at different posterior inclusion probability thresholds. Annotation-specific estimates for BayesR are calculated post-analysis for each group. We also compare the correlation of estimated genetic values with the truth when using BayesRR and BayesR. For this, we conducted estimation of marker effects in an independent data set to compare prediction accuracy. We simulated 10 new phenotypes and computed the genetic value $\hat{g} = X\hat{\beta}$ where X is the genotype matrix and $\hat{\beta}$ is a vector of estimated marker effects for each individual. Figure S2 shows we improve the power of BayesR to recover effect sizes and infer underlying genetic architectures.

Next, we then explored the influence of increasing parallelism in our algorithm. We used the simulated data described above for the randomly sampled 50,000 UK Biobank individuals with imputed genotype data for chromosome 22, where we sampled randomly 4988 evenly spaced markers as causal variants and randomly assigned the effect sizes from a normal distribution with zero mean and variance $0.6/4988$ (the fourth scenario). For each of the 50 simulation replicates, we compared the three chains obtained by running the BayesRR model (with 20 MAF-LD groups) in serial, with a single MPI task and synchronisation rate of 1 (residual updating after sampling each SNP), to three chains obtained by increasing the number of MPI tasks to 4 and then to 8, with synchronisation rates of 10 and 20 sampling steps before residual updating. For each simulation, we ran three chains of our BayesRR model with different starting values for 3000 iterations. Like with all MCMC chains of regression models, convergence and sampling properties will be problem specific and dependent upon the LD of the markers, LD among the causal variants, the phenotypic variation attributable to the SNP markers across the MAF and LD spectrum, the study sample size, the degree of data parallelism per total marker number, and the synchronisation rate. Thus, the aim here is to simply show a series of diagnostic tests that can be utilized to explore the properties of the posterior to highlight how the different metrics can be used to identify convergence issues. We use the distribution, across simulations, of the proportion of effective samples obtained for the hyperparameter estimate of the proportion of phenotypic variance attributable to the markers of each group. This shows that for all ranges of parallelism, we achieve more effective samples for low MAF and low LD variants. As high MAF SNPs are interchangeable in the model to a large degree, their entry and exit from the model is correlated across iterations, and thus this is entirely expected and is actually a consequence of the model mixing. With high synchronisation rates, where many marker updates occur before residual updating by message passing a reduction in effective sample sizes occurs. We also use the distribution of the Gelman-Rubin test statistic for the three chains, a general metric to monitor convergence that compares within- and among-chain variance, as the number of iterations increases. Finally, a Geweke statistic value can be used to test the equality of the means of the first and last part of the Markov chains. We present the results of this simulation in Figure S3 also including the distribution of z-scores of the posterior distribution of the phenotypic variance attributable to the markers for each MAF-LD group from the simulated values, which show stability of the estimates obtained with increasing data parallelism (tasks), but that a very high synchronisation rate with high parallelism can lead to poor convergence rates, meaning that the chains would have to be run for longer (Figure S3).

Finally, we investigated the importance of controlling for multicollinearity for the control of population genetic and data structure effects. Consider, two populations and a single focal SNP marker that has frequency p_1 in population 1 and frequency p_2 in population 2. The difference in allele frequency between the two populations is $\delta = p_1 - p_2$ and the average allele frequency across all the data is $\hat{p} = 0.5(p_1 + p_2)$. We define F_{ST} as $F_{ST} = \frac{0.5(p_1+p_2)^2}{\hat{p}(1-\hat{p}}$ and note that under this definition, F_{ST} scales with allele frequency, with common variants showing higher average F_{ST} than rare variants. The populations may have different mean value for a given trait with the difference $\bar{y}_1 - \bar{y}_2 = 2\beta(p_1 - p_2) + \Delta$, with β the effect size of the marker and Δ the non-genetic environmental contribution to the phenotypic difference. Eq. 2.3 of the Supplementary Note of [52] gives the expected bias of an effect size from a linear regression as $\hat{\beta} = \beta + \frac{\frac{1}{2}\Delta(p_1-p_2)}{2\bar{p}(1-\bar{p}(1+\frac{1}{2}F_{ST})}$, with β the

true effect size and we note that the bias term $c = \frac{\frac{1}{2}\Delta(p_1-p_2)}{2\bar{p}(1-\bar{p}(1+\frac{1}{2}F_{ST})}$ is proportional to the allele frequency. In principle, a MLMA approach will control for bias with correlated markers (either local or long-range LD) fitted as random when testing for the effects of a focal SNP. For two markers, $\mathbf{X}_1$and$\mathbf{X}_2$ in LD correlation $\rho_{\mathbf{X}_1,\mathbf{X}_2}$, with $\beta_2 = 0$ we can express the MLMA fixed effect solution as a partial regression coefficient of the phenotype regressed onto the focal SNP after adjusting for $\mathbf{X}_2$ estimated as $u_{\mathrm{X}_2} = \frac{\mathbf{X}_2^T\mathbf{y}}{\mathbf{X}_2^T\mathbf{X}_2+\lambda\mathbf{I}}$. Following our derivation above for a shrinkage estimator of a partial regression coefficient the effect size of $\mathbf{X}_1$ is estimated as $\hat{\beta}_{y,X_1|X_2} = \frac{N}{\mathbf{X}_1^T\mathbf{X}_1} \times \rho_{y,\mathrm{X}_1} - \frac{\rho_{\mathrm{X}_1,\mathrm{X}_2}\frac{1}{N}\mathrm{X}_2 y}{1-\rho_{\mathrm{X}_1,\mathrm{X}_2}}$ and in this two-SNP example the bias is accounted for in the term $\frac{\rho_{\mathrm{X}_1,\mathrm{X}_2}\frac{1}{N}\mathrm{X}_2 y}{1-\rho_{\mathrm{X}_1,\mathrm{X}_2}}$ when the fixed effect is estimated. Multicollinearity acts to increase the σ_G term of λ, reducing the denominator $\mathbf{X}_2^T\mathbf{X}_2 + \lambda\mathbf{I}$ in the estimation of u_{X_2}, and increasing the variance of the estimates of common markers in high LD, those with the highest average F_{ST} and the greatest potential bias from population stratification.

To confirm this, we conducted a simulation study using real genomic data from chromosome 22 where 10,000 individuals were selected from 2 UK Biobank assessment centres (Glasgow and Croydon). First, causal variants were allocated to 5000 high-LD SNPs with effect sizes simulated from a normal distribution with variance proportional to the F_{ST} among the two populations at each SNP. Second, we selected the same high-LD SNPs as the causal variants, but simulated effect sizes to have correlation 0.5 with the allele frequency differences of the SNPs among the two populations, and thus not only is the effect size proportional to the F_{ST}, but there is also directional differentiation (trait increasing loci tend to be those with higher allele frequency in Croydon, trait decreasing alleles have lower frequency in Croydon). For each of these two scenarios, we simulated 50 replicate phenotypes where the phenotypic variance attributable to the causal SNPs is 0.5, there is a phenotypic difference where Croydon individuals have a phenotype that is 0.5 SD higher than Glasgow individuals (contributing variance 0.05), and residual variance was simulated from a normal with variance 0.45, to give a phenotype with mean of zero and variance of 1. The data were then analysed using a mixed-linear model association (MLMA) and a grouped Bayesian dirac spike and slab models (BayesRR). In the analysis, we either adjusted the phenotype by the first 20 PCs of the genetic data used in the simulation study, or we did not adjust the phenotype for the PCs, to examine the effects of this common methods of population stratification control. In a two-population scenario the leading eigenvector encapsulates the allele frequency differentiation between the populations and thus the expectation is that this should adjust for these differences when estimating the marker associations. The results are presented in Figure S5, where we find that an MLMA approach overestimates the variance attributable to the SNPs under all scenarios, both with and without adjustment for PCs. BayesRR returns accurate estimates when the variance of the marker effects is proportional to F_{ST} and underestimates the variance when there is a directional associations, with this underestimation being less severe with PC adjustment.

UK Biobank data

We restricted our discovery analysis of the UK Biobank to a sample of European-ancestry individuals. To infer ancestry, we used both self-reported ethnic background (UK Biobank data code 21000-0) selecting coding 1 and genetic ethnicity (UK Biobank data code 22006-0) selecting coding 1. We also took the 488,377 genotyped participants and projected them onto the first two genotypic principal components (PC) calculated from 2,504 individuals of the 1,000 Genomes project with known ancestries. Using the obtained PC loadings, we then assigned each participant to the closest population in the 1000 Genomes data: European, African, East-Asian, South-Asian or Admixed, selecting individuals with PC1 projection $<$ absolute value 4 and PC 2 projection $<$ absolute value 3. This gave a sample size of 456,426 individuals.

To facilitate contrasting the genetic basis of different phenotypes, we then removed closely related individuals as identified in the UK Biobank data release. While the BayesRR model can accommodate relatedness similar to mixed linear models, we wished to simply compare phenotypes at markers that enter the model due to LD with underlying causal variants. Relatedness leads to the addition of markers within the model to capture the phenotypic covariance of closely related individuals, and this will vary across traits in accordance with the genetic and environmental covariance for each phenotype. For these unrelated individuals, we used the imputed autosomal genotype data of the UK Biobank provided as part of the data release. We used the genotype probabilities to hard-call the genotypes for variants with an imputation quality score above 0.3. The hard-call-threshold was 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p-value for Hardy-Weinberg test larger than 10-6, as determined in the set of unrelated Europeans) were selected those with minor allele frequency (MAF) > 0.0002 and rs identifier, in the set of European-ancestry participants, providing a data set 9,144,511

SNPs, short indels and large structural variants. From this we took the overlap with the Estonian Genome centre data to give a final set of 8,430,446 markers. From the UK Biobank European data set, samples were excluded if in the UKB quality control procedures they (i) were identified as extreme heterozygosity or missing genotype outliers; (ii) had a genetically inferred gender that did not match the self-reported gender; (iii) were identified to have putative sex chromosome aneuploidy; (iv) were excluded from kinship inference. Information on individuals who had withdrawn their consent for their data to be used was also removed. These filters resulted in a data set with 382,466 individuals.

We then selected the recorded measures of BMI (UK Biobank variable identifier 21001-0.0) and height (variable identifier 50-0.0) collected during initial assessment visit (year 2006-2010). BMI and height phenotypes 6 standard deviations (SD) away from the mean were not included in the analyses. For Type 2 Diabetes (T2D) in UKB we selected as cases very broadly as individuals who have main or secondary diagnosis (UKB fields 41202-0.0 - 41202-0.379 and 41204-0.0 - 41204-0.434) of "non-insulin-dependent diabetes mellitus" (ICD 10 code E11) or self-reported non-cancer illness (UKB field 20002-0.0 - 20002-2.28) "type 2 diabetes" (code 1223). From respondents self-reporting just "diabetes" (code 1220), we selected as cases those who did not self-report "type 1 diabetes" (code 1222) and had no Type 1 Diabetes (ICD code E10) diagnosis. Individuals with self-reported "diabetes" and "type 1 diabetes"/E10 were also left out from controls. We also defined coronary artery disease (CAD) cases broadly as participants with one of the following primary or secondary diagnoses or cause of death: ICD 10 codes I20 to I28; self-reported angina (code 1074) or self-reported heart attack/myocardial infarction (code 1075). Participants with self-reported "heart/cardiac problem" (code 1066) were not included as cases but also excluded from controls. This gave a sample size for each trait of 25,773 T2D cases and 359,730 T2D controls, 39,766 CAD cases and 344,054 CAD controls, 382,402 measures of height and 381,899 measures of BMI.

All phenotypes were adjusted for age of attending assessment centre (UKB code 21003-0.0, factor with levels for each age), year of birth (UKB field 34-0.0, factor with levels for each year), UK Biobank recruitment centre (UKB field 54-0.0, factor with levels for each centre), Genotype batch (UKB field 22000, factor with levels for each batch) and final 20 leading principal components of 1.2 million LD clumped markers from the 8,430,446 markers included in the analysis, calculated using flashPCA (see Code Availability). The residuals were then converted to z-scores with 0 mean and variance of 1. Similarly as for relatedness, population stratification is also accounted for within the BayesRR model through the addition of a background of marker effects entering the model, however we also wished to account for this in the standard manner by adjusting for the leading 20 PCs of the SNP data to get as close as possible to the inclusion of markers in the model that reflect LD with the causal variants. We note that as with any association model, while we take steps to adjust for known spatial (UKB centre), batch, and ancestry effects, and that the effects of each SNP is estimated jointly (and thus conditionally on the effects of all the other SNPs) environmentally induced covariance between SNP markers and a phenotype is still possible.

We partition SNP markers into 7 location annotations using the knownGene table from the UCSC browser data (see Code Availability), preferentially assigned SNPs to coding (exonic) regions first, then in the remaining SNPs we preferentially assigned them to intronic regions, then to 1kb upstream regions, then to 1-10kb regions, then to 10-500kb regions, then to 500-1Mb regions. Remaining SNPs were grouped in a category labelled "others" and also included in the model so that variance is partitioned relative to these also. Thus, we assigned SNPs to their closest upstream region, for example if a SNP is 1kb upstream of gene X, but also 10-500kb upstream of gene Y and 5kb downstream for gene Z, then it was assigned to be a 1kb region SNP. This means that SNPs 10-500kb and 500kb-1Mb upstream are distal from any known nearby genes. We further partition upstream regions to experimentally validated promoters, transcription factor binding sites (tfbs) and enhancers (enh) using the HACER, snp2tfbs databases (see Code Availability). All SNP markers assigned to 1kb regions map to promoters; 1-10kb SNPs, 10-500kb SNPs, 500kb-1Mb SNPs are split into enh, tfbs and others (un-mapped SNPs) extending the model to 13 annotation groups. Within each of these annotations, we have three minor allele frequency groups (MAF<0.01, 0.01>MAF>0.05, and MAF>0.05), and then each MAF group is further split into 2 based on median LD score. This gives 78 non-overlapping groups for which our BayesRR-RC model jointly estimates the phenotypic variation attributable to, and the SNP marker effects within, each group. For each of the 78 groups, SNPs were modelled using five mixture groups with variance equal to the phenotypic variance attributable to the group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1). We conducted a series of convergence diagnostic analyses of the posterior distributions to ensure we obtained estimates from a converged set of four Gibbs chains, each run for 6,000 iterations with a thin of 5 for each trait (Figure S6, S7, S8, S9).

Estonian Genome Centre data

For the Estonian Genome Centre Data, 32,594 individuals were genotyped on Illumina Global Screening (GSA) arrays and we imputed the data set to an Estonian reference, created from the whole genome sequence data of 2,244 participants [53]. From 11,130,313 markers with imputation quality score > 0.3, we selected SNPs that overlapped with the UK Biobank, resulting in a set of 8,433,421 markers.

We selected height and BMI measures from the Estonian Genome Centre data, in 32,594 individuals genotyped on GSA array and converted them to sex-specific z-scores after applying the same outlier removal procedure as in UKB and adjusting for the age at agreement. Prevalent cases of CAD and T2D in the Estonian Biobank cohort were first identified on the basis of the baseline data collected at recruitment, where the information on prevalent diseases was either retrieved from medical records or self-reported by the participant. The cohort was subsequently linked to the Estonian Health Insurance database that provided additional information on prevalent cases (diagnoses confirmed before the date of recruitment) as well as on incident cases during the follow-up.

As the UK Biobank marker effects are estimated from traits that were standardized to a z-score prior to analysis, all effect sizes obtained are on the SD scale. Thus when we create a genomic predictor, for say coding SNPs, by multiplying SNPs mapped to coding regions genotyped in Estonia to the effect sizes obtained in the UK Biobank for each iteration, to obtain a genetic predictor for each iteration, providing a posterior predictive distribution that is also on the SD scale. For each trait, we created 2000 genomic predictors for each individual in the Estonian Biobank, at each of the 13 annotation groups, by selecting effect size estimates obtained every tenth iteration from the last 3000 iterations of each of the four Gibbs chains and combining them together in a single posterior.

We calculated prediction accuracy as the proportion of phenotypic variation explained by the genomic predictor, and area under the receiver operator curve (AUC) for T2D and CAD using each individual's mean genetic predictor. For each of the 13 annotation groups, we calculated the partial correlation of the genetic predictor of each of the 2000 iterations and the phenotype, and then used this to estimate the independent proportional contribution of each group to the total prediction accuracy, providing a metric of replication for our UK Biobank enrichment results.

For height and BMI, we determined the probability that each Estonian individual's predictor accurately reflected their phenotypic value. To do this, we calculated the proportion of posterior samples with $\text{abs}(\hat{g} - y)$ of less than 1 for each individual, which gives a measure of the degree to which each posterior predictive distribution overlaps with the phenotype within +/- 1 SD.

For T2D and CAD, we extended the PCF metric, typically defined as the proportion of cases with larger estimated risk the then top p^{th} percentile of the distribution of genetic risk in the general population. We calculated the proportion of posterior samples for each individual with values in the top 25% of the distribution of genomic predictors for each trait. Thus for each individual, we calculate the probability that the posterior predictive distribution is in the top 25% of the distribution of genetic risk in the general population.

Posterior summaries and discovery

The ability of the additive regression model outlined and applied here to infer the underlying distribution of genomic effects is limited unless an additive model with many 0 coefficients holds as approximately true and the true number of underlying nonzero coefficients is $<< n$. Various ad hoc penalty functions in machine learning, and the range of proper priors employed by members of the Bayesian alphabet and beyond, all impose a restriction on the size of the regression coefficients, and while these restrictions differ, they all provide shrinkage estimators that by their definition are biased as they are shrunk toward zero (this true of mixed-linear association models also). In other words, the penalty function (prior) will be important and will influence the inference made here. Thus, the inference we obtain can only be made with respect to our *a priori* assumption that many marker effects are zero, and that the effects of those that are not zero can be reflected by a mixture of zero centred Gaussian distributions. Given this, we focused on comparing the posterior distributions of different traits obtained under the same model, focusing on the hyper-parameter estimates obtained for MAF-LD-annotation groups, and comparing these across traits. It has been shown in Bayesian penalized regression models that what is learned about β is a function of what is learned about $\mathbf{X}\boldsymbol{\beta}$ and thus by placing separate hyper-parameters over different genomic groups we can obtain inference as to the variance contributed by each group [54]. As we show through theory and simulation, MAF-LD-annotation specific hyper-parameters likely results in improved inference as to the distribution of genetic effects. However, with the exception of very rare variants with LD ~ 0, we cannot treat each β_j as independent and thus here we outline a strategy to identify associated genes, or genomic regions within a probabilistic framework.

For a simple example, consider two markers in LD that are correlated with a single causal variant, where either or both markers may be in the model at any one iteration and the expected posterior inclusion probability of each SNP is 0.5. In this scenario, we cannot use the posterior inclusion probability of each marker to assess association and thus instead, we take an approach of assessing the contribution of different genomic regions to trait variation whilst controlling the posterior type I error rate (PER), which is more suitable controlling for false positives, than controlling the genome-wide error rate (GER). Many papers have discussed the advantages of controlling the false discovery rate (FDR), and related measures rather than controlling GER [55] and here we follow [24] where the posterior probability that β_j is nonzero for at least one SNP j in a window or genomic segment is used to make inferences on the presence of an association in that segment.

Briefly, following [24], we will refer to this probability as the window posterior probability of association (WPPA). The underlying assumption is that if a genomic window contains a marker in LD with a causal variant, one or more SNPs in that window will have nonzero β_j. Thus, WPPA, which is estimated by counting the number of MCMC samples in which β_j is nonzero for at least one SNP j in the window, can be used as a proxy for the posterior probability that the genomic region contains a causal variant. Because WPPA for a given window is a partial association conditional on all other SNPs in the model, including those flanking the region, the influence of flanking markers on the WPPA signal for any given window will be inversely related to the distance k of the flanking markers. Thus, as the number of markers between a causal variant and the focal window increases, the influence of the causal variant on the WPPA signal will decrease and so WPPA computed for a given window can be used to locate associations for that given window [24].

This measure can be shown to control the PER, which in frequentest statistics would be associated with the test of a hypothesis. The null hypothesis in this case is that the genomic region does not contain any SNPs associated with the trait. Using this notation, WPPA is the conditional probability that the null is false given the observed data, while PER is the conditional probability that the null hypothesis is true given that it has been rejected based on some statistical test. Suppose the test is based on WPPA and the null is rejected whenever WPPA is larger than some value t. Then, PER is the probability that that the null hypothesis is true given WPPA is larger than t, and it can be written as:

$$\text{PER} = \Pr(H_0 \text{is true} | \text{WPPA} > t) = E[(1 - \text{WPPA}) | \text{WPPA} > t] \tag{45}$$

Thus, for any interval with WPPA $> t$ the proportion of false positives among significant results will be $\leq (1 - t)$. Here, we are interested in detecting genes and genomic regions that explain more than some proportion v of the total phenotypic variance attributable to the SNP markers (genetic variance). The genomic segment variance is defined as the sum of the squared partial regression coefficient estimates at each iteration and these are divided by the sum of all the squared partial regression coefficient estimates genome-wide to give a proportion for each genomic region at each iteration. Then we simply count the proportion of MCMC samples where the proportion of genetic variance is greater than a thresholds of 0.001% and we denote this metric as the posterior probability of window variance (PPWV). We estimate the PPWV of 50kb regions across the genome, then map SNPs to the coding region of genes, and to the closest gene +/- 50kb from the SNP position labelling them as located in a coding region, an intron, 1kb upstream of a gene using our functional annotations (Figure **??**). Remaining snps are labelled as located in a cis-region (up to +/- 50kb from a gene). Finally, we mapped SNPs with greater than 50% posterior inclusion probability (PIP) across all 4 chains labelling them using our 7 location annotations (Figure S13). We report SNPs with PIP $> 95\%$ and their corresponding p-values from UKB GWAS summary statistics (fastGWA, see Code Availability) with 'body mass index' entry for BMI, 'standing height' for HT, 'angina / heart attack' for CAD and 'diabetes' for T2D (Supplementary Table S6).

We also validate the use of PPWV in simulation study, first simulating 500 replicate data sets of 10,000 SNP markers for 5,000 individuals for each of two scenarios. In the first scenario, 1000 SNPs are randomly selected to be causal variants and all 10,000 SNP markers are LD independent. In the second, the 1000 causal variants are each in LD with four other variants with LD = 0.95, with the remaining 5000 variants having zero effect size and LD = 0. For each scenario, we simulate effect sizes as an equally spaced sequence from an effect size of -0.04 SD, to 0.04 SD giving genetic variance of 0.55, and we simulate residual variance from a normal distribution with zero mean and variance 0.45, to give a phenotype with zero mean and unit variance. For the first scenario, we calculate the posterior inclusion probability of each causal SNP. For the second scenario, we calculate the PPWV for each 5-SNP group. Across the 500 replicates of each scenario, we take the mean PPWV and mean PIP for each of the 1000 different effect sizes and compare these in Figure S12. Additionally, we grouped SNPs in 50kb regions and selected the number of regions that explain at least 0.1%, 0.01% and 0.001% of the variance attributed to all SNP markers in 0.8% to 100% of the iterations

using the simulated data described above for the Multiple group enrichment scenario for chromosome 22 in the UK Biobank. We then calculated the false discovery rate (FDR), defined as the proportion of 50kb regions identified that do not contain a causal variant, at PPWV thresholds ranging from 0.8% to 100%. We compare these in Figure S12 where as we lower the PPWV variance threshold, the number of false discoveries in the model increases but remains at $\leq 5\%$ when the PPWV is $\geq 95\%$.

Data availability

This project uses UK Biobank data under project 35520. The Estonian Genome Centre data are available upon request from the cohort authors with appropriate research agreements. Summaries of all posterior distributions obtained are provided in Supplementary data sets. Full posterior distributions of the SNP marker effects sizes for each trait are deposited on Dryad `https://datadryad.org/`

Code availability

Our BayesRR-RC model is implemented within the software Hydra, with full open source code available at: `https://github.com/medical-genomics-group/hydra`.

UCSC Table Browser `https://genome.ucsc.edu/cgi-bin/hgTables`

flashPCA `https://github.com/gabraham/flashpca`

Plink1.90 `https://www.cog-genomics.org/plink2/`

GCTA `https://cnsgenomics.com/content/software`

HACER database `http://bioinfo.vanderbilt.edu/AE/HACER/`

snp2tfbs database `https://ccg.epfl.ch//snp2tfbs/`

fastGWA database `http://fastgwa.info/ukbimp/phenotypes/`

Computing environment `https://www.epfl.ch/research/facilities/scitas/hardware/helvetios/`

Author contributions

MRR conceived and designed the study. MP, DTB, and AK contributed to the study design. MP and MRR conducted the experiments and analyses with input from DTB, AK, SEO, JS, PMV, RM and LR. MRR, DTB, SEO, and LR derived the equations and the algorithm. EJO and DTB developed the software, with contributions from MRR, MP, SEO, AK, and GM. MRR, MP, and DTB wrote the paper. RM provided study oversight and contributed data to the analysis. All authors approved the final manuscript prior to submission.

Author competing interests

The authors declare no competing interests.

Acknowledgements

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. We would like to thank the participants of the cohort studies, and the Ecole Polytechnique Federal Lausanne (EPFL) SCITAS for their excellent compute resources, their generosity with their time and the kindness of their support. PMV acknowledges funding from the Australian National Health and Medical Research Council (1113400) and the Australian Research Council (FL180100072). LR acknowledges funding from the Kjell Märta Beijer Foundation (Stockholm, Sweden). We also would like to acknowledge Simone Rubinacci, Oliver Delanau, Alexander Terenin, Eleonora Porcu, and Mike Goddard for their useful comments and suggestions.

Supplementary Online Material

Marion Patxot, Daniel Trejo Banos, Athanasios Kousathanas, Etienne J. Orliac, Sven E. Ojavee, Gerhard Moser, Julia Sidorenko, Zoltan Kutalik, Reedik Mägi, Peter M. Visscher, Lars Rönnegård, Matthew R. Robinson

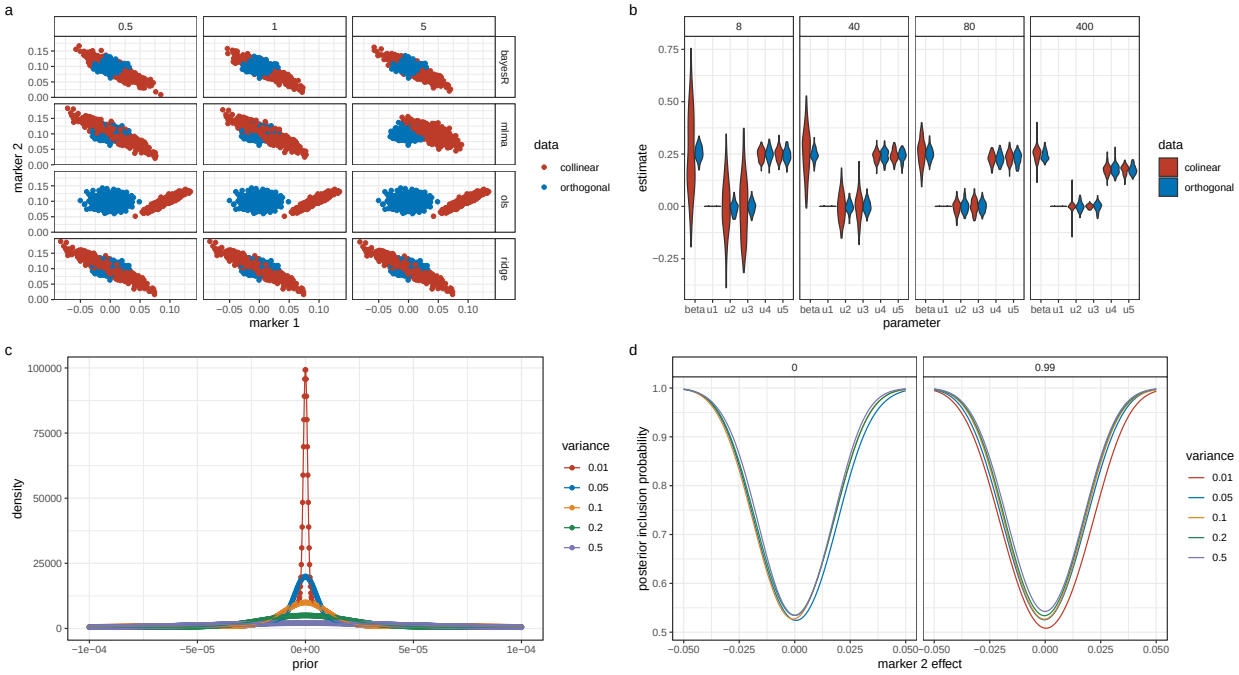

Figure S1. Theory and simulation study of SNP marker model parameters. (a) accompanies Eq. (23) and shows the distribution of the point estimates of the effect sizes of two correlated markers of effect size (0,0.1) under orthogonality (LD = 0) and collinearity (LD = 0.99) across 2500 replicates (50 independent genomic regions for 5,000 individuals within each of 50 replicates) for a range of different models: a dirac spike and slab mixture of regressions model (bayesR), a mixed linear association model (MLMA), single-marker ordinary least squares (OLS), and ridge regression (Ridge). Panels give the lambda shrinkage parameter of the model, the error variance divided by the phenotypic variance attributable to the SNP markers, showing that as lambda decreases the variation of the estimates increases under multicollinearity. (b) accompanies Eq.(26) and shows the marker estimates obtained from Henderson's mixed model equations for a MLMA with the focal marker as fixed (beta) and random (u1), with four other markers in the model. Markers were either uncorrelated (orthogonal, LD=0) or the focal marker was correlated with the first two out of the four other markers (collinear, LD=0.99). Panels give the lambda shrinkage parameter, showing that as lambda decreases the variation of the estimates increases under multicollinearity. (c) shows the prior density of the BayesR model for different hyperparameter values of the phenotypic variance attributable to genetic effects (variance), showing that as the variance attributable to the markers decreases, the prior has higher mass around zero. Thus, with a grouped mixture of regressions model (BayesRR), each hyperparameter estimate will be smaller and thus there will be higher prior density around zero. This then has consequences for marker inclusion in the BayesRR model. Higher prior mass around zero makes little difference for the inclusion of uncorrelated markers, but it results in reduced posterior inclusion probability for correlated markers as shown in (d). For (d), we calculated the inclusion probability (PIP) of two markers with LD = 0 and LD = 0.99, as the variance attributable to the SNP markers, and thus the prior distribution, changes assuming a background inclusion probability of 0.1, a sample size of 5000, and an effect size of 0.01 SD for marker 1 (see Methods). (d) shows that the PIP of the second marker is reduced across a range of possible effect size values (the average of 1000 replicated simulations for 1000 marker 2 effect values for each line) as the hyperparameter estimate decreases, and thus the smaller hyperparameter estimates in a BayesRR model means that correlated markers are less likely to enter the model, controlling better for the effects of multicollinearity.

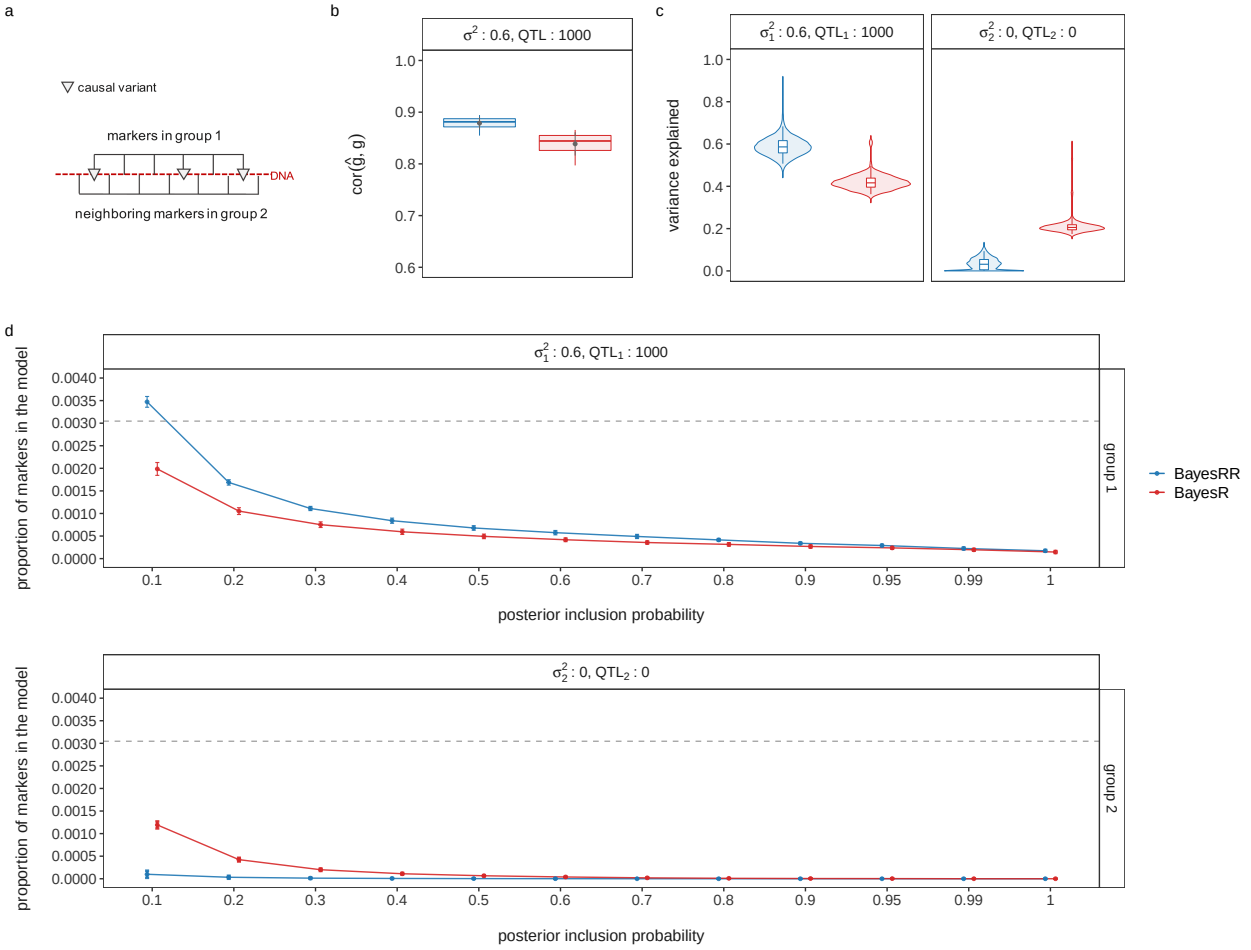

Figure S2. Classification power of BayesRR. Grouping effects in a BayesRR model improves the power of BayesR to estimate effect sizes and infer the genetic architecture of common complex traits and diseases. This setting compares 10 simulations of 5 chains with different starting values (chain length : 2500, burn-in : 500, thin : 5) executed using BayesRR and our BayesR software. (a) Each simulation has two groups in high LD with an interdigitated structure where one in two SNPs is assigned to group 1 and all genetic variance is assigned to group 1 with 1000 QTL. Annotation-specific estimates for BayesR are calculated post-analysis for each group. (b) Estimation of markers effects in an independent data set. BayesRR improves on correlation between predicted and simulated genetic values. This increase in prediction implies that adding functional information to BayesR better fits the data and improves prediction accuracy. (c) Genetic variance and (d) proportion of markers entering the model at posterior inclusion probability (pip) thresholds summarized across 10 simulations for group 1 and group 2. The proportion of markers included in the model is closer to the truth (dotted grey line) when using BayesRR compared our BayesR software. Effects are thus more likely attributed to the correct group using our approach, which also explains why we estimate more accurately the group genetic variance compared to the baseline. Simulation setting: $N = 20,000$ unrelated European individuals from the UK Biobank, $M = 328,385$ markers (chromosome 2). Dots in box plots show the mean of the correlation between predicted and simulated genetic values.

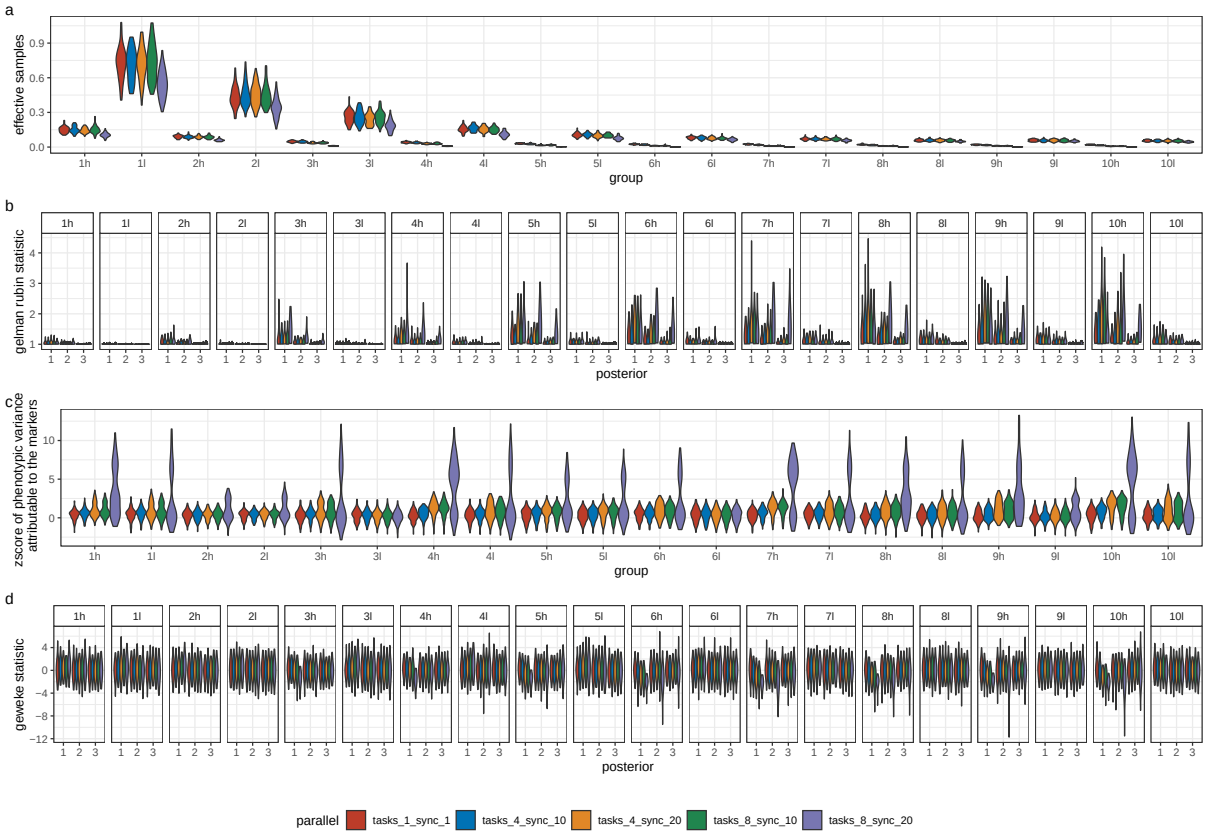

Figure S3. Simulation study of increasing task parallelism and increasing message passing rate for our hybrid-parallel sampling scheme. We aimed to compare (a) the effective samples obtained, (b) the convergence rate of the algorithm, (c) the accuracy of the estimation, and (d) the stability of the estimates obtained as data parallelism increases within a burn-in period of the initial 3000 iterations. For 50,000 randomly selected UK Biobank individuals, and 111,425 imputed SNP markers of chromosome 22, we simulated 50 replicate phenotypes by randomly selecting 4,988 SNPs as causal variants and randomly allocating effect sizes from a normal distribution, with SNP heritability of 0.5. For each simulation, we ran three chains of our BayesRR model with different starting values for 3000 iterations. The SNP marker data was grouped into deciles of the distribution of minor allele frequency (MAF) and within each decile the markers were further grouped these into two groups based on the distribution of linkage disequilibrium (LD), giving twenty groups in total (1l = MAF decile 1, low LD; 1h = MAF decile 1, high LD; ...; 10l = MAF decile 10, low LD; 10h = MAF decile 10, high LD). We repeated the three chains, but with increasing data parallelism: (1) in serial where one MPI task is used and the residual is updated after each marker is sampled (tasks_1_sync_1); (2) where the markers were split across four MPI processes with synchronisation occurring by message passing after 10 markers have been updated (task_4_sync_10); (3) where the markers were split across four MPI processes with synchronisation occurring after 20 markers have been updated (task_4_sync_20); (4) with 8 MPI processes and synchronisation of 10 (task_8_sync_10); and (5) with 8 MPI processes and synchronisation of 20 (task_8_sync_20). (a) shows the distribution across simulations of the proportion of effective samples obtained for the hyperparamter estimate of the proportion of phenotypic variance attributable to the markers of each group. For all ranges of parallelism, we achieve more effective samples for low MAF and low LD variants. With high synchronisation rates, where many marker updates occur before residual updating by message passing a reduction in effective sample sizes occurs. (b) gives the distribution of the Gelman-Rubin test statistic for the three chains, a general metric to monitor convergence that compares within- and among-chain variance, as the number of iterations increases. On the x-axis, 1 gives the distribution of the statistic across chains and MAF-LD groups for the first 500 iterations showing divergence of the chains (y-axis value >> 1) across all MAF-LD groups, 2 gives the distribution for the first 1000 iterations, and 3 gives the distribution for the whole chain showing convergence of the chains by the end of this initial 3000 iteration sampling period irrespective of the data parallelism, with the exception of a few groups with infrequent synchronisation and high data parallelism which have yet to converge within this burn-in phase. (c) gives the distribution of z-scores of the posterior distribution of the phenotypic variance attributable to the markers for each MAF-LD group from the simulated values, showing stability of the estimates with increasing data parallelism (tasks), but not with infrequent synchronisation within the 3000 iterations run here. (d) shows the distribution of the Geweke statistic value which is a test of the equality of the means of the first and last part of the Markov chains. On the x-axis, 1 gives the distribution of the statistic calculated using all iterations across all MAF-LD groups, 2 gives the distribution discarding the first 500 iterations, and 3 gives the distribution discarding the first 1000 iterations. (a) - (d) suggest that our hybrid-parallelism sampling scheme achieves the same accuracy and convergence rates as a serial sampling scheme, provided that frequent synchronisation occurs and data parallelism is kept moderate. At high data parallelism and infrequent synchronisation, our theory shows that we are more likely to make a sampling mistake, preventing chains from converging and requiring longer sampling times. Convergence and accuracy of the MCMC Gibbs sampling chain will be problem specific and dependent upon the LD of the markers, LD among the causal variants, the phenotypic variation attributable to the SNP markers across the MAF and LD spectrum, the study sample size, the degree of data parallelism per total marker number, and the synchronisation rate. Therefore, like with all MCMC chains, a series of diagnostic tests can be utilized to explore the properties of the posterior and here we show how different metrics can be used to identify convergence issues.

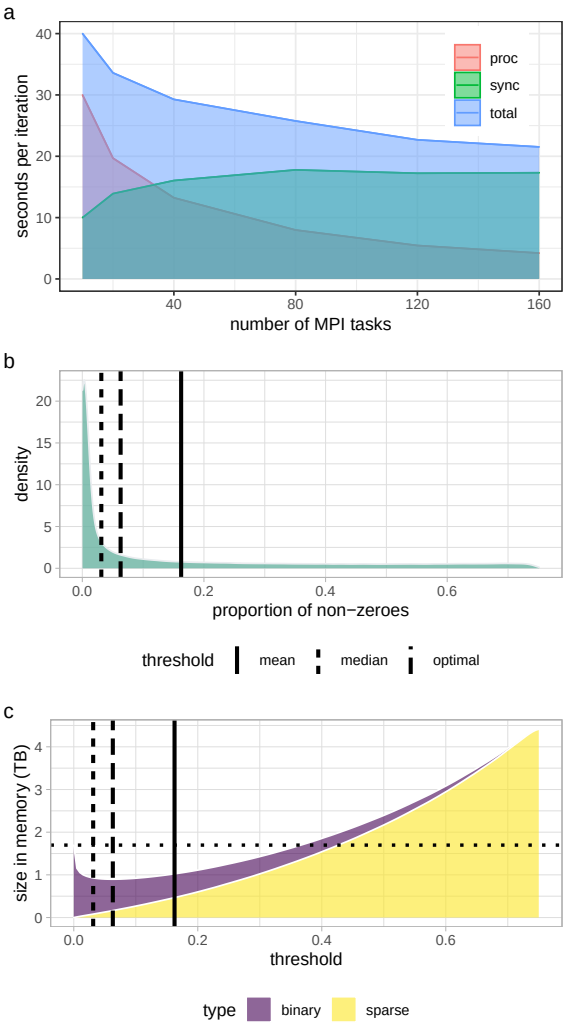

Figure S4. A mixed representation bulk synchronous hybrid-parallel Gibbs sampling scheme for genomic data.(a) The minimum seconds per iteration achieved for 382,466 unrelated individuals from the UK Biobank data genotyped at 8,430,466 markers, with an increasing number of message-passing interface (MPI) tasks used. The total seconds is given in blue and this is subset into (i) the time taken to process the markers and estimate all of the 8,433,421 marker effects and hyper-parameters (proc), and (ii) the time taken to synchronise the estimates as they are being obtained (sync). With increasing data parallelism parameter estimation times drop quickly to less than 5 seconds with 160 MPI tasks, however the time taken to synchronise the estimates increases as the number of tasks increases. The SD was 1 second, with variation in sampling times induced by fluctuations in networking speed that influenced the synchronisation times. Each MPI task was able to used 4 CPUs. (b) the distribution of the proportion non-zeros per column of a genotype matrix for $\sim 4 \times 10^5$ individuals and $\sim 1.5 \times 10^7$ SNPs taken from UKB, with solid line representing the mean of the distribution and dashed line the median. (c) the size in memory in TB of the data as the coding of the SNP markers moves from binary to the sparse indexed format, the optimal threshold is achieved between mean and median of the distribution of non-zeros in the genotype matrix. Above this threshold columns are coded in binary format below in sparse index. Through a combination of a mixed data representation and highly vectorized look-up tables, memory usage is reduced while maintaining fast computational speed.

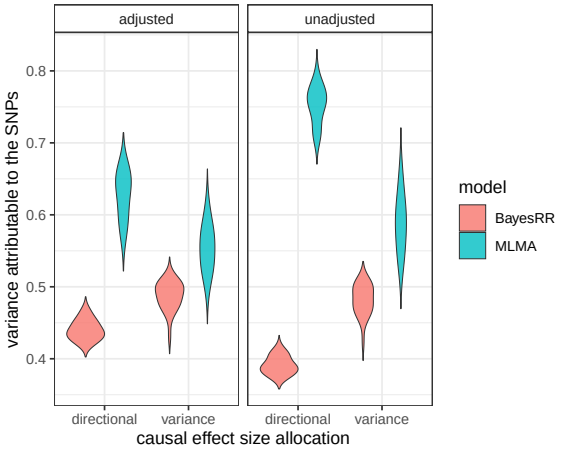

Figure S5. Comparison of a mixed-linear association model (MLMA) and a grouped dirac spike and slab model (BayesRR) when genetic effects have variance proportional to F_{ST} (labelled 'variance'), or correlated with allele frequency differentiation across populations (labelled 'directional'). Simulation study using real genomic data from chromosome 22 where 10,000 individuals were selected from 2 UK Biobank assessment centres (Glasgow and Croydon). First, causal variants were allocated to 5000 high-LD SNPs with effect sizes simulated from a normal distribution with variance proportional to the F_{ST} among the two populations at each SNP (labelled 'variance', see Methods). Second, we selected the same high-LD SNPs as the causal variants, but simulated effect sizes to have correlation 0.5 with the allele frequency differences of the SNPs among the two populations, and thus not only is the effect size proportional to the F_{ST}, but there is also directional differentiation (trait increasing loci tend to be those with higher allele frequency in Croydon, trait decreasing alleles have lower frequency in Croydon). For each of these two scenarios, we simulated 50 replicate phenotypes where the phenotypic variance attributable to the causal SNPs is 0.5, there is a phenotypic difference where Croydon individuals have a phenotype that is on average 0.5 SD higher than Glasgow individuals (contributing variance 0.05), and residual variance was simulated from a normal with variance 0.45, to give a phenotype with mean of zero and variance of 1. The distribution across simulations of the estimated phenotypic variance attributable to the SNP markers is shown for each of the two causal effect size allocation scenarios when the data was analysed using a mixed-linear model association (MLMA, distribution of the point estimates) and a grouped Bayesian dirac spike and slab models (BayesRR, distribution of the posterior means). In the analysis, we either adjusted the phenotype by the first 20 PCs of the genetic data used in the simulation study ("adjusted") or we did not adjust the phenotype for the PCs ("unadjusted").

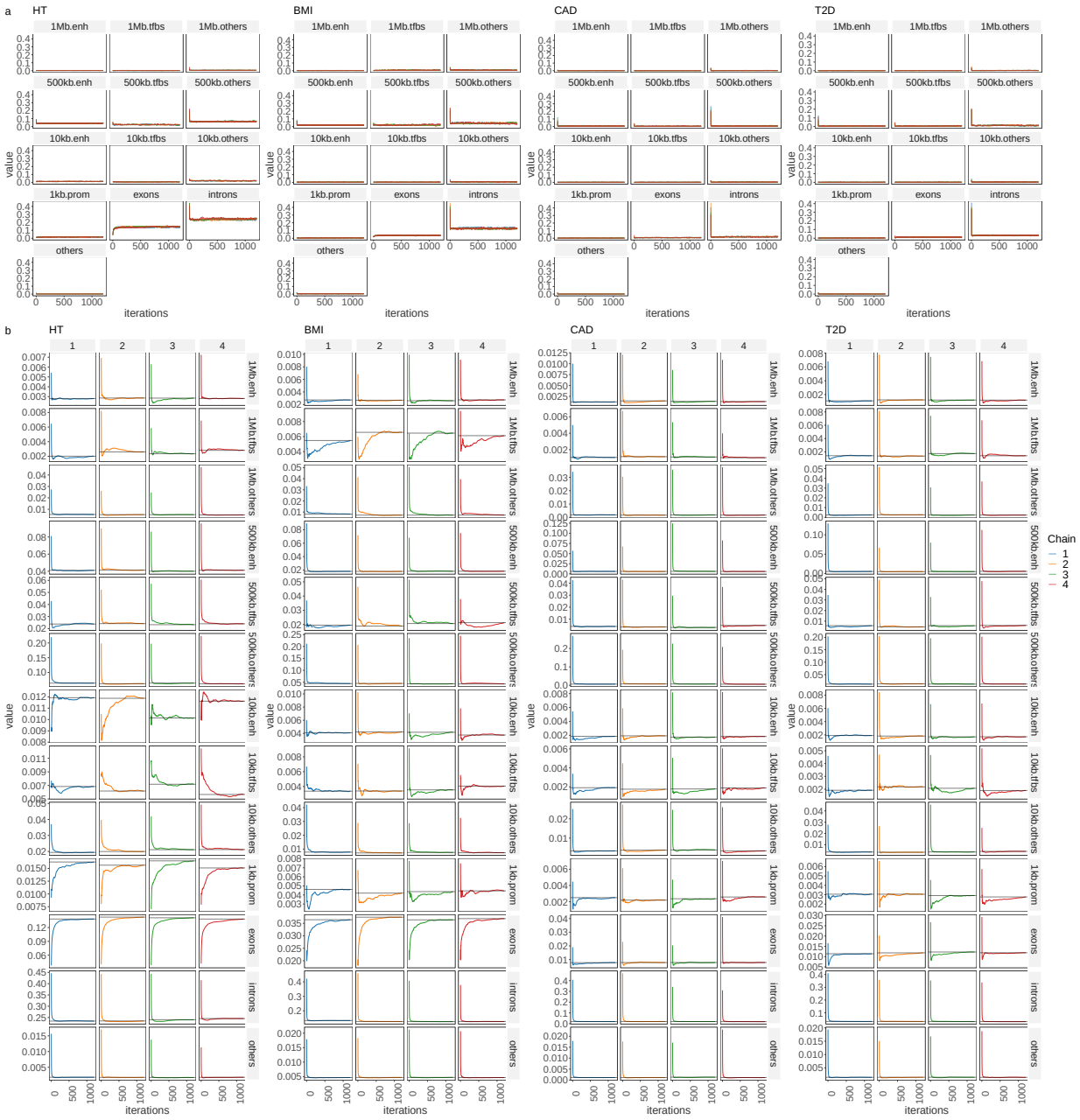

Figure S6. Convergence diagnostics of model chains for UK Biobank analysis.(a) traceplot of the phenotypic variance attributable to SNP markers for each trait across functional annotation of exonic regions, intronic regions, promoters (prom) 1kb upstream of coding regions, enhancers (enh) 1kb to 10kb upstream of coding regions, transcription factor binding sites (tfbs) 1kb to 10kb upstream of coding regions, other snps 1kb to 10kb upstream of coding regions, enh 10kb to 500kb upstream, tfbs 10kb to 500kb upstream, other snps 10kb to 500kb upstream, enh 500kb to 1MB upstream,tfbs 500kb to 1Mb upstream, other snps 500kb to 1Mb upstream and SNP markers elsewhere in the genome (other), with colours representing the different chains. (b) a time series of the running mean of each chain, for each annotation group and each trait showing all chains approach the same mean value for each parameter.

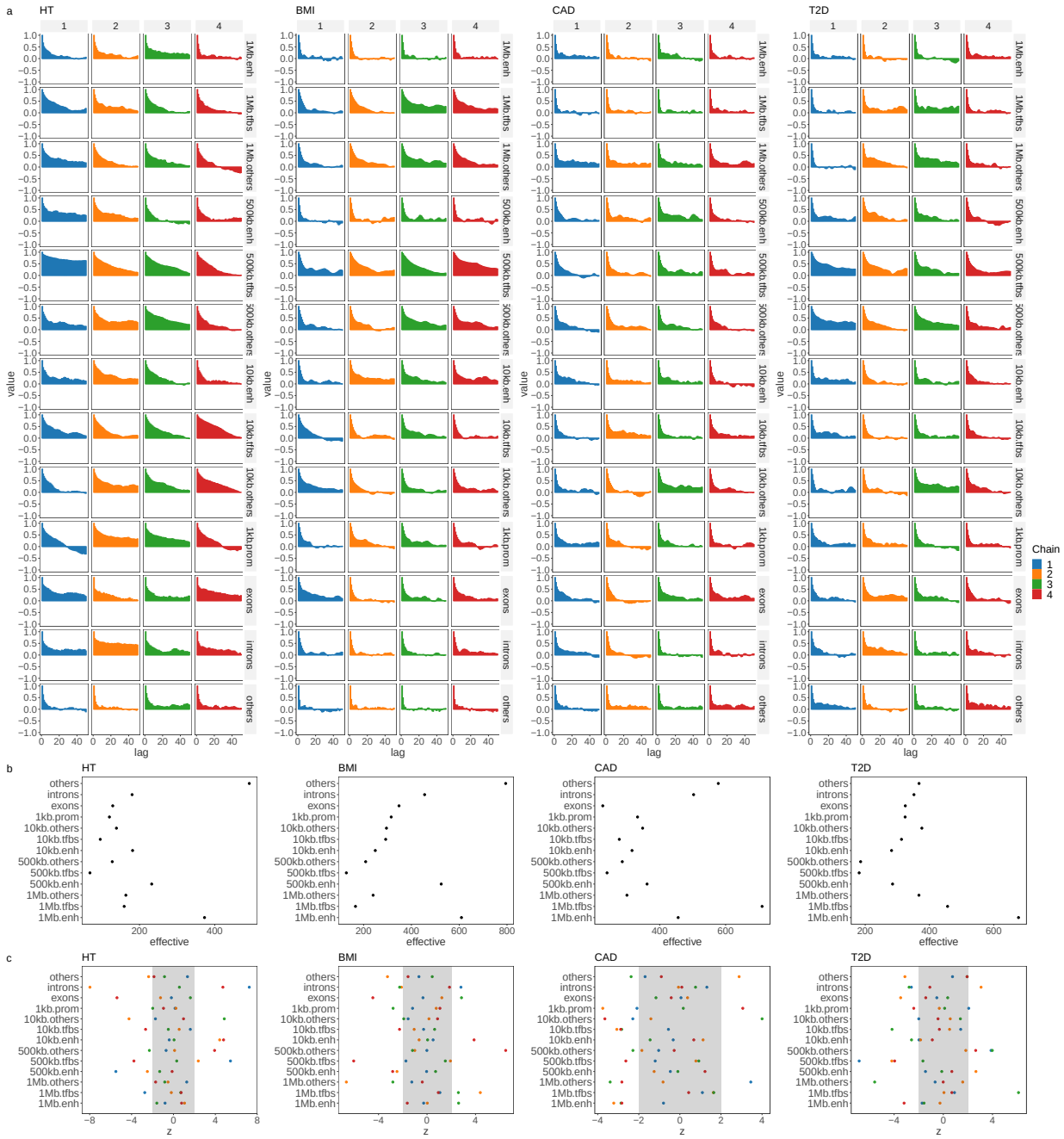

Figure S7. Convergence diagnostics of model chains for UK Biobank analysis.(a) lagged autocorrelation plot of each chain, for each annotation group and each trait and (b) effective number of uncorrelated sampled obtained for each annotation group and each trait. As phenotypic variance is being partitioned it is not expected that posterior estimates obtained are entirely uncorrelated. (c) Geweke z-score statistic comparing the initial part of the chain to the final part, for each annotation group and each trait.

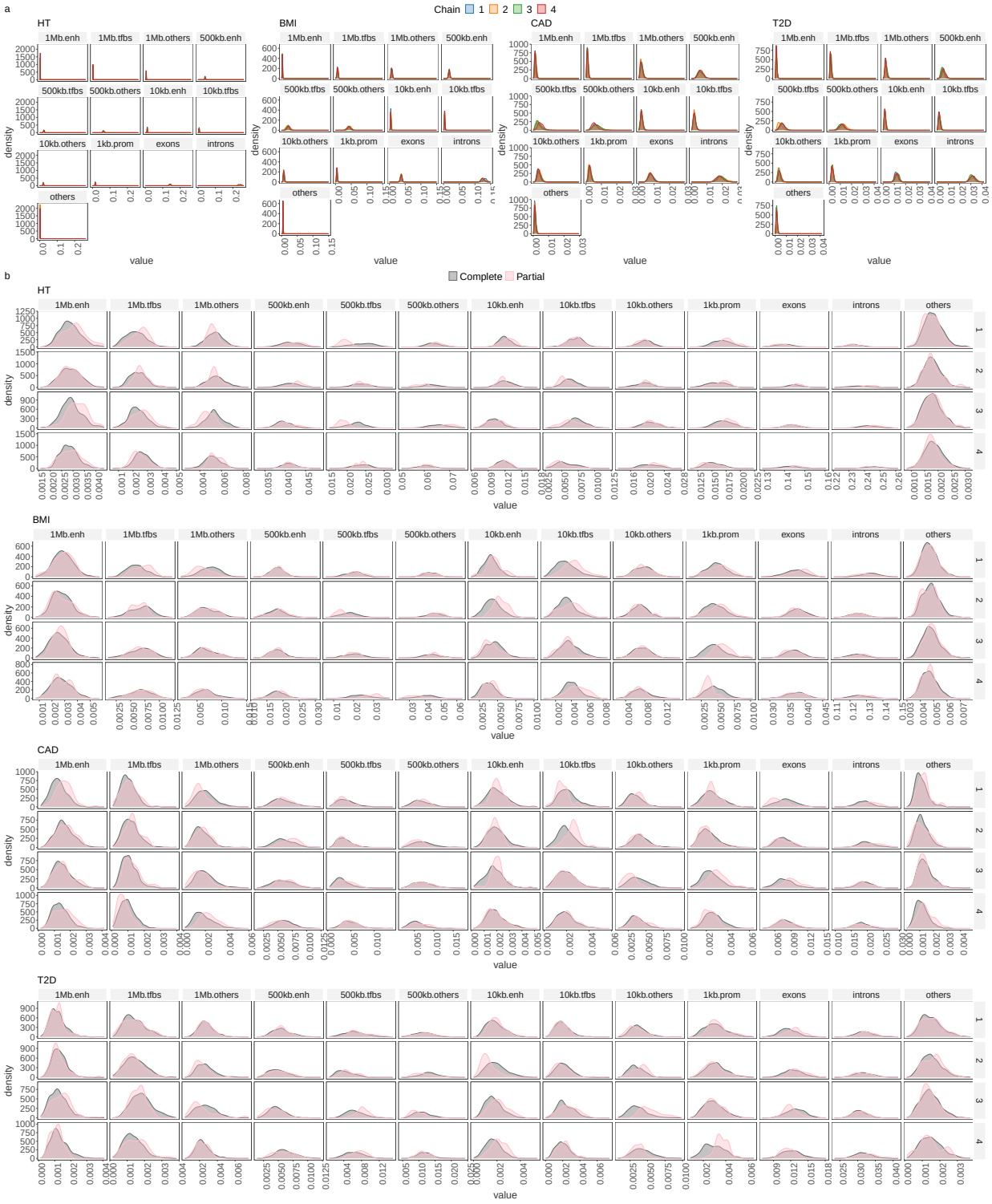

Figure S8. Convergence diagnostics of model chains for UK Biobank analysis. (a) overlapped density plots to compare the target distribution by chain showing each chain has converged in a similar space, for each annotation group and each trait. (b) overlapped density plots comparing the last 10 percent of the chain (green), with the whole chain (pink), showing that the initial and final parts of the chain are sampling the same target distribution for each annotation group and each trait.

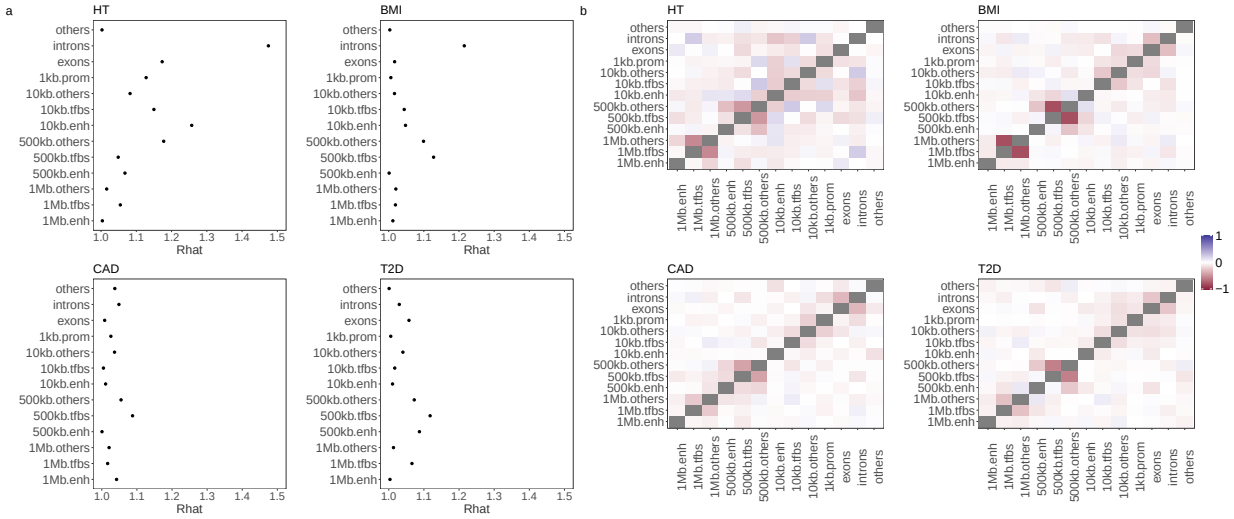

Figure S9. Convergence diagnostics of model chains for UK Biobank analysis. (a) the potential scale reduction factor comparing the among- and within-chain variance for each annotation group and each trait. (b) the cross-correlation between all parameters for each annotation group and each trait.

Figure S10. Genetic architecture of height, body-mass-index (BMI), cardiovascular disease (CAD) and type-2-diabetes (T2D). (a) Shows violin plots with boxplots giving the 95% credible intervals for the posterior mean of the phenotypic variance attributable to the SNP markers in each trait. We find that SNPs contribute 57.66% (95%CI 56.09, 59.14) for height, 28.74% (95%CI 27.62, 30.00) for BMI, 5.94% (95%CI 5.30, 6.67) for CAD and 8.45% (95%CI 7.83, 9.18) for T2D. Values are summed over annotation, MAF and LD groups. (b) Violin plot with boxplots giving the 95% credible intervals of the proportion of the total genetic variance attributable to each annotation group. Values are summed over MAF and LD groups. All four traits show the same pattern of annotation-specific genetic variance, with main contributions from intronic regions, exonic regions, and SNPs located 10kb to 500kb upstream of genes to the genetic variance in the population. (c) Bar plots with error bars giving the 95% credible intervals for the proportion of variance of each annotation group that is attributable to each of the four non-zero mixtures for each trait. Values are summed over MAF and LD groups. (d) Bar plots with error bars giving the 95% credible intervals for the proportion of variance of each annotation group that is attributable to each of the three MAF groups for each trait. Values are summed over LD groups. Within each annotation, variation is (c) attributable predominantly to variants with MAF>0.05 and (d) attributable predominantly to small (0.0001) to moderate (0.001) effect sizes variants with little differences across traits, except for BMI which has higher polygenicity compared to height, CAD and T2D.

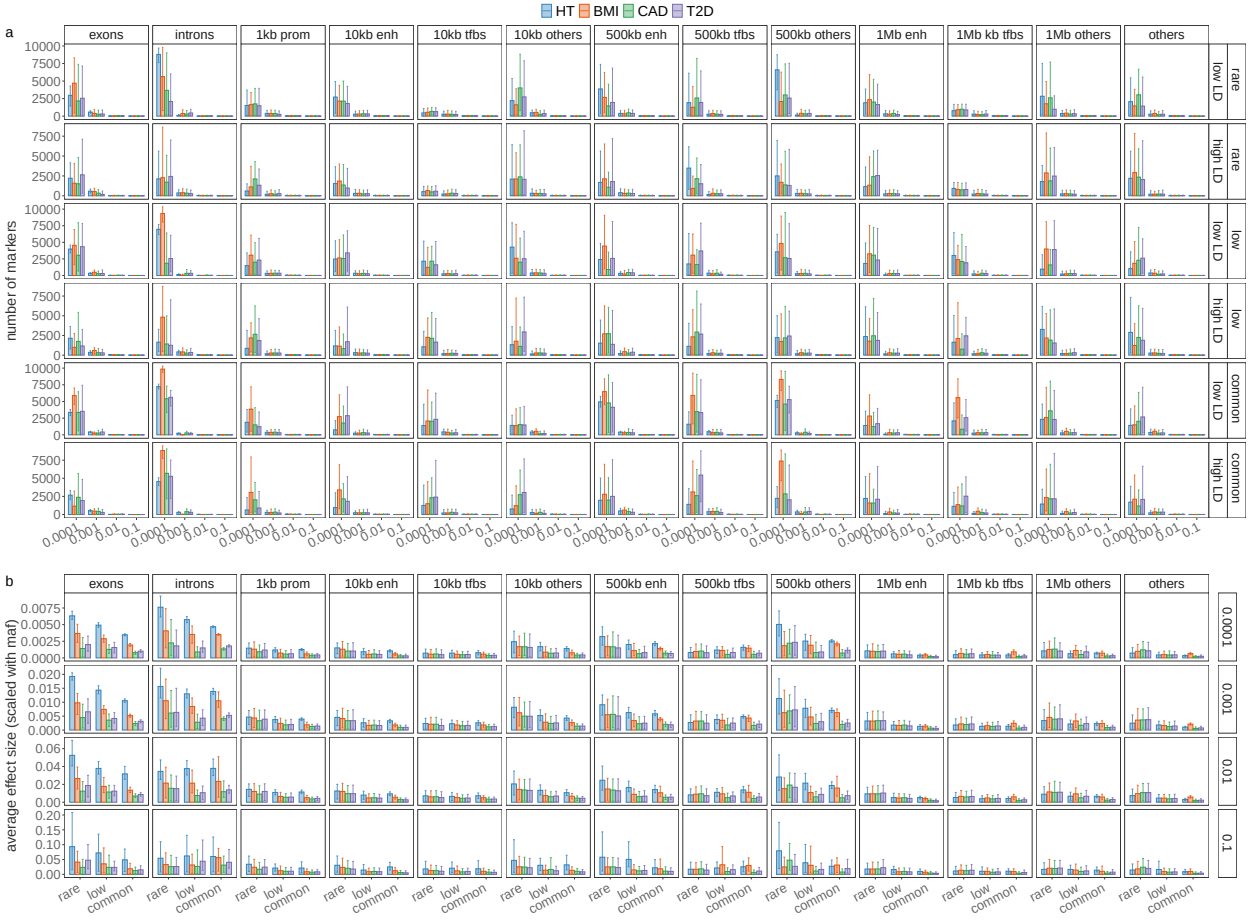

Figure S11. Marker inclusion and effect estimate overview.(a) Barplots of the number of markers entering the model for each mixture group (x-axis), within each MAF-LD group (y-axis facets, with top row MAF and bottom row LD), within each annotation (x-axis facets). Mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1. (b) Boxplots of the posterior distribution of the average effect size of markers in the model for each annotation group, scaling the effects to their frequency and split by mixture.

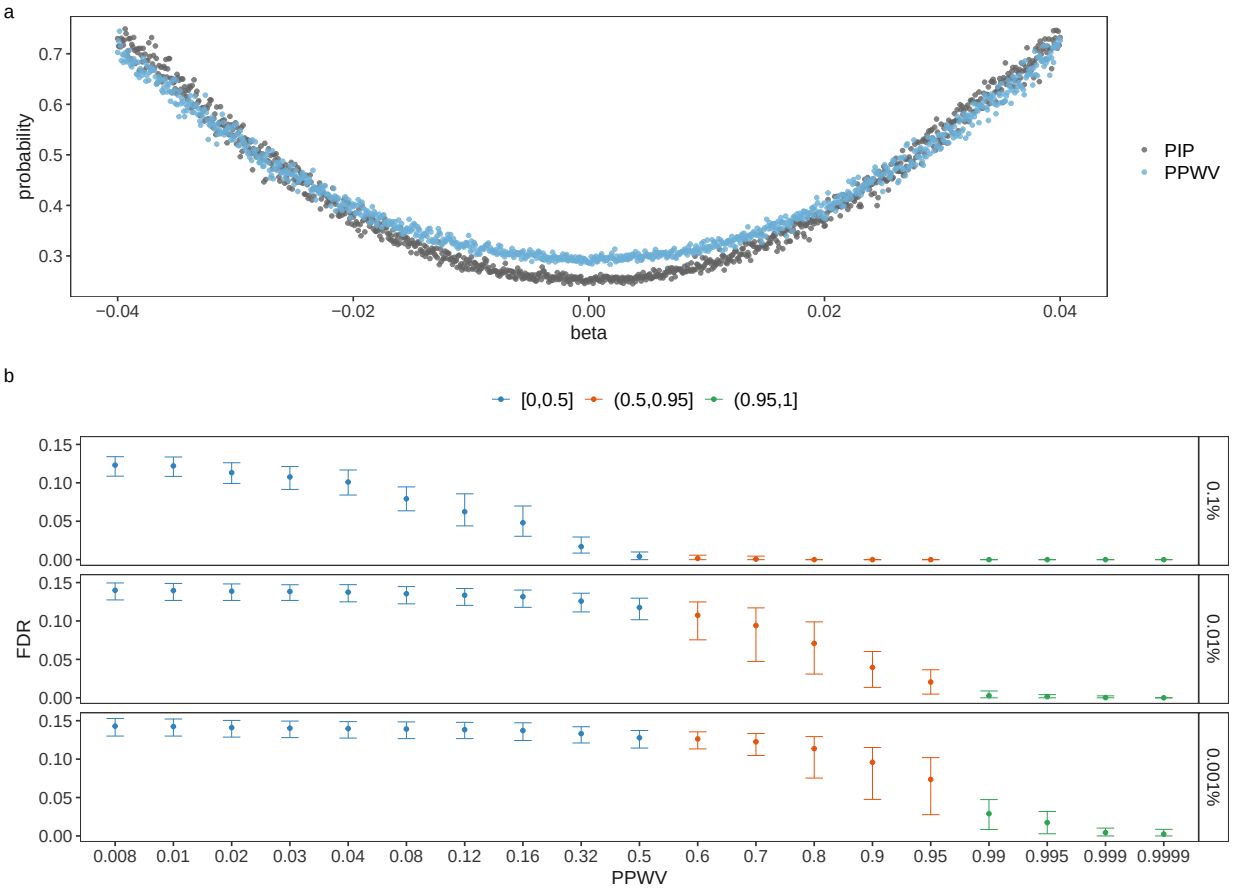

Figure S12. Posterior inclusion probability (PIP) and posterior probability of window variance (PPWV).(a) We validate the use of PPWV in simulation study, first simulating 500 replicate data sets of 10,000 SNP markers for 5,000 individuals for each of two scenarios. In the first scenario, 1000 SNPs are randomly selected to be causal variants and all 10,000 SNP markers are LD independent. In the second, the 1000 causal variants are each in LD with four other variants with LD = 0.95, with the remaining 5000 variants having zero effect size and LD = 0. For each scenario, we simulate effect sizes as an equally spaced sequence from an effect size of -0.04 SD, to 0.04 SD giving genetic variance of 0.55, and we simulate residual variance from a normal distribution with zero mean and variance 0.45, to give a phenotype with zero mean and unit variance. For the first scenario, we calculate the posterior inclusion probability of each causal SNP. For the second scenario, we calculate the PPWV for each 5-SNP group. Across the 500 replicates, we take the mean PIP for each SNP of the 1000 different effect sizes for the first scenario and the mean PPWV of each of the 1000 5-SNP windows for the second scenario, and these are the points on the figure. (b) Shows mean and 95% credible interval of the false discovery rate (FDR), defined as the proportion of regions identified that do not contain a causal variant, at PPWV thresholds ranging from 0.8% to 100%. Here, we grouped SNPs in 50kb regions and selected the number of regions that explain at least 0.1%, 0.01% and 0.001% of the variance attributed to all SNP markers in 0.8% to 100% of the iterations using simulated data for chromosome 22 in the UK Biobank (see Methods). We compare the FDR at these different PPWV thresholds and as we lower the PPWV variance, the number of false discoveries in the model increases, but remains at $\leq 5\%$ at PPWV $\geq 95\%$.

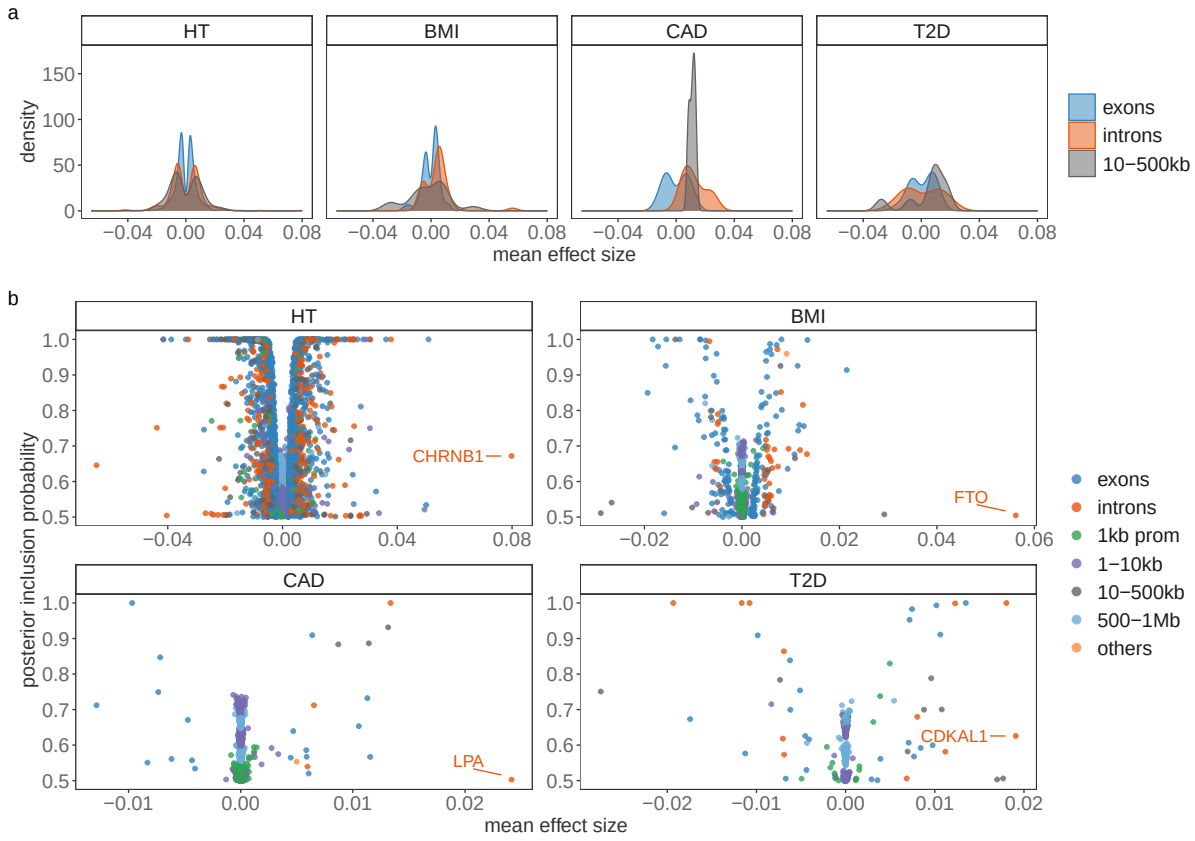

Figure S13. Contribution of SNPs with posterior inclusion probability (PIP) > 0.5 to height, body-mass-index (BMI), cardiovascular disease (CAD) and type-2-diabetes (T2D). (a) Shows the distribution of mean effect sizes for SNPs with PIP > 0.5 attributed to exons, introns and 500kb upstream of genes in each trait. (b) We then plot the relationship between mean effect size and posterior inclusion probability for SNPs with PIP > 0.5 attributed to the annotation groups (exons, introns, SNPs located 1kb, 1-10kb, 10-500kb and 500-1Mb upstream of genes and other un-mapped SNPs). We labelled the closest gene to the SNP with the highest mean effect size in each trait.

References

1. Luke M Evans, Rasool Tahmasbi, Scott I Vrieze, Gonçalo R Abecasis, Sayantan Das, Steven Gazal, Douglas W Bjelland, Teresa R De Candia, Michael E Goddard, Benjamin M Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737–745, 2018.

2. Doug Speed, Na Cai, Michael R Johnson, Sergey Nejentsev, David J Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature Genetics*, 49(7):986, 2017.

3. Doug Speed, John Holmes, and David J Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, 2020.

4. Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics*, page 1, 2019.

5. Steven Gazal, Carla Marquez-Luna, Hilary K Finucane, and Alkes L Price. Reconciling s-ldsc and ldak functional enrichment estimates. *Nature Genetics*, 51(8):1202–1204, 2019.

6. Doug Speed and David J. Balding. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics*, 51(2):277–284, feb 2019.

7. Longda Jiang, Zhili Zheng, Ting Qi, Kathryn E. Kemper, Naomi R. Wray, Peter M. Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, 2019.

8. Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284, 2015.

9. Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Gonçalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv*, 2020.

10. Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, 2018.

11. Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R. Day, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R.B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J. Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

12. M. Erbe, B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7):4114–4129, 2020/05/10 2012.

13. Gerhard Moser, Sang Hong Lee, Ben J. Hayes, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genetics*, 11(4):1–22, 04 2015.

14. Daniel Trejo Banos, Daniel L McCartney, Marion Patxot, Lucas Anchieri, Thomas Battram, Colette Christiansen, Ricardo Costeira, Rosie M Walker, Stewart W Morris, Archie Campbell, et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nature Communications*, 11(1):1–14, 2020.

15. Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

16. Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264, Feb. 2016.

17. Ismaël Castillo, Johannes Schmidt-Hieber, Aad Van der Vaart, et al. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

18. IM MacLeod, PJ Bowman, CJ Vander Jagt, M Haile-Mariam, KE Kemper, AJ Chamberlain, C Schrooten, BJ Hayes, and ME Goddard. Exploiting biological priors and sequence variants enhances qtl discovery and genomic prediction of complex traits. *BMC Genomics*, 17(1):144, 2016.

19. Rasmus Froberg Brøndum, Guosheng Su, Mogens Sandø Lund, Philip J Bowman, Michael E Goddard, and Benjamin J Hayes. Genome position specific priors for genomic prediction. *BMC Genomics*, 13(1):543, 2012.

20. Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

21. Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

22. Po-Ru Loh, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, Naomi R Wray, Kenneth S Kendler, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385, 2015.

23. M. Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245 EP –, 08 2009.

24. Rohan Fernando, Ali Toosi, Anna Wolc, Dorian Garrick, and Jack Dekkers. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. *Journal of Agricultural, Biological and Environmental Statistics*, 22(2):172–193, 2017.

25. Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang, and Peter M. Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications*, 10(1):5086, 2019.

26. Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

27. Nicholas Mancuso, Malika K Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, 51(4):675–682, 2019.

28. Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177 – 1186, 2017.

29. James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020.

30. Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

31. Sebastian Zöllner, Xiaoquan Wen, and Jonathan K Pritchard. Association mapping and fine mapping with treeld. *Bioinformatics*, 21(14):3168–3170, 2005.

32. Mark J Minichiello and Richard Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, 79(5):910–922, 2006.

33. Gemma E. Moran, Veronika Ročková, and Edward I. George. Variance prior forms for high-dimensional bayesian variable selection. *Bayesian Anal.*, 14(4):1091–1119, 12 2019.

34. Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

35. Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.

36. Yali Amit and Ulf Grenander. Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, 37(2):197–222, 1991.

37. C. M. Theobald. Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):103–106, 1974.

38. Robert M. Maier, Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, Naomi R. Wray, Jian Yang, Peter M. Visscher, and Matthew R. Robinson. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications*, 9(1):989, 2018.

39. Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna AE Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John RB Perry, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, 47(10):1114, 2015.

40. C.R. Henderson. Best linear unbiased prediction of breeding values not in the model for records. *Journal of Dairy Science*, 60(5):783 – 787, 1977.

41. Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.

42. Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

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

44. Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

45. Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

46. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

47. John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016.

48. Bala Rajaratnam, Doug Sparks, Kshitij Khare, and Liyuan Zhang. Uncertainty quantification for modern high-dimensional regression via scalable bayesian methods. *Journal of Computational and Graphical Statistics*, 28(1):174–184, 2019.

49. Matthew Johnson, James Saunderson, and Alan Willsky. Analyzing hogwild parallel gaussian gibbs sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2715–2723. Curran Associates, Inc., 2013.

50. Elaine Angelino, Matthew James Johnson, Ryan P Adams, et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.

51. Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Bogdan Pasaniuc, and Sriram Sankararaman. Scalable multi-component linear mixed models with application to snp heritability estimation. *bioRxiv*, page 522003, 2019.

52. Matthew R Robinson, Gibran Hemani, Carolina Medina-Gomez, Massimo Mezzavilla, Tonu Esko, Konstantin Shakhbazov, Joseph E Powell, Anna Vinkhuyzen, Sonja I Berndt, Stefan Gustafsson, Anne E Justice, Bratati Kahali, Adam E Locke, Tune H Pers, Sailaja Vedantam, Andrew R Wood, Wouter van Rheenen, Ole A Andreassen, Paolo Gasparini, Andres Metspalu, Leonard H van den Berg, Jan H Veldink, Fernando Rivadeneira, Thomas M Werge, Goncalo R Abecasis, Dorret I Boomsma, Daniel I Chasman, Eco J C de Geus, Timothy M Frayling, Joel N Hirschhorn, Jouke Jan Hottenga, Erik Ingelsson, Ruth J F Loos, Patrik K E Magnusson, Nicholas G Martin, Grant W Montgomery, Kari E North, Nancy L Pedersen, Timothy D Spector, Elizabeth K Speliotes, Michael E Goddard, Jian Yang, and Peter M Visscher. Population genetic differentiation of height and body mass index across europe. *Nature Genetics*, 47(11):1357–1362, 2015.

53. Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M. Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, Jaak Vilo, and Lili Milani. Genetic variation in the estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442–454, 2019.

54. Daniel Gianola. Priors in whole-genome regression: The bayesian alphabet returns. *Genetics*, 194(3):573–596, 2013.

55. Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.