We develop a Bayesian model (BayesRR-RC) that provides robust SNP-heritability estimation, an alternative to marker discovery, and accurate genomic prediction, taking 22 seconds per iteration to estimate 8.4 million SNP-effects and 78 SNP-heritability parameters in the UK Biobank. We find that only ≤10% of the genetic variation captured for height, body mass index, cardiovascular disease, and type 2 diabetes is attributable to proximal regulatory regions within 10kb upstream of genes, while 12-25% is attributed to coding regions, 32-44% to introns, and 22-28% to distal 10-500kb upstream regions. 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 ≥95% probability of contributing ≥0.001% to the genetic variance of these four traits. Our open-source software (GMRM) provides a scalable alternative to current approaches for biobank data.
s 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. Rather than relying on subsets of the SNPs, methods should fully utilise the data, exploiting computational power to facilitate discovery of additional genomic regions, to improve understanding of the genomic architecture of common disease, and to provide more informative genomic prediction.

For example, when estimating the proportion of phenotypic variance attributable to different categories of genetic markers (the SNP-heritability, \( h^2_{\text{SNP}} \) of a genomic region), recent studies\(^1\)\(^–\)\(^4\) highlight the importance of accounting for minor allele frequency (MAF) and LD structure of the genomic data. Generally, 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 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, as this would yield a better understanding of the polygenicity of genomic effects across different genomic annotation groups.

Furthermore, statistical inference usually follows a multi-step approach. 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 without the use of MAF-LD-annotation information, and then estimating the marker effect sizes one-by-one as fixed effects in a second step\(^7\)\(^–\)\(^10\). Following this initial mixed-model association step, statistical inference (variance components, fine mapping and risk prediction) is then typically conducted on the summary statistics generated. The advantage of a multi-step approach is that large sample size can be easily obtained through meta-analyses, combining summary statistics from different studies and avoiding the need for individual-level data sharing. However, as large-scale biobank data is increasingly available, methods that provide joint estimates of the marker effects in a single step by estimating the effect sizes one-by-one as

\[
y = \mu + \sum_{q=1}^{p} \mathbf{X}_q \beta_q + \epsilon,
\]

where there is a single intercept term \( \mu \) and a single error term, a vector (\( N \times 1 \)) of residuals \( \epsilon \), with \( e_i \sim \mathcal{N} \left( 0, \sigma^2 \right) \). An \( N \times p \) matrix of single nucleotide polymorphism (SNP) genetic markers, centred and scaled to unit variance, which we denote as \( \mathbf{X}_q \). The effects are allocated into groups (1, \( \ldots \), \( \Phi \)). Each group has a set of model parameters \( \Theta_q = (\beta_q, \pi_q, \sigma^2_{\text{SNP}, q}) \), with \( \beta_q \) as a \( p_q \times 1 \) vector of partial regression coefficients, where \( \beta_{q \ell} \) is the effect of a 1 SD change in the \( \ell \)th covariate within the \( q \)th group. The spike and slab prior, contains what is called a Dirac spike\(^14\)\(^–\)\(^15\) for \( \beta_q \), 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 centred at zero constitute the slab component. The slab shrinks the non-zero coefficients towards zero according to the slab’s width, and by having a scale mixture of Gaussians, the distribution has heavier tails and can accommodate big and small effects\(^16\). Therefore, each \( \beta_{q \ell} \) is distributed according to:

\[
\beta_{q \ell} \sim \pi_{q \ell} \delta_0 + \pi_{q \ell} \mathcal{N} \left( 0, \sigma^2_{\text{SNP}, q} \right) + \pi_{q \ell, p} \mathcal{N} \left( 0, \sigma^2_{\text{p}, q} \right) + \ldots + \pi_{q \ell, p_q} \mathcal{N} \left( 0, \sigma^2_{\text{p}_q, q} \right),
\]

where for each SNP marker group \( \{ \pi_{q \ell}, \pi_{q, 1 \ell}, \ldots, \pi_{q, p_q} \} \) are the mixture proportions and \( \{ \sigma^2_{\text{SNP}, q}, \sigma^2_{\text{p}, q}, \ldots, \sigma^2_{\text{p}_q, q} \} \) are the mixture-specific variances proportional to

\[
\begin{bmatrix}
\sigma^2_{1 \ell} \\
\vdots \\
\sigma^2_{p_q, q}
\end{bmatrix}
= \begin{bmatrix}
C_{1q} \\
\vdots \\
C_{p_q, q}
\end{bmatrix},
\]

with \( \sigma^2_{\text{p}, q} \) the phenotypic variance associated with the SNPs in group \( p \), which, like all the other parameters, is estimated directly from the data. Here, we use 78 MAF-LD-annotation SNP marker groups. SNPs are partitioned into seven location annotations preferentially to coding (exonic) regions first, then to intronic regions, then to 1 kb upstream regions, then to 1–10 kb regions, then to 10–50 kb regions, then to 50–100 Mb regions, then to 500–1 Mb 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 1 kb upstream of gene X, but also 10–500 kb upstream of gene Y and 5 kb downstream for gene Z, then it was assigned to be a 1 kb region SNP. This ensures that SNPs 10–500 kb and 500–1 Mb upstream are distal to any known gene. We further partition upstream regions to experimentally validated promoters, transcription factor binding sites (tfs) and enhancers (enh) using the HACER, snp2tfs databases.
to cing computational complexity of a single Gibbs step from sparse-index representation (see Supplementary Note 2), reducing schemes by deriving sampling steps to utilise whole interface (MPI) tasks. We extend previous sparse residual within and across compute nodes in a series of message-passing parallel (BSP) Gibbs sampling scheme for large-scale genomic data, as the view that Bayesian approaches to large-scale genomic data is the view that the computation of a posterior distribution is too expensive. In Supplementary Note 2, we derive a Bulk Synchronous hybrid approach to enable a direct comparison, (ii) a Haseman–Elston (HE) regression using the same 78 group model implemented in the LDAK software, which are suggested to outperform all other approaches. In comparison to the best LDAK predictor, BayesRR-RC can obtain or even exceed the theoretical expectation of prediction accuracy under ridge regression assumptions (Fig. 2b, see “Methods” section). We then conduct a number of follow-up simulation studies. Recent work has highlighted differences in statistical model performance depending upon the relationship of SNP marker effect size, LD and MAF. We explore the performance of our model in theory, with highly correlated genetic markers in Supplementary Note 4. We also conducted another large-scale, but well-powered, simulation study to explore the model performance of BayesRR-RC as compared to existing approaches.

Simulation study. We first compare the model performance of BayesRR-RC to existing approaches across 18 different genetic architectures. We randomly selected 40,000 unrelated UK Biobank individuals and used 596,741 imputed SNP markers from chromosomes 19 to 22. We randomly selected either 1000, 10,000 or 100,000 LD independent (LD R^2 < 0.1) causal SNP markers. For each SNP marker set, we then simulated effect sizes from a normal distribution with zero mean and variance of 0.1, 0.3 or 0.6 divided by the number of causal variants and \( \pi N(\rho, \rho(1-\rho) - 0.25) \), with \( \rho \) the allele frequency (see “Methods” section). This simulates stronger effect sizes for rare variants in line with recent empirical estimates and we simulated ten replicate phenotypes for each of the nine different genetic architectures. We then additionally repeat each simulation, sampling the SNP marker effects this time from 13 different distributions, one for each of 13 different genomic annotation groups with different proportions of \( h^2_{SNP} \) to create nine further different genetic architectures. We compare our BayesRR-RC model to the following statistical models: (i) a restricted maximum likelihood (REML) model implemented in the software BoltREML; with the same 78 MAF-LD-annotation groups enabling a direct comparison, (ii) a Haseman–Elston (HE) regression using the same 78 group model implemented in the software RHEMc, (iii) summary statistic linkage disequilibrium score regression (LDSC), with LD scores calculated using the same data, and the same 78 non-overlapping annotations in a 78 component LDSC annotation model, and (iv) summary statistic SumHer (LDAK) with the same 78 non-overlapping annotations.

We find that BayesRR-RC estimates the phenotypic variation attributable to different genomic annotation groups comparable with the BoltREML model, with similar correlation of the estimated and simulated values within each simulation replicate (Fig. 1a). In comparison, RHEmc, which also uses individual-level data, yields estimates with lower correlation with the simulated value, but higher than both summary statistic approaches implemented in LDSC and SumHer (Fig. 1a). We calculate estimates of enrichment, defined as the proportion of \( h^2_{SNP} \) attributable to the annotation divided by the proportion of SNPs mapping to the annotation (for bayesRR-RC, because there is sparsity in the SNP effects, we define enrichment as the proportion of SNPs in the model that map to the annotation, see “Methods” section) and we compare these to the true simulated value. Compared to other approaches, we find that BayesRR-RC gives a lower probability of false enrichment, calculated as the proportion of times within a simulation replicate that an annotation group was incorrectly assigned as having enrichment greater than 2 (Fig. 1b). Thus, BayesRR-RC provides accurate partitioning of genomic enrichment across the genome.

In Supplementary Note 3, we propose a posterior probability window variance (PPWV) approach, which provides a probabilistic determination of association of a given LD block, genomic window, gene, or upstream region, relative to the amount of phenotypic variation attributable to that window. Our PPWV approach determines the posterior inclusion probability that each region and each gene contributes at least 0.001% to the \( h^2_{SNP} \), with theory outlined in Supplementary Note 3 suggesting well controlled FDR. We determine the ability of our PPWV approach to correctly localise an association to LD blocks (defined as groups of markers with LD R^2 ≥ 0.1) that contain causal variants, and compare this to using LD to clump mixed-linear model association estimates obtained using the BoltLMM software (Fig. 2a). We find that a PPWV approach identifies associated LD blocks with higher probability as compared to clumped MLMa associations, for all genetic architectures, with the exception of simulated phenotypes with enrichment and low polygenicity, where the small numbers of relatively large effect size regions are better identified with a single-marker regression approach (Fig. 2a). Thus, BayesRR-RC provides an alternative to standard genome-wide association studies to localise SNP-phenotype associations at the regional level, especially for traits with high polygenicity.

We then also compare the prediction accuracy obtained in an independent sample when creating genomic predictors using (i) effect sizes estimated by BayesRR-RC, (ii) fixed-effect SNP effect sizes estimated in the MLMa approach implemented in bolt, and (iii) effect size estimates obtained from four different genomic prediction models proposed in a recent paper; implemented in the LDAK software, which are suggested to outperform all other current approaches. In comparison to the best LDAK predictor, we find that BayesRR-RC obtains similar or improved prediction accuracy across all genetic architectures, with greater prediction accuracy gains observed under genetic architectures where the SNP effect distributions differed across genomic annotations (Fig. 2b). We find that given sufficient power, BayesRR-RC can obtain or even exceed the theoretical expectation of prediction accuracy under ridge regression assumptions (Fig. 2b, see “Methods” section).
The genetic architecture of four complex traits in the UK Biobank. We apply BayesRR-RC to cardiovascular disease outcomes (CAD), type-2 diabetes (T2D), body mass index (BMI) and height measured for 382,466 unrelated individuals from the
We conducted a series of convergence diagnostic analyses of the year of birth, genotype batch effects, UK Biobank assessment allele frequency >0.0002. We adjust each phenotype for age, sex, introns, and within 500 kb of genes across all traits, and Supplementary Data 2). For all annotation groups in exons, the other phenotypes (Fig.3b; Table1 and Supplementary Fig. 11 thousands of common variants, each of small effect (Fig. 3b expected prediction accuracy based on ridge regression theory. Error bars show the SD in both panels.

Fig. 2 Simulation study for the performance of our BayesRR-RC model implemented in the GMRM software against existing approaches for localisation of associations and genomic prediction. a Probability of detecting genomic regions containing simulated causal variants by a Bayesian regional fine-mapping approach (GMRM: blue) versus standard mixed linear model association (MLMA) testing (BOLT: green). The column facets give the simulated heritability and rows give the number of causal variants and whether the effect sizes differed across genomic annotation groups (enrich) or were randomly assigned (random). b Correlation of a genomic predictor and a phenotype in an independent sample when the genomic predictor is created from GMRM effects sizes (blue), MLMA effect sizes using BOLT (green), and the optimal effect sizes obtained from individual-level and summary statistic models implemented in the Mega-PRS LDG approach (purple). The column facets give the simulated heritability and the number of causal variants. The row facets give whether the effect sizes differed across genomic annotation groups (enrich) or were randomly assigned (random). The red lines give the expected prediction accuracy based on ridge regression theory. Error bars show the SD in both panels.

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 section) and have minor allele frequency >0.0002. We 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 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 6000 iterations with a thin of five for each trait (Supplementary Figs. 7–10).

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 10–500 kb upstream of genes, with proximal (within 10 kb) promoters, enhancers and transcription factor binding sites cumulatively contributing <10% (Fig. 3b and Supplementary Fig. 11), with estimates summed across MAF and LD groups Table 1, and full results in Supplementary Data 2). The large contribution of exonic and intronic annotations to variation is in-line with the fact that these annotations account for ~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 (Fig. 3b; Table 1 and Supplementary Fig. 11 and Supplementary Data 2). For all annotation groups in exons, introns, and within 500 kb of genes across all traits, ≥60% of the $h^2_{SNP}$ attributable to this group is contributed by many thousands of common variants, each of small effect (Fig. 3b and Supplementary Figs. 11 and 12).

Our estimates compare similarly to those obtained by RHEmc and SumHer, but differ to those obtained by LDSC (Table 1 and Supplementary Data 3, 4, and 5 for full results). In addition to providing variance component estimates, our model facilitates assessment of 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 500 kb upstream regions is underlain by many thousands of SNPs that on average each contribute 0.001% (estimates summed across MAF and LD groups in Fig. 3b and Supplementary Figs. 11 and 12). In contrast, the variance is spread more evenly across the mixtures for the other groups, implying that 10–500 kb 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 (Fig. 3 and Supplementary Fig. 12). Therefore, we find that 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.

Across traits, posterior mean 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 10–500 kb upstream regions (Fig. 3c). We find little evidence that SNPs located in proximal promoters, enhancers, and transcription factor binding sites within 10 kb of genes showed average effect sizes that were higher than SNPs located 1 MB away from genes, or those that were not mapped to
a specific category, with perhaps the exception of high MAF variants (Fig. 3c). Generally, all phenotypes simply appear to be predominantly underlain by very many common variants, with SNPs within distal regulatory regions, coding and intronic regions contributing more to the variance. 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, and again we find that rare variants have higher average allele substitution effects than common variants for exonic, intronic, promoters and enhancers (Supplementary Fig. 12b). An exception to these patterns were BMI-associated intronic and 10–500 kb group SNPs, where we find no evidence that the allele substitution effect size differs across frequency groups (Supplementary Fig. 12b).

**Discovery of associated genomic regions.** We then partitioned the variance attributed to SNP markers across 50 kb regions of the genome, then across SNPs annotated to genes, and then to LD blocks of the DNA using our PPWV approach. We find 1660 50 kb regions for height with ≥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 (Fig. 4a and Table 2). We then map the phenotype association to these genomic regions using a BayesRR-RC model for each of the 78 groups with phenotypic variance attributable to each group multiplied by constants (mixture 0 = 0, mixture 1 = 0.0001, 2 = 0.001, 3 = 0.01, 4 = 0.1)

**Fig. 3 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 seven location annotations (coding regions, intronic regions, and windows 1, 1–10, 10–500 kb and 500 kb–1 Mb upstream of genes, with other SNPs grouped in a category labelled “others”). Windows 1–10 kb, 10–500 kb and 500 kb–1 Mb 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 annotated as rare, 0.01 < MAF ≤ 0.05 annotated as low, and MAF > 0.05 annotated as common), and then each MAF group is further split into two 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.
SNPs to their closest gene (+/−50 kb from SNP position) and we use our annotations to label them (see “Methods” section). We find 243 independent coding regions for height with ≥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 regions of genes with 1254 independent cis-regions for height with ≥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 95% posterior probability of explaining at least 0.001% of the $h^2_{SNP}$, 1162 independent intronic regions of genes for BMI, 307 for CAD and 347 for T2D. When we calculate the number of exons, introns, promoters and cis regions with ≥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 (Fig. 4b). Generally, we find that only 1% or less of the available exons and promoter 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 ≥95% probability (Fig. 4b). The variance contributed by each exonic, intronic, promoter, 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%
We then calculate the associations that explain at least 0.001% of the phenotypic variance attributed to all SNP markers. We then map SNPs to the closest gene using our functional annotations (Fig. 3a). Remaining snps are labelled as located in a cis-region (up to 50 kb from the SNP position and labelled them as located in a coding region, an intron, 1 kb upstream of a gene or 50 kb outside of the exon and promotor). 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. We calculated the phenotypic variance contributed by exonic, intronic, promoter region and SNPs +/−50 kb outside of the exon and promotor regions (cis) for each gene. Bar plots show the correlation among the variance explained by the groups across genes. Error bars show the SD.

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 (Fig. 4d). With the exception of height, there was no evidence for a relationship among the following groups: (i) SNPs in the exons of each gene and SNPs +/−50 kb outside of the exon and promotor regions; (ii) SNPs in the exons of each gene and SNPs in proximal promotor; and (iii) intronic SNPs and SNPs in promotor regions; (ii) SNPs in the exons of each gene and SNPs in proximal 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 largely independent of the effects of SNPs in their closest exon, as they do not align in terms of the variance they explain (Fig. 4d). 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 (Fig. 4d). 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-level and gene-level associations, 360
Table 2 Summary of findings for height (HT), body mass index (BMI), type-2 diabetes (T2D) and cardiovascular disease (CAD).

| Findings                  | Method               | HT   | BMI  | CAD  | T2D  |
|---------------------------|----------------------|------|------|------|------|
| Associated SNPs           | COJO-plink2          | 1673 | 517  | 34   | 85   |
|                           | COJO                 | 2131 | 565  | 34   | 84   |
|                           | BoltLMM              | 2134 | 555  | 34   | 82   |
|                           | COJO-Regenie         | 1660 | 520  | 70   | 87   |
| 50 kb regions             | BayesRR-RC           | 2578 | 2956 | 1478 | 1581 |
| (PPWV ≥ 95%)              |                      |      |      |      |      |
| Genic regions             | BayesRR-RC           | 2578 | 2956 | 1478 | 1581 |
| (PPWV ≥ 95%)              |                      |      |      |      |      |
| Exons                     |                      | 243  | 29   | 5    | 13   |
| Introns                   |                      | 1072 | 1162 | 307  | 347  |
| cis                       |                      | 1254 | 1765 | 1166 | 1221 |
| SNPs (PIP ≥ 95%)          | BayesRR-RC           | 360  | 20   | 2    | 9    |
| Exons                     |                      | 216  | 16   | 1    | 4    |
| Introns                   |                      | 73   | 2    | 1    | 5    |
| 10-500 kb                 |                      | 48   | 1    | 0    | 0    |
| LD clumps with r² = 0.1  | BayesRR-RC           | 1220 | 206  | 16   | 19   |
| (PPWV ≥ 95%)              |                      |      |      |      |      |

SNPs are top loci with a p-value < 5 × 10⁻⁸ from the fastGWAS UK Biobank summary statistic data for standing height, BMI, angina/heart attack and type-2 diabetes (fastGWA, see “Code availability”). This highlights that selecting on the top SNPs markers identified by standard association studies would give a different set of variants than those obtained from selecting high PIP SNPs.

Out-of-sample prediction into another European healthcare system. We 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 compared to recently proposed summary statistic approaches, 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 (Supplementary Fig. 14a). The area under the receiver operator curve (AUC) for T2D was 0.67 and 0.57 for CAD. In comparison, using the 64 BLD-LDAK annotations recommended by a recent study, the highest prediction accuracy obtained from MegaPRS was 0.55 for height, 0.32 for BMI, 0.10 for T2D, and 0.05 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 10–500 kb upstream regions contribute proportionally more to the prediction accuracy than other genomic groups, replicating our results from the UK Biobank (Supplementary Fig. 14). We find evidence for zero/low correlations of genomic predictors created from different annotation groups, which supports our results from the UK Biobank (Supplementary Fig. 14e). 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.

Additionally, for height and BMI we also determined the proportion of the posterior predictive distribution for each individual that was within +/-1 SD of their true phenotypic value. On average 67.5% of an individuals posterior predictive distribution is within +/-1 SD of their true phenotype and 75% for height, with similar prediction accuracy across individuals (Supplementary Fig. 14e). For T2D and CAD, we extended the PCF metric, typically defined as the proportion of cases with larger estimated risk than the top pth 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 (Supplementary Fig. 14d). We find 25 individuals for T2D and 15 individuals for CAD where ≥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, giving an odds ratio of 20 and 18 between the ≥90% and ≤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 ≤50% of their distribution within the high-risk category. These results highlight how variation contained within a posterior predictive distribution that is typically ignored in human genomic prediction can be used. We show that genomic prediction for personalised medicine with patient-specific predictions or stratification of patients is currently extremely limited.

Discussion

There is no single statistical model appropriate for all settings and thus there will always be a situation where a model poorly fits the data. We have provided theoretical and empirical evidence that a grouped Dirac spike-and-slab model (which we term BayesRR-RC), has a prior that is flexible enough to show robust model performance across the data analysed here, improving inference in many settings over commonly applied approaches. 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. This has enabled us to compare and contrast the inferred underlying genetic distribution for four complex phenotypes under this prior, providing novel insight into the genetic architecture of these traits. 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.

There has been debate on how to best estimate SNP heritability and here we validate that one approach could be to split SNPs markers by LD to improve genetic effect size estimates. 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 found 29 independent coding, 2888 independent intronic, and 5406 independent cis regions with ≥95% probability of contributing at least 0.001% of the SNP heritability. Understand how these coding, intronic and proximal and distal regulatory regions combine to contribute to...
phenotypic variance remains a substantial challenge 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\textsuperscript{24}, 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. Here, we present an approach for analysing large-scale biobank data, which is becoming increasingly available. However, a substantial number of GWAS have already been conducted, with associated published genome-wide summary association statistic estimates. Many methods have been developed to take advantage of these estimates, with downstream analysis models making use of various summary statistics resources in efficient and flexible ways. We show here that two leading summary statistic approaches perform poorly as compared to individual-level models for estimation of enrichment and genomic prediction. Despite this, the sample sizes obtained in consortia study meta-analyses will exceed those from single biobanks, especially for disease, and thus the genomic prediction accuracy of consortia study meta-analysis summary statistic prediction models may exceed those from individual-level analyses. Combining the posterior distribution obtained from BayesRR-RC across different individual-level biobank studies would alleviate this issue.

Additionally, 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. 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, then introns, and then to their nearest upstream gene position. These SNPs are most likely to be allocated accurately, with 1 and 1–10 kb 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 promoter, enhancer and tfbs regions assisting to some degree in alleviating this. Rather, 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. 10–500 kb distal regions may contribute more variance as marker density and marker coverage is higher in these regions, with missing variation within 10 kb upstream as causal variants are poorly correlated with SNPs. The posterior distributions for the variance explained by 1 kb, 1–10 kb regions, and 10–500 kb regions are negatively correlated (Supplementary Fig. 8, meaning that these groups are competing with each other, as if variance goes to one then it is being taken away from the other because 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 genomic regions are uncorrelated in their contribution to variance with the promoter and upstream regions still holds as does our probabilistic inference on the associations of each gene and their contribution to the phenotypic variation.

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 step toward maximising the probabilistic inference that can be obtained from large-scale Biobank studies.

**Methods**

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

\[ y = 1 + \frac{y_i}{\alpha} + \frac{y_i}{\gamma} \quad \alpha \quad \gamma \quad \beta \quad \epsilon, \]

(3)

where there is a single intercept term \( 1 \) and a single error term \( \epsilon \) but now SNPs are allocated into groups \( \{y_1, \ldots, y_d\} \), each of which having it’s own set of model parameters \( \beta_{\phi} = [\beta_{\phi 1}, \beta_{\phi 2}, \ldots, \beta_{\phi N}]^T \). As such, each \( \beta_{\phi} \) is distributed according to

\[
\beta_{\phi} \sim \pi_{\phi} \delta_0 + \pi_1 N(0, \sigma_{\phi 1}^2) + \pi_2 N(0, \sigma_{\phi 2}^2) + \ldots + \pi_p N(0, \sigma_{\phi p}^2),
\]

(4)

for each SNP marker group \( \{\pi_1, \pi_2, \ldots, \pi_p\} \) are the mixture proportions and \( \{\sigma_{\phi 1}^2, \sigma_{\phi 2}^2, \ldots, \sigma_{\phi p}^2\} \) are the mixture-specific variances proportion to

\[
\frac{\sigma_{\phi}^2}{C_1} = \frac{\sigma_{\phi}}{C_2} = \ldots = \frac{\sigma_{\phi}}{C_q},
\]

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 BayesR\textsuperscript{25} and BayesRS\textsuperscript{26}), 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 to only 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 \( \sigma_{\phi}^2 \) 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. In this work, we use 78 SNP marker groups, each with five mixture components (including 0).

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 as described in Supplementary Note 4 Equ. 35, which yields:

\[
f(a, y | \pi_{\phi}, \sigma_{\phi}, \{x, \gamma, \epsilon\}) \propto \exp \left\{ \frac{1}{2\sigma_{\phi}} ||y - Xa||^2 + \frac{1}{2\sigma_{\phi}} ||a||^2 - \log \left( \frac{1}{\pi_{\phi}} \right) ||y||_a \right\},
\]

(5)

where \( \pi_{\phi} \) are the group-specific mixture proportions and \( ||y||_a \) is the cardinality of the group. The corresponding MAP estimate would amount to adding extra penalisation on sparsity through the \( \pi_{\phi} \) terms, while keeping the same level of shrinkage as the baseline BayesR.

In our model the conditional posterior is:

\[
f(a, y | \pi_{\phi}, \sigma_{\phi}, \{x, \gamma, \epsilon\}) \propto \exp \left\{ \frac{1}{2\sigma_{\phi}} ||y - Xa||^2 + \frac{1}{2\sigma_{\phi}} ||a||^2 - \log \left( \frac{1}{\pi_{\phi}} \right) ||y||_a \right\}
\]

(6)

now each marker has a group-specific shrinkage \( \sigma_{\phi} \), which translates to a specific \( \lambda_{\phi} \) 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 \( \Phi \) 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 refs.\textsuperscript{16,27}
it is showed that the former case requires weaker conditions in order to recover the true vector \( \beta \) consistently than the latter. Although the sampling scheme was different, we have showed that a similar mean allele frequency to \( 0.25 \) as used by the authors and which matches the simulation setting. This model is intended to approximate an individual-level REMEL analysis with 78 annotations, but using a different scaling of the relationship matrix, and provides an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.

We used the software SunHer\(^5\). We calculated marker taggings under the same 78 component annotation model. We ignored the LD weights when calculating the taggings as we found this gave the best estimates we could obtain from the simulated data across all scenarios. We set the relationship of a similar model to mean allele frequency to \( 0.25 \) as used by the authors and which matches the simulation setting. This model is intended to approximate an individual-level REMEL analysis with 78 annotations, but using a different scaling of the relationship matrix, and provides an estimate of the variance attributable to SNPs genome-wide and an estimate of the variance attributable to SNP markers of each annotation group.

We ran the following prediction models, using a testing set of 10,000 UK Biobank unrelated individuals, that were also unrelated to the training data, and focusing on the models proposed in a recent paper\(^{21}\). These methods contain two approximations to our BayesRR-RC model and the authors claim to outperform all other existing methods, including individual-level models. The models are:

- An individual-level bayesR model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS\(^{21}\). This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- An individual-level boltREML model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS\(^{21}\). This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- A summary statistic boltREML model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS\(^{21}\). This provides estimates of the SNP marker effects for creating a genetic risk predictor.
- A summary statistic boltREML model using genomic annotation SNP variance estimates from the SumHer models as implemented in the software MegaPRS\(^{21}\). This provides estimates of the SNP marker effects for creating a genetic risk predictor.

First, we compared the correlation of the simulated and estimated proportion of polygenic variance attributable to the 13 genomic annotation groups across all models in Fig. 1. We determined the ability of the approaches to correctly identify enriched regions of the DNA by estimating the probability within each simulation replicate that a SNP marker group would have an estimated enrichment above a threshold (i.e., being described as having average effect sizes that are twice as large as expected) when the simulated value was \( \leq 1 \). As BayesRR-RC induces sparsity in the SNP effect estimates, with some markers always remaining in the variance = 0 spike, we propose a different enrichment definition where the proportion of \( h^2_{SNP} \) is divided by the proportion of markers that are in the model for the SNP group, rather than the proportion of markers mapping to the SNP group.

In Supplementary Note 3, we propose a posterior probability window variance (PPWV) approach\(^{26}\), which provides a probabilistic determination of association of a given LD block, genomic window, gene, or upstream region, relative to the annotation-specific polygenic variance explained by that window. The PPWV approach determines the posterior inclusion probability that each region and each gene contributes at least 0.001% to the \( h^2_{SNP} \), with theory and small-scale simulations outlined in Supplementary Note 3 suggesting well controlled FDR. We partitioned the 596,741 imputed SNP markers in LD blocks, defined as groups of markers with \( R^2 \geq 0.1 \). Within each simulation replicate, we estimated the probability that LD blocks containing a causal variant were identified by PPWV. We compared this to MLMA estimates obtained using the BoltLMM software, by estimating the probability that LD blocks containing a causal variant were identified as having a SNP with \( p \)-value \( \leq 5 \times 10^{-8} \), the standard genome-wide significance threshold. We present these results in Fig. 2a.

We then compare the prediction accuracy obtained in a testing set of 10,000 UK Biobank unrelated individuals, that were also unrelated to the training data. We predicted phenotype using SNP marker effect sizes obtained from BayesRR-RC, MLMA effect sizes from BoltLMM, and the four MegaPRS methods outlined above implemented in the LDAK software. While we would suggest that fixed-effect MLMA estimates are improper for prediction we include this comparison as polygenic risk scores have often been created from fixed-effect SNP estimates. We calculate the correlation between the simulated phenotype in the testing set and the genomic predictor within each simulation replicate and we compare the mean correlation across the 18 different genomic annotations in Fig. 2. Additionally, to provide a benchmark, we compare to the theoretical expectation under ridge regression approximations\(^{28}\), with the number of markers set to the number of causal variants.

Relationship between effect size, minor allele frequency and LD. We then conducted another large-scale, but this time well-powered simulation study, where we ascertained the causal variant SNP markers in different ways and varied the relationship of the phenotype effect size, minor allele frequency and LD. We used the same randomly selected 40,000 unrelated individuals and all 596,741 imputed (version 3) genomic markers from chromosomes 19 through 22 from the UK Biobank.
simulated a wide-range of different possible underlying genetic effect size distributions as follows:

- We chose either 5000 or 10,000 imputed SNP markers for which to assign a genetic effect size, providing two different levels of polygenicity.
- We selected these 5000 or 10,000 markers in two different ways. Either, we selected SNP markers randomly, or we selected the marker of highest minor allele frequency per LD block of the genome, with an LD block defined as a group of SNP markers with absolute LD of at least 0.05. Randomly allocating markers creates a set of associated variants with the same distribution of LD and MAF as the SNP data, which is composed of predominantly low frequency variants. Selecting only the highest frequency marker per LD block creates a setting where for each set of markers in LD with each other, there is only one causal genetic variant, and where the distribution of associated markers differs to that of the SNP markers as a whole.
- Having created four different ways of selecting associated markers (5000 or 10,000 and high-MAF or random) we then created five different ways of assigning effect sizes to them:
  - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) with no relationship to the LD or MAF of the markers. Thus, effects had variance \( \alpha_0(N, \sigma^2(p(1-p)^9) \) with \( \lambda \) the LD score of the marker and \( p \) the allele frequency.
  - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) \( \alpha_0(N, \sigma^2(p(1-p)^9) \). This simulates stronger effect sizes for rare variants and those in low LD.
  - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) \( \alpha_0(N, \sigma^2(p(1-p)^9) \). This simulates stronger effect sizes for rare variants and higher LD.
  - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) \( \alpha_0(N, \sigma^2(p(1-p)^9) \). This simulates equivalent effect sizes for common and rare variants, and greater effects for markers in low LD.
  - We simulated effect sizes from a normal distribution with zero mean and variance 0.6 divided by the number of markers (5000 or 10,000) \( \alpha_0(N, \sigma^2(p(1-p)^9) \). This simulates equivalent effect sizes for common and rare variants, and greater effects for markers in high LD.

- For each of the four different sets of markers, each with five different effect size sampling schemes, we then created two additional settings. In the first setting markers were sampled from the various normal distribution, as described above, for the five different effect size sampling schemes. In the second setting, for each of the five different effect size sampling schemes, we simulated effects from 13 different distributions, one for each of 13 different genomic annotation groups with different proportions of total SNP heritability (\( \mathcal{H}_{SNV} \)). For each of the five different effect size sampling schemes, we partitioned the LD to MAF and SNPS maintained the same, but the total variance attributed to the SNP markers was partitioned across annotation groups as follows for exonic variants (\( \mathcal{H}_{SNV} = 0.1 \)), intrinsic variants (\( \mathcal{H}_{SNV} = 0.2 \)), 1 kb promoter variants (\( \mathcal{H}_{SNV} = 0.05 \)), 1-10 kb enhancers (\( \mathcal{H}_{SNV} = 0.025 \)), 10-50 kb enhancers (\( \mathcal{H}_{SNV} = 0.05 \)), 10-500 kb transcription factor binding sites (\( \mathcal{H}_{SNV} = 0.05 \)), 10-500 kb other variants (\( \mathcal{H}_{SNV} = 0 \)), 500 kb-1 Mb enhancers (\( \mathcal{H}_{SNV} = 0.05 \)), 500 kb-1 Mb transcription factor binding sites (\( \mathcal{H}_{SNV} = 0.05 \)), 500 kb-1 Mb other variants (\( \mathcal{H}_{SNV} = 0 \)), and other non-annotated SNPs (\( \mathcal{H}_{SNV} = 0 \)). Four of these distributions had zero variance indicating that no associations were created for these groups. In the first setting, this simulates variance explained by annotation groups that is on average proportional to the number of SNPs within each annotation. In the second scheme, the variance and average effect size differs across annotation groups. We refer to these as two different enrichment settings: “random”, or “enriched”.
- This created four different sets of associated markers (5000 or 10,000 and high-MAF or random), each with five different marker effect size sampling schemes, which we refer to in the main text as the 20 different generative genetic models (Table 1), each of which has two enrichment settings. This gave 40 different sampling settings that were effects that we simulated ten replicates for each setting, giving a total set of 400 simulated phenotypes.
- For each generative model the total genetic variance was 0.6 and we sampled individual-level environmental (residual) variance from a normal distribution with zero mean and variance 0.4 to give phenotypes with zero mean and unit variance.

This range covers generative genetic models discussed in the literature and provide a platform for comparing the performance of different methods without violating the assumptions of the random variance component statistical models. This includes both individual-level and summary statistic approaches, that are currently applied in the literature for estimation of the variance attributable to the SNP markers, for testing association of genetic markers with phenomena genome-wide, and for genomic prediction. Re-simulation provided a set of different scenarios, which we can explore the model performance of BayesRR-RC and compare it to existing approaches. In Supplementary Fig. 1, we compare the \( h_{SNP}^2 \) estimation, estimation of the summary genetic variance along with the RMSE of the estimates, and the estimated average effect size.

First, we calculate z-scores of the marker effect estimates from their true simulated value. As MLMA approaches estimate marker effects one-at-a-time, we calculated the z-score of the estimate from the true simulated value for the causal variants in each simulation replicate, across generative genetic models. For the Bayesian methods, at any one iteration, LD among the markers is controlled for (see Supplementary Note 4). However across iterations as the chain mixes, markers in LD will enter and leave the model, with their posterior inclusion probabilities reflecting their association with the trait. Thus, we summed the squared regression coefficient estimates of SNPs in the model at each iteration for each LD block (markers in LD \( R^2 \geq 0.1 \) within 1 MB) of each simulation replicate, took the mean across iterations and then calculated the z-score estimate from the simulated value. This metric provides an assessment of the ability of BayesRR-RC to accurately estimate the contribution of a genomic region to the phenotypic variance as compared to MLMA approaches. We present these results in Supplementary Fig. 2, where we find that the z-scores of the estimated BayesRR-RC effects are generally stable across generative genetic models and comparable to those obtained from BayesR with but with slightly elevated variance in many cases as the model is less sparse (Supplementary Fig. 2a). We find that SNP effect size estimates from MLMA models have higher estimation error, especially when the causal variant is rare, or in high-LD with many other SNPs (Supplementary Fig. 2a). (see Supplementary Note 4). However across iterations as the chain mixes, markers in LD will enter and leave the model, with their posterior inclusion probabilities reflecting their association with the trait. Thus, we summed the squared regression coefficient estimates of SNPs in the model at each iteration for each LD block (markers in LD \( R^2 \geq 0.1 \) within 1 MB) of each simulation replicate, took the mean across iterations and then calculated the z-score estimate from the simulated value. This metric provides an assessment of the ability of BayesRR-RC to accurately estimate the contribution of a genomic region to the phenotypic variance as compared to MLMA approaches. We present these results in Supplementary Fig. 2, where we find that the z-scores of the estimated BayesRR-RC effects are generally stable across generative genetic models and comparable to those obtained from BayesR with but with slightly elevated variance in many cases as the model is less sparse (Supplementary Fig. 2a). We find that SNP effect size estimates from MLMA models have higher estimation error, especially when the causal variant is rare, or in high-LD with many other SNPs (Supplementary Fig. 2a). (see Supplementary Note 4). However across iterations as the chain mixes, markers in LD will enter and leave the model, with their posterior inclusion probabilities reflecting their association with the trait. Thus, we summed the squared regression coefficient estimates of SNPs in the model at each iteration for each LD block (markers in LD \( R^2 \geq 0.1 \) within 1 MB) of each simulation replicate, took the mean across iterations and then calculated the z-score estimate from the simulated value. This metric provides an assessment of the ability of BayesRR-RC to accurately estimate the contribution of a genomic region to the phenotypic variance as compared to MLMA approaches. We present these results in Supplementary Fig. 2, where we find that the z-scores of the estimated BayesRR-RC effects are generally stable across generative genetic models and comparable to those obtained from BayesR with but with slightly elevated variance in many cases as the model is less sparse (Supplementary Fig. 2a). We find that SNP effect size estimates from MLMA models have higher estimation error, especially when the causal variant is rare, or in high-LD with many other SNPs (Supplementary Fig. 2a).
within the 10,000 UK Biobank individual selected for out-of-sample prediction. We also used the MegaPRS methods implemented in the software LDAK running the four different models described above. We compute the correlation of predicted and simulated genetic value across approaches for each of the 400 simulated phenotypes (Supplementary Fig. 2d).

The influence of population structure and relatedness. We then investigated the importance of controlling for multicollinearity for the control of population genetic and disease-principle. A MAF-based PPR approach will control for correlated markers (either local or long-range LD) fitted as random when testing for the effects of a focal SNP. For two markers, \(X_1\) and \(X_2\) in LD correlation \(\rho_{X_1,X_2}\), with \(\beta_2 = 0\) we can express the MLMaMA effect solution as a partial regression coefficient of the phenotype regressed onto the focal SNP after adjusting for \(X_1\) as

\[
\hat{\beta}_{X_1,X_2} = \frac{\sum_{i=1}^{n} (X_1[i] - \bar{X}_1) (X_2[i] - \bar{X}_2) (y[i] - \bar{y})}{\sum_{i=1}^{n} (X_1[i] - \bar{X}_1)^2}
\]

Following our derivation above for a shrinkage estimator of a partial regression coefficient the effect size of \(X_1\) is estimated as

\[
\hat{\beta}_{X_1,X_2} = \frac{\sum_{i=1}^{n} (X_1[i] - \bar{X}_1) (X_2[i] - \bar{X}_2) (y[i] - \bar{y})}{\sum_{i=1}^{n} (X_1[i] - \bar{X}_1)^2}
\]

and in this two-SNP example the bias is accounted for in the term

\[
\frac{\rho_{X_1,X_2}}{1 - \rho_{X_1,X_2}}
\]

when the fixed effect is estimated. Multicollinearity acts to increase the \(\sigma_1^2\) term of \(\lambda\), reducing the denominator \(X_1'X_1 + \lambda I\) in the estimation of \(\hat{\beta}_{X_1,X_2}\), and increasing the variance of the estimates of common markers in high LD, those with the highest average \(F_{ST}\).

We conducted a simulation study using real genomic data from chromosome 22 where 10,000 individuals were selected from two UK Biobank assessment centres (Glasgow, Croydon). Croydon individuals were assigned to 5000 1000 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 the effect size, but also the direction of effects (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 scenarios where the phenotype variance attributable to the causal SNPs is 0.5, there is a phenotypic differ in which 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 (MLMai in GCTA) and a Bayesian implementation of the MLM approach to identify variant associated regions, irrespective of the genetic and relatedness and with increasing family effects.

Localisation and fine-mapping of SNP phenotype associations. We further validate the use of PPWV in another simulation study with 500 replicate data sets of 10,000 SNP markers for 5000 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 range from 0.04 SD, to 0.15 SD with 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 SNP group. For the first 1000 replicates of each scenario, we take the mean PPWV across 100 samples for each of the 1000 different effect sizes and compare these in Fig. 5a. 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 for each of the iterations of the simulated data. We then applied 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 50 kb regions identified that do not contain a causal variant, at PPWV thresholds ranging from 0.8 to 100%. We compare these in Supplementary Fig. 6b where as we lower the PPWV variance threshold, the number of false discoveries in the model increases but remains at ≤5% when the PPWV is ≤55%. This further demonstrates that our proposed PPWV approach is an appropriate metric of summarising the posterior distribution to identify associated genomic regions, irrespective of the genetic region used.

We also focused on the ability of our approach to fine-map associated regions to identify candidate SNPs and to provide a probabilistic assessment of the most likely associated set of SNP markers. To do this we used our large-scale simulation data and focused on seven focal regions within a blocks of chromosome 22. We allocated effect sizes to the following SNPs: rs131529 with MAF 0.32 which had LD \(r^2 = 0.15\) with 348 other SNPs, rs2096637 with MAF 0.14 which had LD \(r^2 = 0.15\) with 5 other SNPs, rs1131538 with MAF 0.05 which had LD \(r^2 = 0.15\) with 12 other SNPs, rs114962840 with MAF 0.007 which had LD \(r^2 = 0.15\) with 11 other SNPs, rs117873986 with MAF 0.02 which had LD \(r^2 = 0.15\) with 12 other SNPs, and rs9606483 with MAF 0.005 which had LD \(r^2 = 0.15\) with 3 other SNPs, and rs20966462 with MAF 0.005 which had LD \(r^2 = 0.15\) with 1 other SNP. To these scenarios, we assigned the four effect sizes in four different scenarios, \(\sigma_3 = 0.05, 0.025, 0.0125,\) or 0.01 on the SD scale. On the remainder of chromosomes 19, 20, 21 and 22, we randomly selected 1000 SNPs as causal variants to give a polygenic background, sampling their effects from a normal distribution with zero mean and variance 0.5/1000. We repeated each of the four scenarios 20 times. We selected these regions to compare the performance of BayesRR-RC to the fine-mapping approach SuSiE as outlined in a recent paper\(^22\). For BayesRR-RC, we calculate the PPWV of the LD blocks containing the seven focal SNPs, and then prune these blocks based on the LD among the markers in the block (described as ‘purity’ in the SuSiE paper\(^22\)) to identify a credible set with LD \(r^2 = 0.9\). We then count the proportion of times across the simulations that each causal variant was contained of one of the credible sets. For SuSiE, we ran the model from the individual-level data of the whole block of chromosome 22 using the suggested settings and setting \(K = 10\). We then calculate the proportion of times that the identified credible sets contained one of the seven causal variants. We present these results in Supplementary Fig. 6c.

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 2 ancestry (UK Biobank data code 22000-0) also. We also used the imputed autosomal genotype data of the UK Biobank provided as part of the 1000 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 PC projection \(<\) absolute value 4 and PC 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. We tested the BayesRR model for an additional relatedness 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 phenotype 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 0.09 as missing. From the good quality markers (with missingness less than 5% and \(p\)-value for Hardy–Wright 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 these, we took the overlap with the Estonian Genome data to a data set of 8,490,466 variants. For 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; (ii) were identified to have putative sex chromosome aneuploidy; (iv) were excluded from follow-up if we had no information on their location annotations (Supplementary Data 1). If 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 2002-0.0) and compared our BMI and height with the standard BMI and height we used to remove the BMI and height away from the mean were not included in the analyses. For Type 2 Diabetes (T2D) in UKB, we selected cases very broadly as individuals who have main or secondary disease (UKB fields 41202-0.0 – 41202-0.379 and 41204-0.0 – 41204-0.434) of “non-insulin dependent diabetes, diabetes” (ICD code E11) or self-reported non-colon cancer (UKB field 20002-0.0 – 20002.28) “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 128; self-reported angina (code 1074) or self-reported heart attack/myocardial infarction (code 1075). Participants with self-reported “heart/cardiac problem” (code 1086) were not included as cases but also excluded from controls. This gave a sample set 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. UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants. We preferentially assigned SNPs to coding (exonic) regions, then to 1 kb upstream regions, then to 10 kb regions, then to 1 Mb regions. Remaining SNPs were grouped in a category labelled “remaining SNPs”.

We partition SNP markers into seven location annotations using the knownGene table from the UCSC browser data (see “Code availability” section). 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 2002-0.0) and compared our BMI and height with the standard BMI and height we used to remove the BMI and height away from the mean were not included in the analyses. For Type 2 Diabetes (T2D) in UKB, we selected cases very broadly as individuals who have main or secondary disease (UKB fields 41202-0.0 – 41202-0.379 and 41204-0.0 – 41204-0.434) of “non-insulin dependent diabetes, diabetes” (ICD code E11) or self-reported non-colon cancer (UKB field 20002-0.0 – 20002.28) “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 128; self-reported angina (code 1074) or self-reported heart attack/myocardial infarction (code 1075). Participants with self-reported “heart/cardiac problem” (code 1086) were not included as cases but also excluded from controls. This gave a sample set 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.

These are labelled as located in a coding region, an intron, 1 kb upstream of a gene using our functional annotations. Remaining SNPs are labelled as located in a cis-
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, we 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. We 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 \( p < 0.1 \) 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 than the top \( p \)-value 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. As a comparison, we also estimated a bolt-LMM prediction model using MegaPRS\(^2\) as recommended by the authors and as shown to have the best prediction performance out of the MegaPRS approaches in our simulation study. We clumped SNPs with \( r^2 \) threshold of 0.5 resulting in 1,508,624 SNP markers to be included in the analysis and randomly selected 20,000 individuals to compute the LDW weight. We then computed the tagging file using the same data set as reference and the 64 BLD-LDAK annotations. Here, weights are models as an extra annotation and we save the heritability matrix. We then regress the plink\(^2\) summary statistics for height, BMI, CAD and T2D onto the tagging file, saving the per-predictor heritabilities. We then created four reference panels with the same 1,508,624 SNP markers but randomly selecting different 5000 related individuals from the UK Biobank and we used these to: (i) calculate predictor-predictor correlations with a window size of 3000 kb to estimate the LD structure; (ii) use in the construction of the prediction model; (iii) estimate effect sizes specifying a Bolt-LMM model for height, BMI, CAD and T2D, using the predictor-predictor correlations, the per-predictor heritabilities, the plink\(^2\) summary statistics and training pseudo summary statistics, whilst including ambiguous allele and specific variance available due to the per-predictor distributions to determine the most accurate model and obtain the best effect sizes. These effects were then calculated for each individual in the Estonian Biobank using the best effect sizes. We report the squared correlations between the genomic predictor and phenotypes.

**Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

This project uses UK Biobank data under project 35520. The Estonian Genome Centre data are protected and are not available due to data privacy laws. The Estonian Genome Centre data can be made available under restricted access upon request from the cohort author R.M. with appropriate research agreements. Summaries of all posterior distributions generated in this study are provided in Supplementary Data tables. Full posterior distributions of the SNP marker effects sizes and estimated variance components for each trait are deposited on Dryad with https://doi.org/10.5061/dryad.sq9s4n51.

**Code availability**

Our BayesRR-RC model is implemented within the software GMRM, with full open code source available at: https://github.com/medical-genomics-gmrn/gmrn. UCSC Table Browser https://genome.ucsc.edu/cgi-bin/hgTables. flashPCA https://github.com/gebhr/flashspca. Plink\(^1\) 90 https://www.cog-genomics.org/plink2/. GCTA https://www.cog-genomics.org/gcta/. snp2tfbs database https://ccg.epfl.ch/snp2tfbs. fastGWAS database https://fastgwa.info/ukbimph2phenotypes/. Computing environment https://www.epfl.ch/research/facilities/scitas/hardware/helvetios/.

Received: 10 February 2021; Accepted: 5 November 2021; Published online: 30 November 2021

**References**

1. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* 50, 737–745 (2018).
2. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* 49, 986 (2017).
3. Speed, D., Holmes, J. & Bolding, D. J. Improving and evaluating heritability models using summary statistics. *Nat. Genet.* 52, 458–462 (2020).
4. Hou, K. et al. Accurate estimation of SNP heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* 51, 1244–1251 (2019).
5. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling s-lDc and ldak functional enrichment estimates. *Nat. Genet.* 51, 1202–1204 (2019).
6. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* 51, 277–284 (2019).
7. Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* 51, 1749–1755 (2019).
8. Loh, P.-R. et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284 (2015).
9. Mbatchou, J. et al. Computationally Efficient Whole Genome Regression for Quantitative and Binary Traits (Nature Publishing Group, 2020).
10. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341 (2018).
11. Erbe, M. et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129 (2012).
12. Moser, G. et al. Simultaneous discovery, estimation and prediction analysis of genome-wide association summary statistics. *PLoS Genet.* 11, 1–22 (2015).
13. Banos, D. T. et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat. Commun.* 11, 1–4 (2020).
14. George, E. I. & McCulloch, R. E. Variable selection via gibbs sampling. *J. Am. Stat. Assoc.* 88, 881–889 (1993).
15. Malsiner-Walli, G. & Wagner, H. Comparing spike and slab priors for bayesian variable selection. *Austrian J. Stat.* 40, 241–264 (2016).
16. Castillo, I. et al. Bayesian linear regression with sparse priors. *Ann. Stat.* 43, 1986–2015 (2015).
17. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* 47, 1385 (2015).
18. Pazokitorouidi, A. et al. Efficient variance components analysis across millions of genotypes. *Nat Commun* 11, 4020 (2020).
19. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235 (2015).
20. Fernando, R., Toosi, A., Wolc, A., Garrick, D. & Dekkers, J. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. *J. Agric. Biol. Environ. Stat.* 22, 172–193 (2017).
21. Zhang, Q., Privé, F., Vilhjálmsson, B. J. et al. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Commun* 12, 4192 (2021).
22. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. 82*, 1273–1300 (2020).
23. Lloyd-Jones, L. R. et al. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.* 10, 5086 (2019).
24. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits from summary statistics. *PLoS Genet.* 14 (2018).
25. MacLeod, I. M. et al. Exploiting biological priors and sequence variants in whole-genome prediction methods for genome-wide association studies: a bayesian approach. *Genet. Epidemiol.* 35, 378–389 (2012).
26. Purcell, S. et al. A tool for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
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. P.M.V. acknowledges funding from the Australian National Health and Medical Research Council (1113400) and the Australian Research Council (FL180100072). L.R. 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.

Author contributions
M.R.R. conceived and designed the study. M.P., D.T.B. and A.K. contributed to the study design. M.P. and M.R.R. conducted the experiments and analyses with input from D.T.B., A.K., S.E.O., A.H., J.S., P.M.V., R.M. and L.R. M.R.R., D.T.B., S.E.O. and L.R. derived the equations and the algorithm. EJO and DTB developed the software, with contributions from M.R.R., M.P., S.E.O., A.K. and G.M. M.R.R., M.P. and DTB wrote the paper. RM and ZK provided study oversight and contributed data to the analysis. All authors approved the final manuscript prior to submission.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27258-9.

Correspondence and requests for materials should be addressed to Matthew R. Robinson.

Peer review information Nature Communications thanks Luke O’Connor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.