Supplementary Information for

Improving GWAS discovery and genomic prediction accuracy in biobank data

Orliac EJ, Trejo Banos D, Ojavee SE, Läll K, Mägi R, Visscher PM, Robinson MR.

Matthew R. Robinson.
E-mail: matthew.robinson@ist.ac.at

This PDF file includes:

Figs. S1 to S9
Table S1

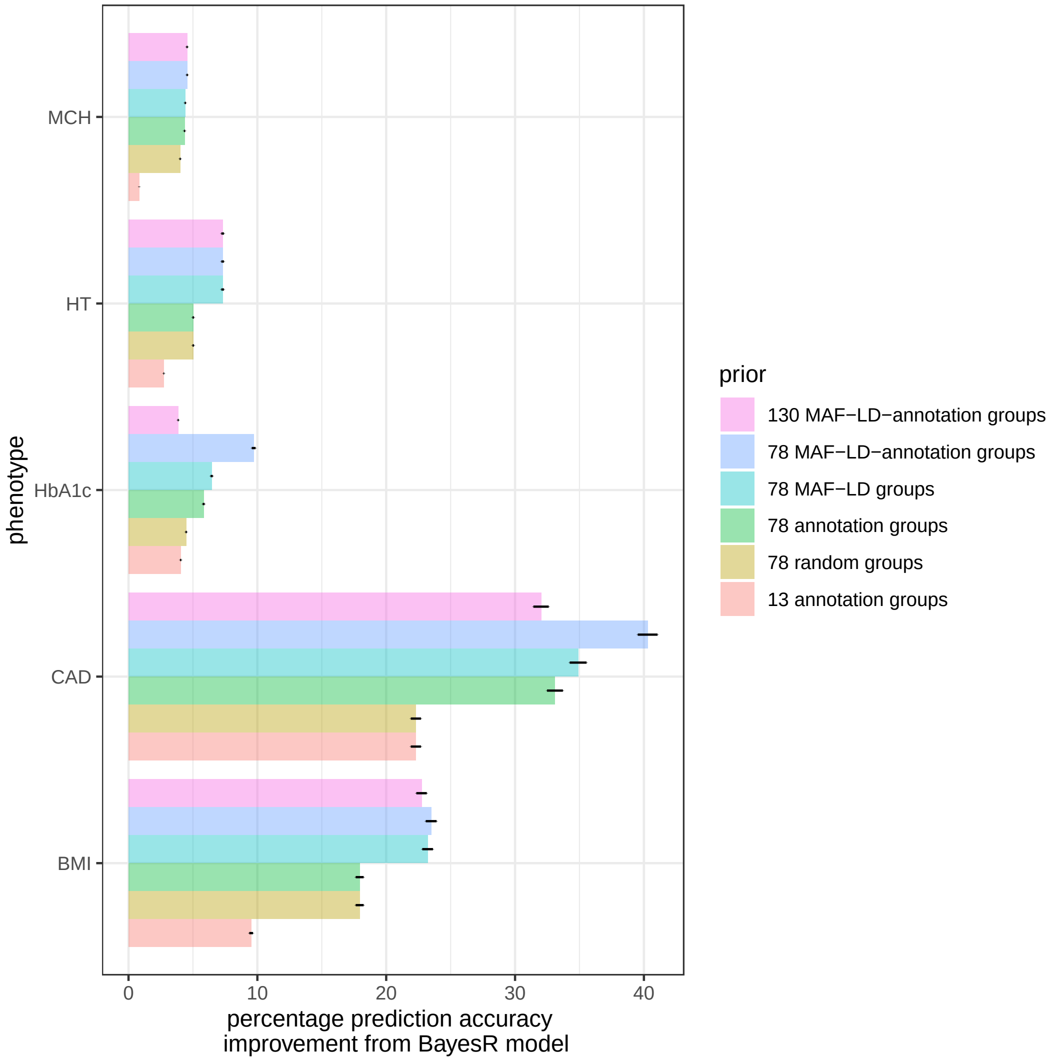

Fig. S1. Selection of priors for the GMRM. Prediction accuracy in a UK Biobank hold-out set, as a percentage improvement over a BayesR model for different prior choices. The prior: "13 annotation groups" represents splitting SNP markers into 13 annotation groups, with effect sizes modelled using a mixture of four normal distributions and a dirac delta spike at zero for each group. "78 random groups" represents a prior of markers grouped randomly into 78 groups, with effect sizes of each group modelled using a mixture of four normal distributions and a dirac delta spike at zero. "78 annotation groups" represents a prior of markers split into 13 annotation groups, with each annotation group further divided into 6 random groups, to give 78 groups in total, where the marker effects of each group are modelled using a mixture of four normal distributions and a dirac delta spike at zero. "78 MAF-LD groups" represents a prior of markers split into 39 groups based on their MAF, with each group further sub-divided into two groups based on the LD Score of the markers, where the marker effects of each group are modelled using a mixture of four normal distributions and a dirac delta spike at zero. "78 MAF-LD-annotation groups" represents the prior presented in the main text of markers divided into 13 annotation groups, with each group further sub-divided into six based on MAF and LD Score, where the marker effects of each group are modelled using a mixture of four normal distributions and a dirac delta spike at zero. Finally, "130 MAF-LD-annotation groups" represents a prior of markers divided into 13 annotation groups, with each group further sub-divided into ten based on MAF and LD Score, where the marker effects of each group are modelled using a mixture of four normal distributions and a dirac delta spike at zero. Error bars give the 95% confidence intervals of the estimate.

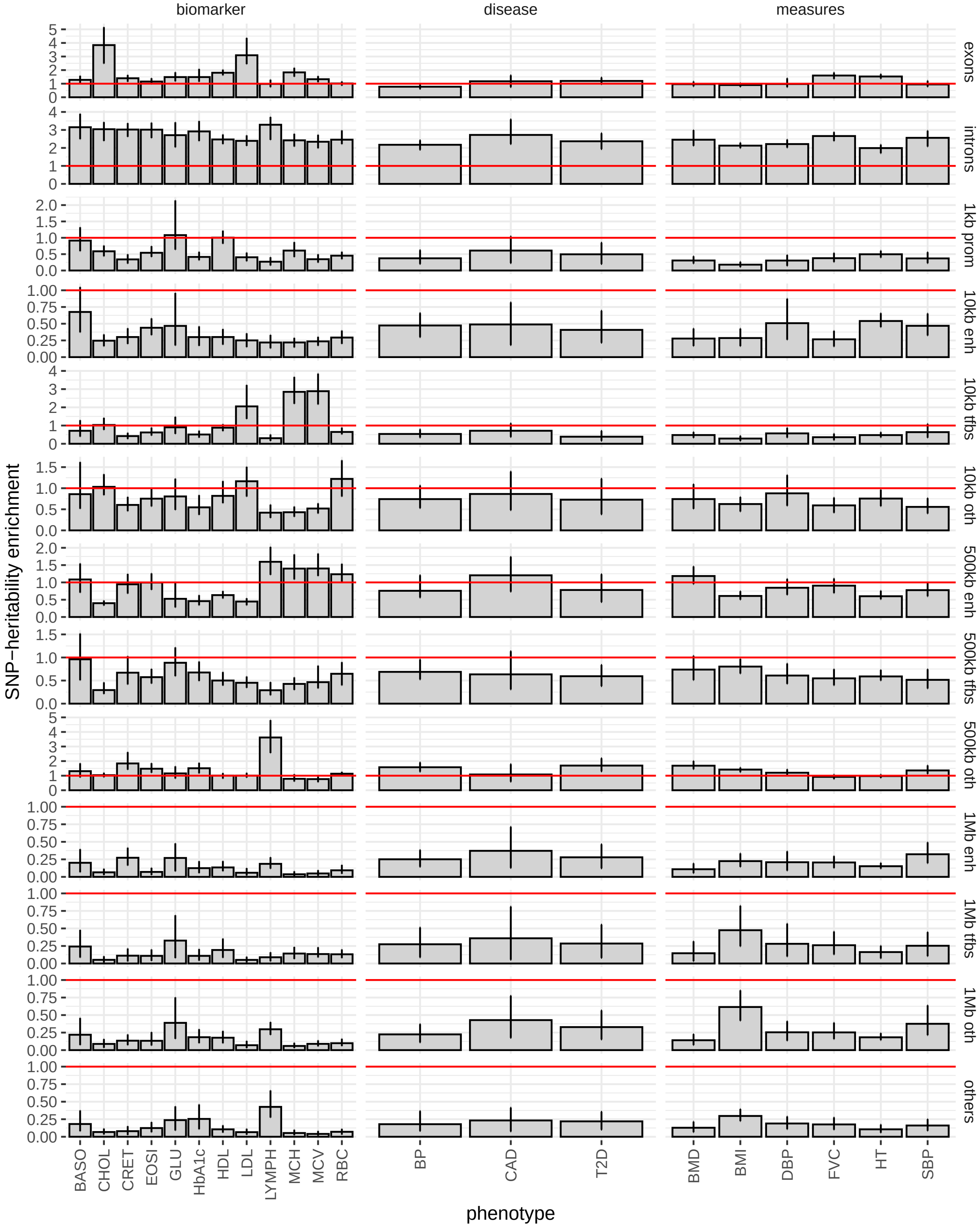

Fig. S2. SNP-heritability enrichment estimated by GMRM. SNP heritability enrichment estimates calculated as the proportion of SNP-heritability attributable to each annotation group divided by the proportion of markers in the model given the total number of markers. If the average effect sizes of markers within a given annotation are larger than expected given the number of markers entering the model for that annotation then the value obtained should be greater than 1 (the red line shown). Conversely, smaller than expected marker effects will yield values less than 1.Error bars in give 95% credible intervals. Full trait codes are given in Supplementary Table 1. exons = SNPs located in exonic regions, introns = SNPs located in intronic regions, prom = SNPs located in promotors, tfbs = SNPs located in transcription factor binding sites, enh = enhancers, oth = markers not mapping to a functional group but to a location from a nearest gene, others = markers not within 1Mb of a gene and that have no known regulatory function.

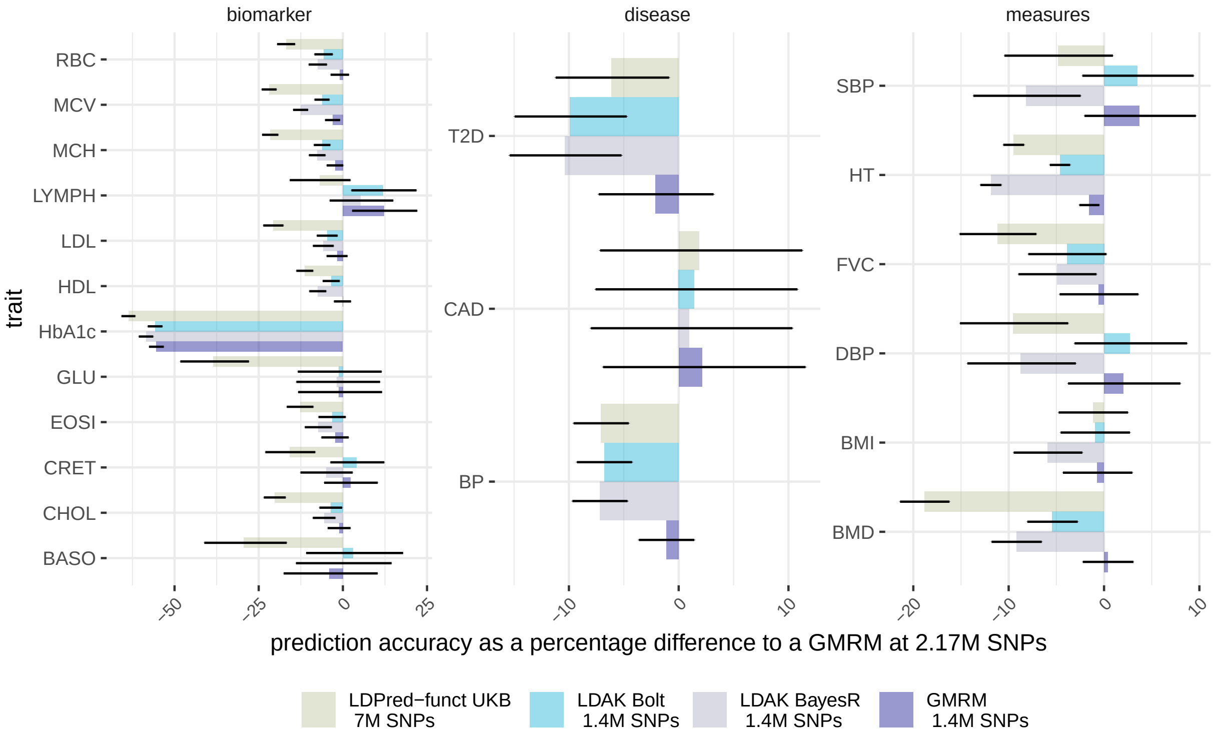

Fig. S3. Prediction accuracy of a GMRM as compared to other approaches at different sets of genetic markers. Prediction accuracy in a UK Biobank hold-out set of different approaches and SNP marker sets as compared to a GMRM at 2.17M SNPs. "GMRM 1.4M SNPs": a GMRM model with 56 MAF-LD-annotation groups at 1,410,525 common SNPs of MAF ≥ 0.01. "LDAK BayesR 1.4M SNPs": an individual-level LDAK implemented BayesR model with BLD-LDAK annotations at 1,410,525 common SNPs of MAF ≥ 0.01. "LDAK Bolt 1.4M SNPs": an individual-level LDAK software implemented Bolt model with BLD-LDAK annotations at 1,410,525 common SNPs of MAF ≥ 0.01. "LDPred-funct UKB" an LDPred-funct model with UK Biobank LD Score annotations for 6,991,095 SNPs. Error bars give the 95% confidence interval of the estimate.

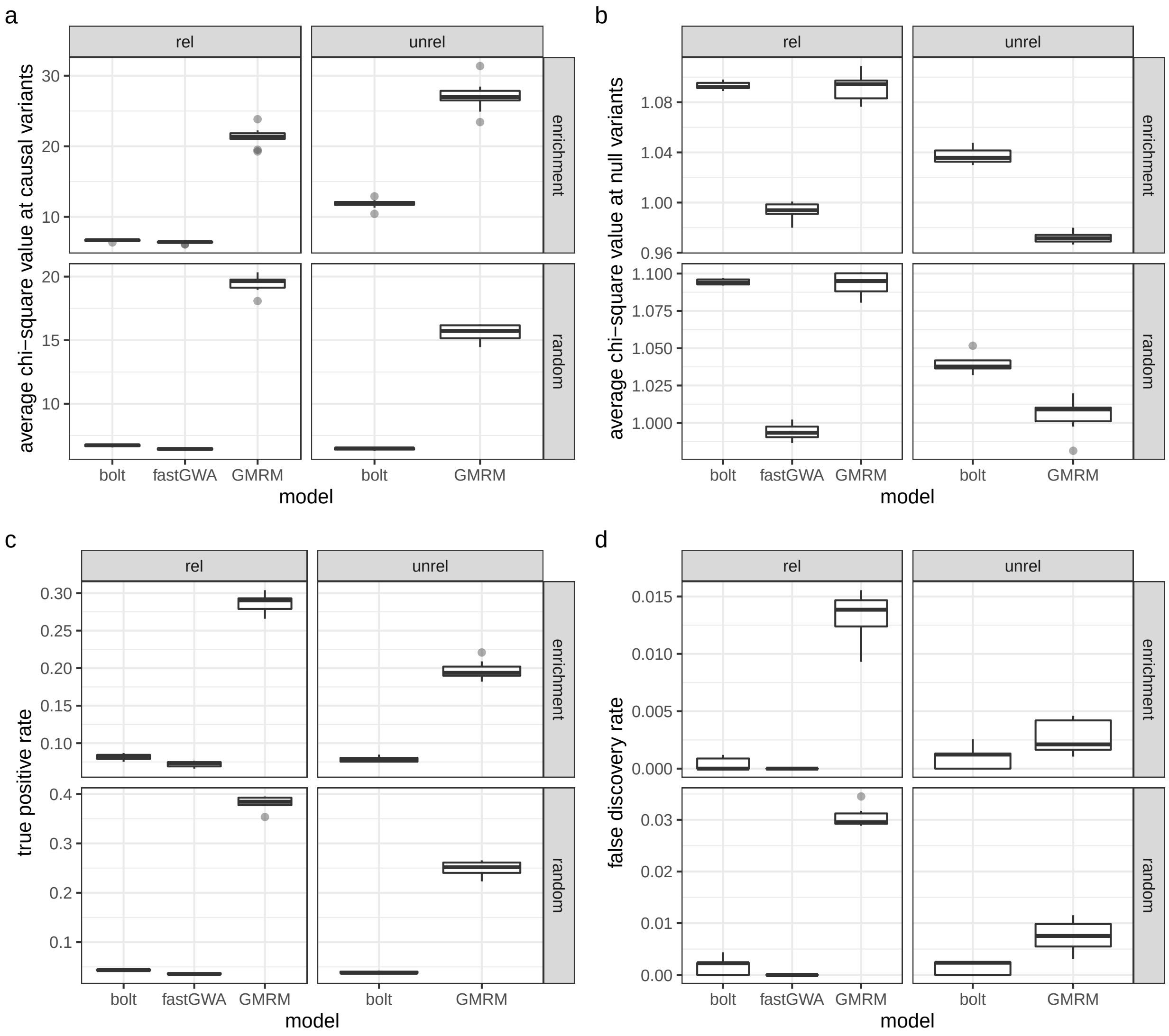

Fig. S4. Simulation study for GMRM. (a) The y-axis gives the average χ^2 values for the approximate mixed linear model test statistics at the simulated causal variants as compared to models run using fastGWA and bolt-LMM. The plot is faceted into two columns representing phenotypes that were simulated using either a mixture of 10,000 randomly sampled UK Biobank sibling pairs and 80,000 randomly sampled unrelated UK Biobank individuals (related), or 100,000 randomly sampled UK Biobank individuals with SNP marker relatedness estimates <0.2 (unrelated). The plot is also faceted into two rows representing $10{,}000$ causal variants from chromosomes 1, 3, 5, 7 and 9 that were either randomly sampled with effect sizes drawn from a normal distribution with zero mean variance $0.5/10000$ (random) or were sampled differently across genomic annotations creating effects size differences across genomic groups (see Methods). Boxplots show the distribution of values obtained for ten simulation replicates. (b) shows the same simulation settings but the average χ^2 values for the approximate mixed linear model test statistics are given for the null chromosomes 2, 4, 6, 8, and 10 where no causal variants were simulated. (c) gives the true positive rate, calculated as the ratio of detected true positives and the total number of causal variants. (d) shows the false discovery rate calculated as the proportion of discoveries that lie on the null chromosomes 2, 4, 6, 8, and 10 where no causal variants were simulated.

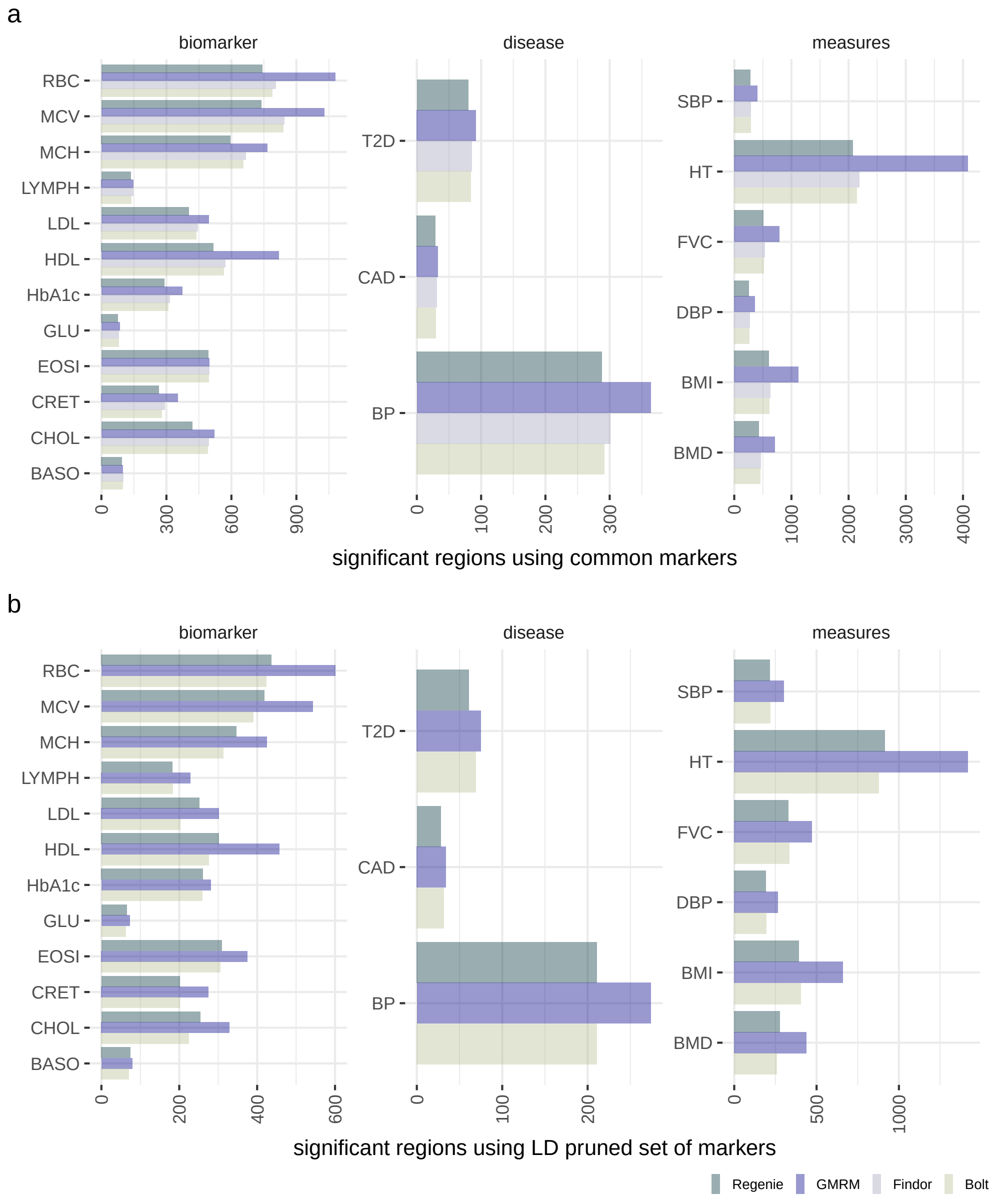

Fig. S5. GWAS discovery for GMRM in comparison to other approaches at different SNP marker sets. (a) Number of LD independent genomic regions identified at $5\text{x}10^{-8}$ by GMRM, as compared to BoltLMM (Bolt), Findor (Findor), and Regenie (Regenie) across 21 traits at 6,991,095 SNPs of minor allele frequency ≥ 0.01. For Findor, we use an LD Score UK Biobank annotation model based on BoltLMM summary statistics. (b) Number of LD independent genomic regions identified at $5\text{x}10^{-8}$ by GMRM, as compared to BoltLMM (Bolt) and Regenie (Regenie) across 21 traits at 2,174,071 LD clumped tagging SNP set. Full trait code descriptions are given in Supplementary Table 1.

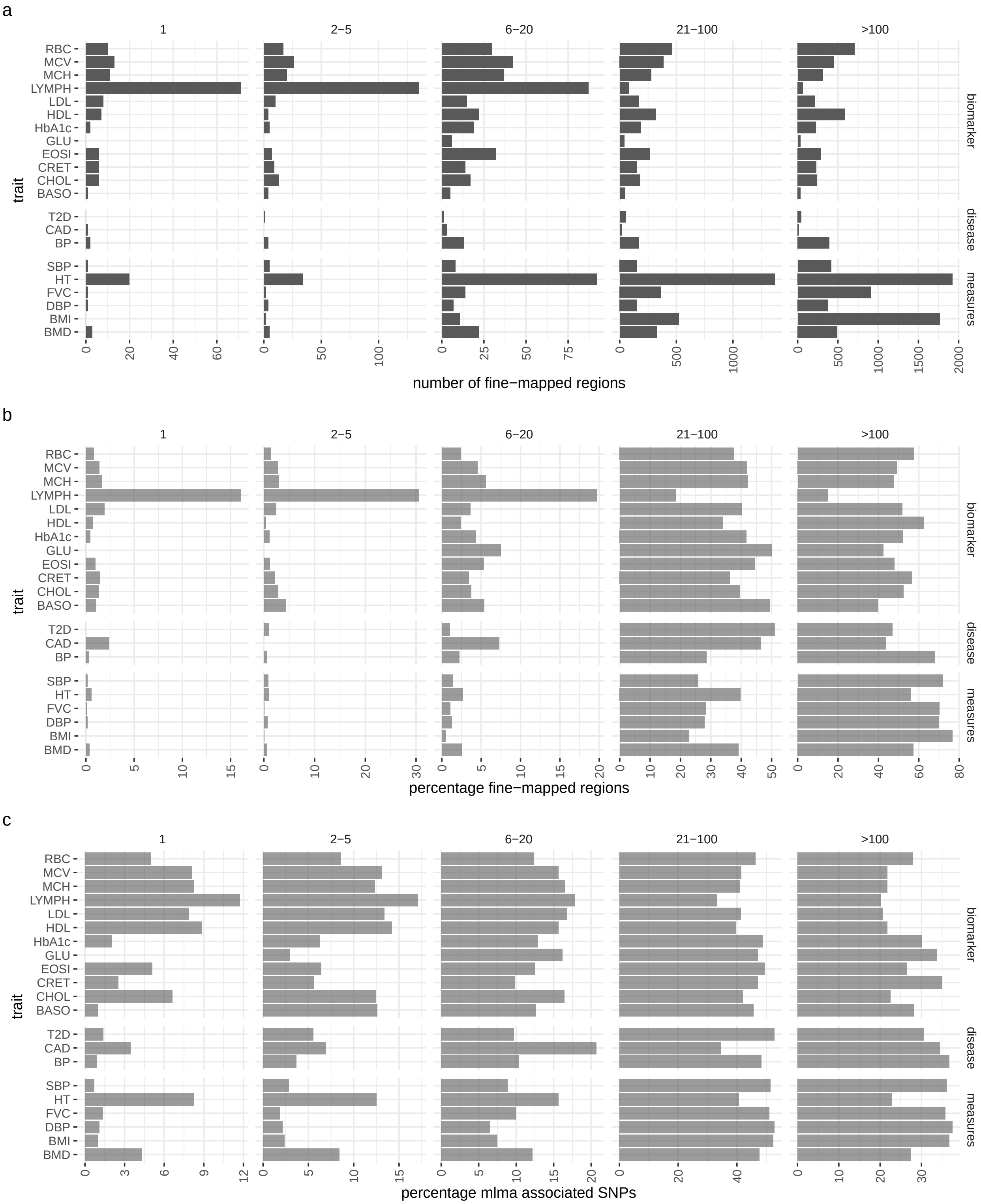

Fig. S6. GMRM Bayesian fine-mapping approach to localise the inclusion of markers in the model into SNP sets based on their LD. (a) The trait acronyms are given on the y-axis and the plot is faceted into five columns representing set of SNPs that are correlated at LD $R^2 \geq 0.1$, from some genomic regions having only a single SNP in LD with no others, to ≥ 100 markers in LD. Barplots give a count of the number of SNP sets for which the posterior probability of explaining greater than 0.001% of the phenotypic variance was greater than 95%. (b) For each trait, the percentage of SNP sets for which the posterior probability of explaining greater than 0.001% of the phenotypic variance was greater than 95% for different sizes of SNP set. (c) GMRM-MLMA significantly associated SNPs grouped into the same SNP sets based on LD. Across traits, a lower percentage of associations fine-map to either single SNPs, or to groups of five or less SNPs in LD as compared to the percentage of the discoveries from the GMRM-MLMA approach.

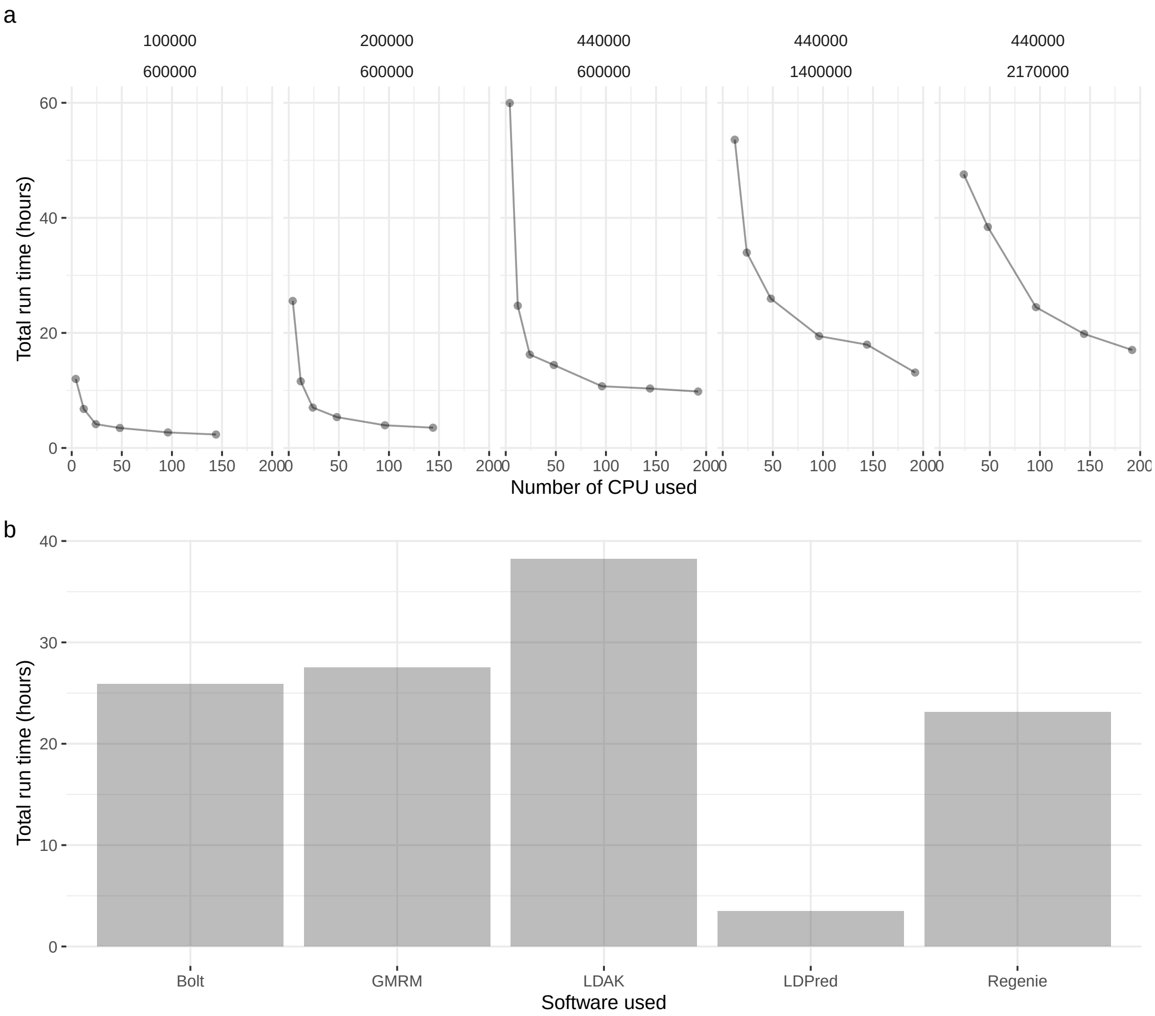

Fig. S7. Scaling of GMRM software with sample size and marker number. (a) The total run time for 3,000 iterations is given for an analysis of human height, with the sample size given by the first row of the plot header, and the marker number given in the second row. For all runs the total RAM use was approx. 1.02 times the size of the Plink binary data on disk, being 15.29 Gb, 28.38 Gb, 66.16 Gb, 147.48 Gb, and 227.32 Gb for the five plots respectively. In the last plot for 440,000 individuals and 2.17M markers, the RAM use would fill the available memory of a typical modern HPC node, and thus we only considered scenarios of 24 or 48 CPU usage. For each setting we used 2 CPU per MPI process. (b) Comparisons of average total run time for BoltLMM, GMRM, LDAK, LDPred-funct (LDPred), and Regenie across the 21 traits at 1,410,525 markers, using a single compute node and 48 CPU (threads). For GMRM, this represents the average time to produce both the posterior mean SNP effects and estimate the MLMA SNP estimates for traits analysed individually. For LDPred-funct, this is the average time taken to produce genomic predictors from existing summary statistics. For Regenie, we ran each trait independently, but traits can be combined and the presented time can be divided by the number of traits analysed.

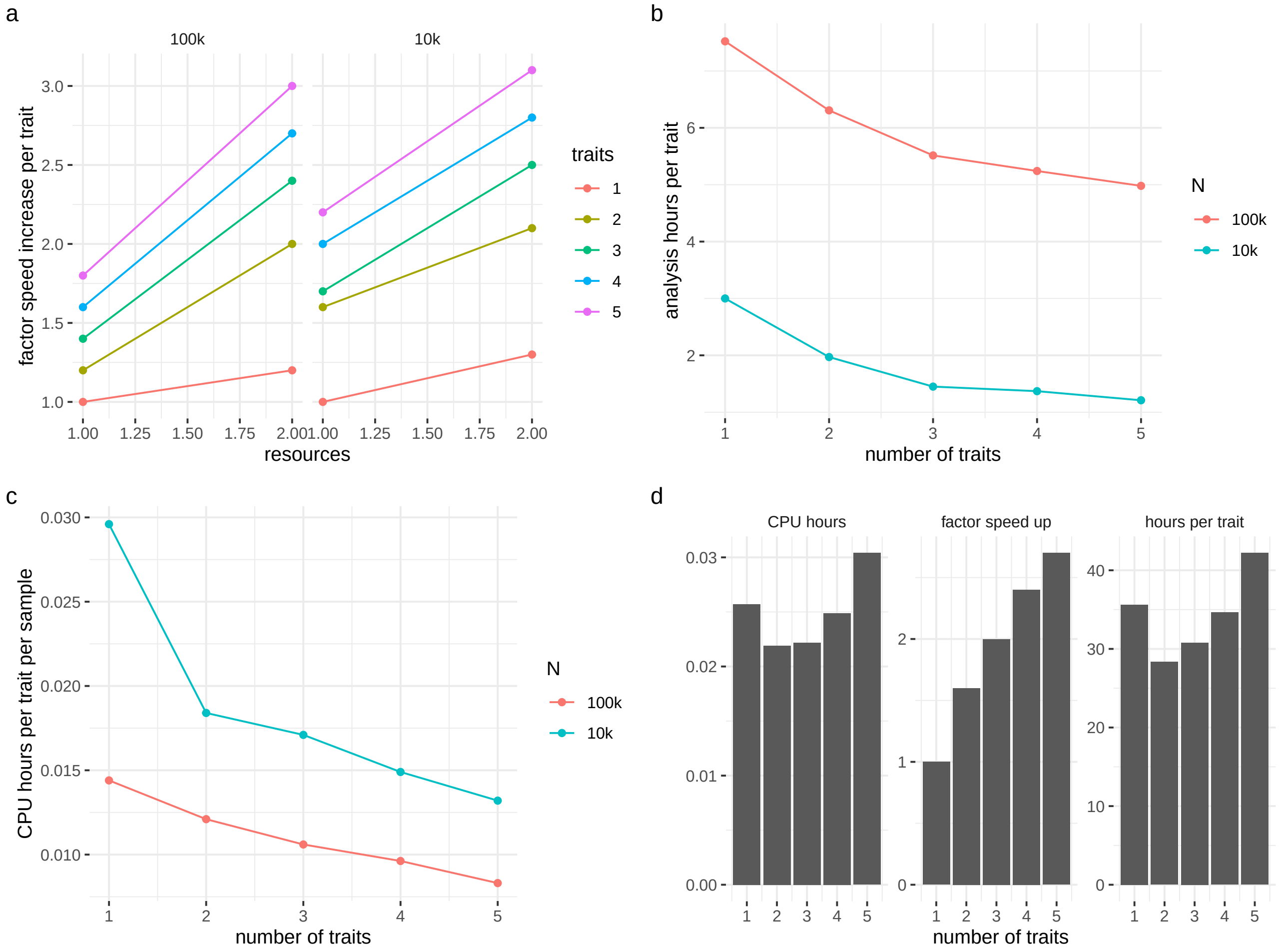

Fig. S8. Scaling of GMRM software for an analysis of 2.2M markers with multiple traits across many MPI processes. (a) For two resource settings of (1) 12 MPI processes each using 4 CPU cores on a single computer and 24 MPI processes each using 4 CPU cores across two computers, labelled as 1 and 2 on the x-axis, the factor speed-up per trait analysed is given on the y-axis. Speed-up factors are presented relative to the time taken to estimate 5,000 draws from the posterior distribution of a single trait in resource setting 1 (baseline time). We calculated the speed-up times for data from 10,000 individuals (10k) and 100,000 individuals (100k) separately. For example, for the analysis of 5 traits at the same time for 10k setting in resource setting 2 the value presented was calculated as: $t_{1,1}/(t_{5,2}/nt)$, where $t_{1,1}$ is the time at baseline for 1 trait in resource environment 1, $t_{5,1}$ is the time for analysing 5 traits simultaneously in resource environment 2, and nt is the number of traits analysed. Thus we obtain the proportional increase in time obtained relative to the baseline. (b) The number of hours required to obtain 5,000 posterior samples at two different sample sizes, decreases steeply but then begins to tail-off with a large number of simultaneously analysed traits when using 48 MPI processes, with each MPI process using 4 CPU. (c) In the configurations presented here when using 48 MPI processes, with each MPI process using 4 CPU, our algorithm displays sub-linear scaling of CPU core hours per trait per sample, with improved CPU use efficiency as sample size and number of traits analysed increases, essentially becoming more efficient as sample size and the number of traits grows. These examples were specifically designed to remain within memory and last level memory cache restrictions of the high-performance compute hardware used here in order to demonstrate the performance of our algorithm. Mostly, with the multi-trait version some MPI communication is shared among the traits, explaining the sub-linear scaling observed here. (d) In our hardware, at 400,000 individuals, 2.2M markers, and 72 MPI processes last-level cache size overflowed at greater than two traits analysed simultaneously, resulting in increased CPU hours per trait per sample (CPU hours), reduced speed-up, and an increase in the analysis time per trait. Thus, performance gains from our algorithm will become hardware limited. If the epsilon vectors stay in the cache we get continually improved performance as shown in (a) through (c). However, with more individuals, more traits, and more tasks per socket the hardware limits the sub-linear scaling to the point where degraded performance is observed as compared to single trait processing. The main considerations are the last-level cache size, the number of individuals and the number of traits to be analysed, and we advise users to set 1 MPI task per socket and estimate how many epsilon vectors can be held in the last level cache of the CPU.

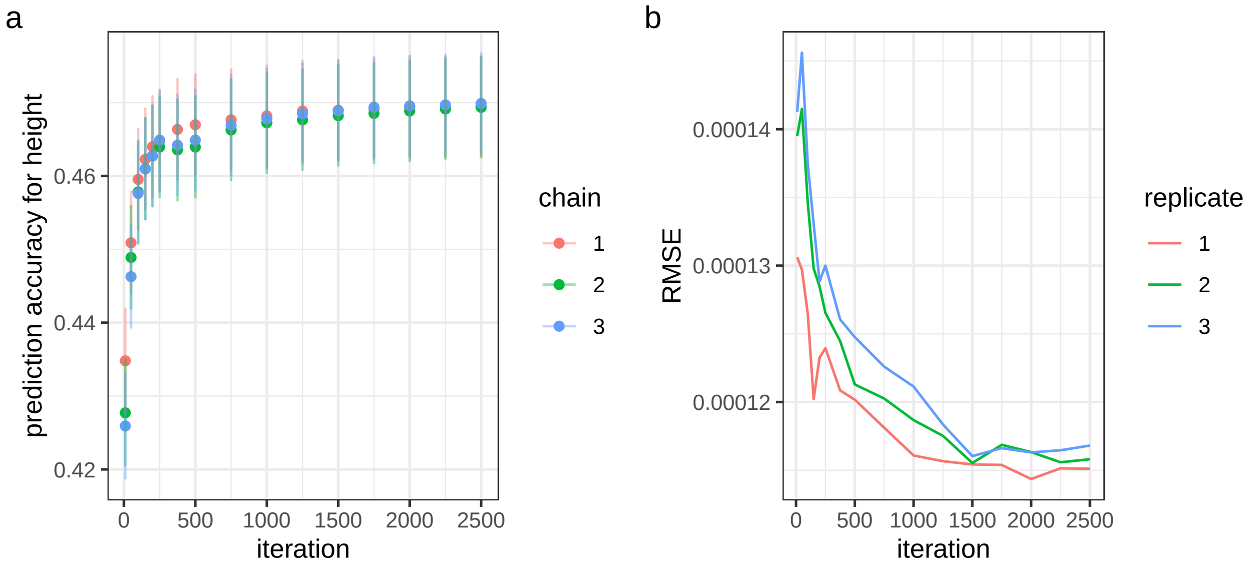

Fig. S9. Convergence of GMRM software. (a) Convergence of prediction accuracy in a hold out sample of 30,000 UK Biobank individuals for human height across three independent bulk synchronous Gibbs chains with synchronisation occurring after the update of 20 markers across 60 MPI process for 2.17 million SNP markers and 428,747 individuals. (b) The root mean square error of the marker effect size estimates of randomly assigned LD independent causal variants in simulation study of 100,000 unrelated individuals from the UK Biobank for three replicates.

Algorithm 1 GMRM MLMA algorithm

```
Matrix $\mathbf{X}$ of genotypes in plink binary format (.bed/.bim/.fam), matrix of posterior Bayesian marker estimates in binary format (.bet file), a file linking column number to rs SNP identifiers (.link), and file of phenotypic information (.phen file).
Assign markers present in SNP rs-identifier file to blocks across $k$ MPI tasks
Assign the posterior Bayesian marker estimates to the markers of each task $k$ based on the rs SNP identifiers
task $k$ Calculate means $\bar{x}_j$ and SD($x_j$) of markers
Initialise $\mathbf{g}_{k,l}$ and $\bar{\mathbf{g}}_k$
iteration $l$ in iter
marker $j$
if $\beta_j == 0$ thennext
    if $\beta_j ! == 0$ thencalculate $\hat{\mathbf{g}}_j$ as $\frac{\mathbf{x}_j - \bar{x_j}}{\mathrm{SD}(x_j)}\beta_j$;
add $\hat{\mathbf{g}}_j$ to $\mathbf{g}_{k,l}$  Add $\mathbf{g}_{k,l}$ to $\mathbf{g}_k$  Communicate $\mathbf{g}_k$ to task 0 where task 0 calculates $\mathbf{g} = \sum_{l=1}^{l}\sum_{k=1}^{k}\mathbf{g}_{k,l}$
Receive $\mathbf{g}$ from node 0
Subtract $\mathbf{g} - \mathbf{g}_k$ then divide by iter to give $\bar{\mathbf{g}}_k$
Calculate $\tilde{\mathbf{y}}_{\mathrm{block}_k}$ as $\tilde{\mathbf{y}}_{\mathrm{block}_k} = \mathbf{y} - \bar{\mathbf{g}}_k$
Calculate $\sigma^2_{\tilde{\mathbf{y}}_{\mathrm{block}_k}}$ as $\frac{1}{N}\tilde{\mathbf{y}}^T_{\mathrm{block}_k}\tilde{\mathbf{y}}_{\mathrm{block}_k}$ where $N$ is the number of individuals
marker $j$
calculate xtx as dot product $\mathbf{x}_j^T\mathbf{x}_j$
calculate xty as $\mathbf{x}_j^T\tilde{\mathbf{y}}_{\mathrm{block}_k}$
calculate $\beta_{j,k}$ as $[\mathrm{xtx}]^{-1}\mathrm{xty}$
calculate $t_j$ as $\frac{\mathrm{xty}}{[\sigma^2_{\tilde{\mathbf{y}}_{\mathrm{block}_k}}\mathrm{xtx}]^{0.5}}$
calculate p-value $j$ from $\chi^2_1$ using $t_j^2$
output $\beta_{j,k}$,$t_j$ and corresponding $\chi^2_1$ p-values.
```

Type	Phenotype	Code	Biobank codes	Adjusted covariate codes	MLMA sample size
Biomarker	Basophill count	BASO	30160-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	401,452
Biomarker	Cholesterol	CHOL	30690-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	395,025
Biomarker	Creatinine	CRET	30700-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	394,813
Biomarker	Eosinophill count	EOSI	30150-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	401452
Biomarker	Glucose	GLU	30740-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	360,010
Biomarker	Glycated haemoglobin	HbA1c	30750-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	394,912
Biomarker	HDL cholesterol	HDL	30760-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	360,286
Biomarker	LDL direct	LDL	30780-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	394,264
Biomarker	Lymphocyte count	LYMPH	30120-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	401,452
Biomarker	Mean corpuscular haemoglobin	MCH	30050-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	402,201
Biomarker	Mean corpuscular volume	MCV	30040-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	402,202
Biomarker	Red blood cell count	RBC	30010-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	402,204
Disease	Cardiovascular disease	CAD	3894-0.0, 3627-0.0, 42000-0.0, 6150-0.0, 40001-0.0, 41202-0.0, 41204-0.0, 41270-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	401,000 (78,283)
Disease	High blood pressure	BP	6150-0.0, 2966-0.0, 6153-0.0, 6177-0.0, 40001-0.0, 41201-0.0, 41202-0.0, 41204-0.0, 41270-0.0, 41262-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	401,106 (126,352)
Disease	Type-2 diabetes	T2D	2443-0.0, 2976-0.0, 6153-0.0, 6177-0.0, 40001-0.0, 41201-0.0, 41202-0.0, 41204-0.0, 41270-0.0, 41262-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	398,322 (36,865)
Measures	Heel bone mineral density T-score	BMD	3148-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	231,693
Measures	Body mass index	BMI	21001-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	413,595
Measures	Diastolic blood pressure	DBP	4079-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	377,358
Measures	Forced vital capacity	FVC	3062-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	376,724
Measures	Standing height	HT	50-0.0	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20	414,055
Measures	Systolic blood pressure	SBP	4080-0.1	Sex, Age, East-West coordinates, UK Biobank Centre, Genotype Batch, PCs 1-20, Medication (6177-0.0)	377,347

Table S1. UK Biobank phenotypes used within the study. Columns of the table give the type of trait, the name, the trait code used for the figures, the UK Biobank codes used to construct the trait values, the covariates adjusted for within the analysis, and the sample size of the MLMA GWAS association analysis (case numbers given in brackets).