Joint modelling of whole genome sequence data for human height via approximate message passing

Al Depope[1,*], Jakub Bajzik[1], Marco Mondelli [1,†,*], Matthew R. Robinson[1,†,*]

[1] Institute of Science and Technology Austria, Klosterneuburg, Austria.

*corresponding authors
† indicates joint supervision

Abstract

Human height is a model for the genetic analysis of complex traits, and recent studies suggest the presence of thousands of common genetic variant associations and hundreds of low-frequency/rare variants. However, it has not yet been possible to fine-map the genetic basis of height, since all variant effects have not been modelled jointly leaving correlations unaccounted for. To address this issue, we develop a new algorithmic paradigm based on approximate message passing, *gVAMP*, to directly fine-map whole-genome sequence (WGS) variants and gene burden scores, conditional on all other measured DNA variation genome-wide. We find that the genetic architecture of height inferred from WGS data differs from that inferred from imputed single nucleotide polymorphism (SNP) variants: common variant associations from imputed SNP data are allocated to WGS variants of lower frequency, and there is a stronger relationship of effect size and variant frequency. Thus, even fine-mapped imputed variants are systematically mis-assigned and without the joint analysis of WGS data it remains premature, if not unfounded, to make statements regarding the number of independent associations and their properties. We validate gVAMP on various datasets across UK Biobank traits where it outperforms widely used methods for polygenic risk score prediction and association testing, offering a scalable foundation towards analyzing hundreds of millions of variables measured on millions of people.

Keywords: whole genome regression; joint association testing; fine-mapping; polygenic risk scores; approximate message passing

Introduction

Efficient utilization of large-scale biobank data is crucial for inferring the genetic basis of disease and predicting health outcomes from DNA. The common statistical approach of single-marker or single-gene burden score regression [1–4], gives marginal associations that do not account for linkage disequilibrium (LD), and in whole genome sequence (WGS) data, even weak associations that are physically distant from causal variants will be discovered as significant at scale. While fine-mapping aims to identify causal variants, current methods focus only on genome-wide significant loci within one region at a time [5], in isolation from the rest of the genome, resulting in miscalibration and a compromise of power. Thus, we currently lack accurate statistical models to jointly estimate the effect of each locus, conditional on all other genetic variants. Applying whole genome regression (WGR), where the effect of each variant is estimated conditional on all others, has the potential to resolve these issues and reveal the underlying genetic architecture of complex traits.

Here, we focus on the highly heritable polygenic phenotype of human height, and develop a new framework, gVAMP, which fits tens of millions of WGS variants jointly at scale. Applying gVAMP to WGS data on hundreds of thousands of UK Biobank participants, we find a stronger relationship of effect size and variant frequency, due to common variant associations in imputed SNP data being allocated to WGS variants of lower frequency. These insights could not be obtained from existing statistical approaches, and we additionally validate gVAMP on a number of datasets by benchmarking against the state of the art: gVAMP outperforms widely used summary statistic methods such as LDpred2 [6] and SBayesR [7] for polygenic risk score prediction, and an individual-level REGENIE [1] method for association testing. Additionally, we show that its performance matches that of MCMC sampling schemes [8] but with a dramatic speed-up in time (analysing 8.4M SNPs jointly in under day as opposed to weeks). This lays the foundations for a wider range of analyses in large WGS datasets that are entirely infeasible for other methods.

Results

Overview of the approach

Our focus is on the simple idea of joint association testing controlling for local and long-range LD: we estimate the significance of each variant, conditional on all other observed DNA locations genome-wide. To do this, we consider a general form of whole-genome Bayesian linear regression, common to genome-wide association studies (GWAS) [7, 8], estimating the effects vector $\boldsymbol{\beta} \in \mathbb{R}^P$ from a vector of phenotype measurements $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ given by

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + \epsilon_i, \quad \text{for } i \in \{1, \ldots, N\}. \tag{1}$$

Here, $\boldsymbol{x}_i$ is the row of the normalized genotype matrix $\boldsymbol{X}$ corresponding to the i-th individual, $\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle = \boldsymbol{x}_i^T \boldsymbol{\beta}$ denotes the inner product, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is an unknown noise vector with multivariate normal distribution $\mathcal{N}(0, \gamma_\epsilon^{-1} \cdot \boldsymbol{I})$ and unknown noise precision γ_ϵ^{-1}. To allow for a range of genetic effects, we select the prior on $\boldsymbol{\beta}$ to be of an adaptive spike-and-slab form:

$$\beta_i \sim (1-\lambda)\cdot\delta_0(\cdot) + \lambda \cdot \sum_{i=1}^{L} \pi_i \cdot \mathcal{N}(\cdot, 0, \sigma_i^2), \quad \text{for } i \in \{1, \ldots, P\}. \tag{2}$$

Here, $\lambda \in [0,1]$ is the DNA variant inclusion rate, L is the unknown number of Gaussian mixtures, $(\pi_i)_{i=1}^L$ denote the mixture probabilities and $(\sigma_i^2)_{i=1}^L$ the variances for the slab component.

Current association testing [1–4], fine-mapping [5] and polygenic risk score methods [8, 9] are all based on forms of Equations (1) and (2), with parameters estimated by restricted maximum likelihood (REML), Markov Chain Monte Carlo (MCMC), expectation maximisation (EM), or variational inference (VI). REML and MCMC are computationally intensive and slow; while EM and VI are faster, they trade speed for accuracy with few theoretical guarantees. Furthermore, current software implementations of these algorithms limit either the number of markers or individuals. Mixed linear model association (MLMA) approaches are restricted to using less than one million SNPs to control for the polygenic background [1, 2], resulting in a loss of power [8] and the potential for inadequate control for fine-scale confounding factors [10]. Polygenic risk score algorithms are limited to a few million SNPs, and lose power by modelling only blocks of genetic markers [6, 7]. Likewise, fine-mapping methods are generally limited to focal segments of the DNA [5], and they are unable to fit all genome-wide DNA variants together. Thus, no existing approach can apply the statistical model of Equations (1) and (2) to jointly estimate the effects vector $\boldsymbol{\beta}$ and the genome-wide significance of each element in WGS data.

We overcome this issue by developing a new approach for GWAS inference, dubbed *genomic Vector Approximate Message Passing* (gVAMP). Approximate Message Passing (AMP) [11–13] refers to a family of iterative algorithms with several attractive properties: *(i)* AMP allows the usage of a wide range of Bayesian priors; *(ii)* the AMP performance for high-dimensional data can be precisely characterized by a simple recursion called state evolution [14]; *(iii)* using state evolution, joint association test statistics can be obtained [15]; and *(iv)* AMP achieves Bayes-optimal performance in several settings [15–17]. However, we find that existing AMP algorithms proposed for various applications [18–21] cannot be transferred to biobank analyses as: *(i)* they are entirely infeasible at scale, requiring expensive singular value decompositions; and *(ii)* they give diverging estimates of the signal in either simulated genomic data or the UK Biobank data. To address the problem, we combine a number of principled approaches to produce an Expectation Propagation method tailored to whole genome regression as described in the Methods (Algorithm 1). gVAMP approximates the posterior $\mathbb{E}[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}]$, providing joint effect size estimates and statistical testing via state evolution (see "gVAMP SE association testing" in the Methods). Additionally, we learn all unknown parameters in an adaptive Bayes expectation-maximisation (EM) framework [22, 23], which avoids expensive cross-validation and yields biologically informative inference of the phenotypic variance attributable to the genomic data (SNP heritability, h^2_{SNP}) allowing for the first full characterisation of the genetic architecture of human complex traits in WGS data.

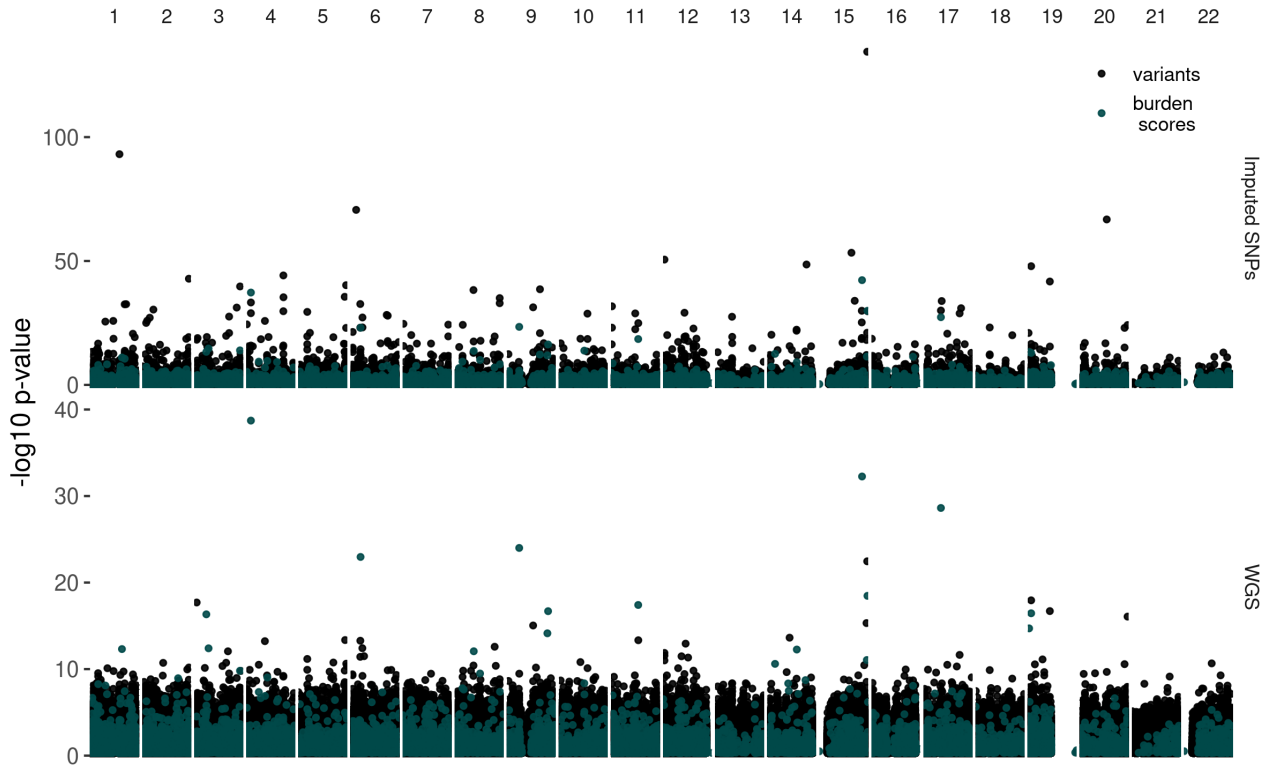

Figure 1. Joint association plot of 8.4 million imputed SNPs and 17 million WGS for human height in 415,000 UK Biobank participants. AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers. In the panel "Imputed SNPs", we combine 8,430,446 autosomal imputed SNP markers with 17,852 whole exome sequencing gene burden scores, estimating the effects jointly within the gVAMP SE testing framework. In the panel "WGS" we combine 16,854,878 whole genome sequence variants with 17,852 whole exome sequencing gene burden scores, again estimating the effects jointly within the gVAMP SE testing framework.

The genetic architecture of human height

When analysing 415,000 UK Biobank individuals, we find that the genetic architecture of human height inferred from 16,854,878 WGS variants differs to that inferred from 8,430,446 imputed SNP markers (Figure 1). gVAMP estimates the proportion of phenotypic variance in human height attributable to WGS data as 0.652, as compared to 0.63 for the imputed SNP data, comparable to a previously published REML estimate in a different WGS dataset [25] and family-based estimates [26]. This confirms that additional phenotypic variation is attributable to variants in WGS data that are missing in imputed SNP data. Surprisingly, despite this increase in attributable height variation, when gVAMP maps effects to single-locus positions across the DNA in WGS data, we find 526 genome-wide significant effects as compared to 930 in the imputed data (Figure 1). This decrease in genome-wide significant loci occurs because the WGS analysis attributes height variation to DNA variants of lower minor allele frequency (MAF), as compared to the imputed SNP data analysis, giving a reduction in association testing power (Figure 2).

For WGS variants significant at different thresholds, we determine whether the same variant, or a variant within a given number of base pairs (distance, x-axis), is identified in the imputed SNP data at the same significance threshold. We find little overlap in the locations of the SNPs that we

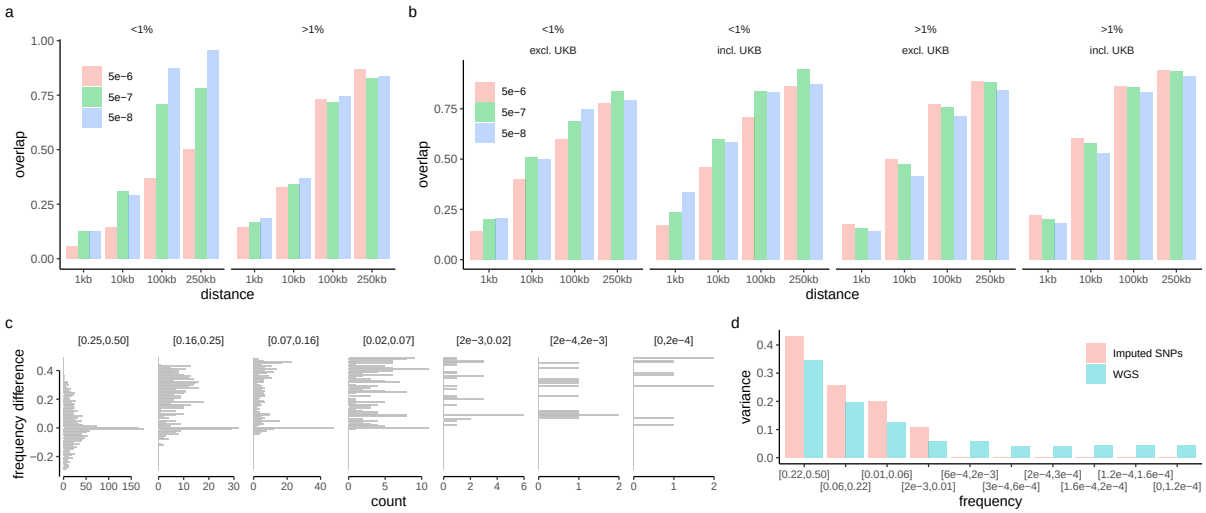

Figure 2. The genetic architecture of human height inferred from 16,854,878 whole genome sequence variants differs to that inferred from 8,430,446 imputed SNP markers in the UK Biobank. *(a)* For each whole genome sequence (WGS) variant discovered at different genome-wide significance thresholds, we determine whether we also identify a variant within a given number of base pairs (distance, x-axis) in the imputed SNP data at the same threshold. The overlap is calculated as the proportion of WGS variants of either $> 1\%$ or $< 1\%$ minor allele frequency (MAF) that are discovered at a given threshold within a certain base pair distance. *(b)* For each whole genome sequence (WGS) variant discovered at different genome-wide significance thresholds, we determine whether a variant was identified within the latest height GWAS study [24] at the same threshold for a given number of base pairs (distance, x-axis), with the overlap calculated as in *(a)*. We select the GWAS marginal summary statistics for European individuals including (incl. UKB), or excluding (excl. UKB) the UK Biobank (for analysis details see [24]). *(c)* For each WGS variant at genome-wide significance level $p \leq 5 \cdot 10^{-6}$, we determine the imputed SNP at $p \leq 5 \cdot 10^{-6}$ with the closest MAF and show a histogram of these frequency differences for WGS variants of different frequencies. *(d)* For all WGS variants and imputed SNPs, we calculate the proportional contribution to phenotypic variance across different MAF groups.

map to single-locus resolution across the two datasets within 1kb (a small proportion maps to the same location), but substantial overlap within 100kb either side of the WGS findings (Figure 2a). Thus, similar DNA regions are identified, but the effects are assigned to different variant locations. When we determine whether the same variant, or a variant within a given number of base pairs (distance, x-axis), was identified in the most recent GWAS study of human height [24], we again find little overlap in the locations of the SNPs that we map to single-locus resolution across the two datasets within 1kb, but substantial overlap within 100kb either side of the WGS findings (Figure 2b). This demonstrates that our results are predominantly replicated in large-scale GWAS studies, but again that in WGS data effects are localised to different DNA variants.

When we then examine the properties of the WGS variants we identify, we find that WGS variants of MAF $\leq 16\%$ are generally always mis-mapped in the imputed data to variants of higher frequency (Figure 2c). For each WGS variant discovered at genome-wide significance level $p \leq 5 \cdot 10^{-6}$, we determine whether there is an imputed SNP at $p \leq 5 \cdot 10^{-6}$ within 250kb, and show a histogram of the imputed variant with closest MAF, separating the discovered WGS variants by their frequency: we find that, while some variants map to the same location across the two datasets (frequency difference of 0), the majority do not and are assigned in the imputed data to variants of higher frequency (Figure 2c). We also find that each WGS variant not discovered in the imputed data can have multiple neighbouring SNPs of various MAF distribution with the same significance level

(Figure S1). As a consequence, the phenotypic variance attributable to different MAF groups differs in WGS as compared to imputed SNP data, with less variance attributable to common SNPs in WGS data (Figure 2d). This shows that for human height, many discoveries in imputed SNP data are attributable to variants of lower frequency in WGS data. Thus, fine-mapped imputed variants can be systematically mis-assigned, without a full joint analysis of WGS data.

Despite the expectation that fine-mapped variants would show elevated marginal test statistics in standard GWAS association testing, we find that many genome-wide significant height-associated rare variants discovered in both the WGS and imputed data were not found in previous UK Biobank analyses in Open Targets including: rs116467226, an intronic variant by TPRG1; rs766919361, an intronic variant by FGF18; rs141168133, an intergenic variant 19kb from ID4; rs150556786, an upstream gene variant for GRM4; rs574843917, a non-coding transcript exon variant in GPR21; rs532230290, an intergenic variant 42kb from SCYL1; rs543038394, intronic in OVOL1; rs1247942912, a non-coding transcript exon variant in AC024257.3; rs577630729, a regulatory region variant for ISG20; and rs140846043, a non-coding transcript exon variant of MIRLET7BHG. 10 out of our top 38 height-associated WGS variants of $\leq 1\%$ MAF were not previously discovered, and only become height-associated when conditioning on the entire polygenic background captured by WGS data within our analysis.

We can directly determine the relationship between effect size and minor allele frequency, which again differs between WGS and imputed SNP data (Figure 3a). For variants of significance $p \leq 5 \cdot 10^{-4}$, the power relationship of effect size and locus variance, denoted as α in the literature [27], is $\alpha = -0.318$, 95% CI $= 0.022$, p-value $\leq 2 \cdot 10^{-16}$ for imputed SNPs, which is consistent with previous estimates. However, this is much lower for WGS variants: $\alpha = -0.566$, 95% CI $= 0.004$, p-value $\leq 2 \cdot 10^{-16}$. Our model allows for different types of DNA observations to be combined and when we include 17,852 WES gene burden scores into the analysis, we have *(i)* $\alpha = -0.826$, 95% CI $= 0.083$, p-value $\leq 2 \cdot 10^{-16}$ for WES burden scores fit alongside WGS variants; and *(ii)* $\alpha = -0.891$, 95% CI $= 0.042$, p-value $\leq 2 \cdot 10^{-16}$ for WES burden scores fit alongside imputed SNPs. Thus, it appears likely that the relationship between effect size and MAF for human height is stronger than previously inferred for imputed SNP data.

We highlight the benefits of joint estimation to explore genetic architecture, where controlling for LD allows effects to be summed over different categories, facilitating gene/annotation analyses. We find a general concordance of the estimated effect sizes of the 17,852 WES gene burden scores when fit alongside either imputed or WGS data, for most but not all genes (Figure 3b). We sum up the joint effects to determine the variation in height attributable to each gene, and again find general concordance across markers annotated to each of the 17,852 genes, conditional on all other markers, across the imputed SNP and WGS analysis (Figure 3c). We see little relationship between the variance attributable to the burden score of a given gene and the SNPs annotated to the gene (Figure S2), with some notable exceptions: ACAN, ADAMTS17, ADAMTS10, and LCORL. The top genes, where gVAMP attributes $\geq 0.04\%$ of height variation in addition to those listed above are EFEMP1, ZBTB38, and ZFAT. All the 38 genome-wide significant gene burden scores are for genes that have previous GWAS height associations linked to them in Open Targets, but our analysis suggests that the effect is attributable to a rare protein coding variant rather than the common variants suggested by current genome-wide association studies.

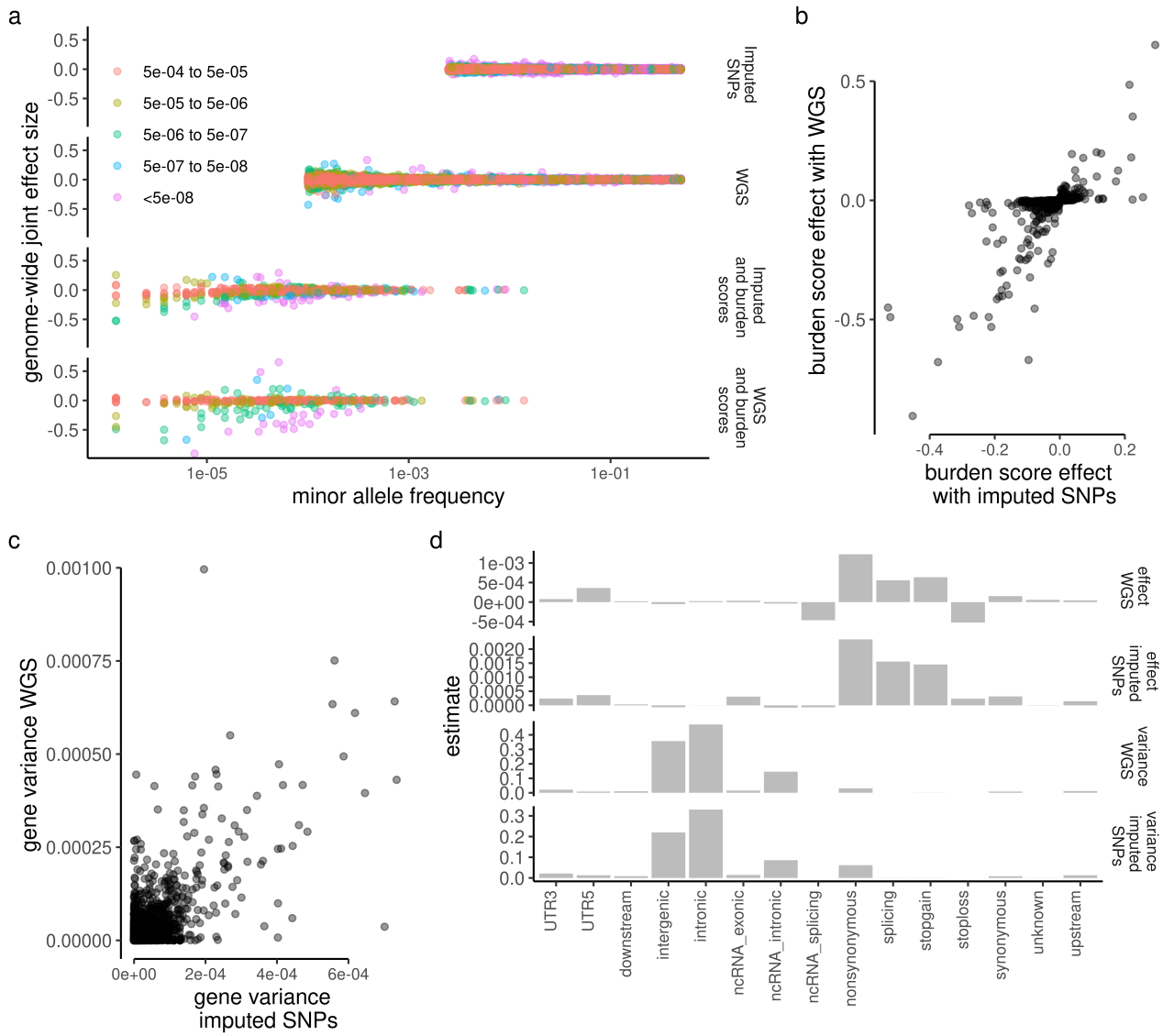

Figure 3. The relationship between effect size and minor allele frequency for human height inferred from whole genome sequence data differs to that inferred from imputed SNP variants in the UK Biobank. *(a)* For different genome-wide significance thresholds, we plot the relationship between joint effect size and minor allele frequency (MAF) for imputed SNPs, whole genome sequence (WGS) variants and whole exome sequence (WES) burden scores fit alongside either imputed SNPs or WGS data. *(b)* Across 17,852 WES gene burden scores, we find general concordance of the estimated effect sizes when fit alongside either imputed or WGS data, for most but not all genes, with squared correlation 0.532. *(c)* Likewise, we also find general concordance of the phenotypic variance attributable to markers annotated to each of the 17,852 genes, when fit conditional on either imputed or WGS variants, with squared correlation 0.494. *(d)* Finally, we show similar patterns of enrichment when annotating markers to functional annotations in either the proportion of variance attributable to each group (labelled "variance"), or in the average effect size relative to the genome-wide average effect size (labelled "effect") from joint estimation of either imputed or WGS data.

Additionally, across annotations, we find that the phenotypic variance attributable to different DNA regions is higher for intergenic and intronic variants (Figure 3d). However, when adjusting for the number of SNPs contained by each category by using the average effect size of the group relative to the average effect size genome-wide, we find that exonic variants that are nonsynonymous, splicing and stop-gain have the largest average effects of the WGS variants included within our

model (Figure 3d). Taken together, joint association testing of all WGS variants resolves many previously discovered height-associated DNA regions to rare DNA variants where exonic variants have large effect sizes, insights that cannot be provided by other GWAS approaches at a scale of 17M DNA variants.

Validation and benchmarking of gVAMP

Table 1. Polygenic risk score prediction accuracy R^2 for 13 different traits from statistical models trained in the UK Biobank data and tested in a UK Biobank hold-out set. Training data sample size and trait codes are given in Table S1 for each trait. The sample size of the hold-out test set is $15,000$ for all phenotypes. LDpred2 and SBayesR give estimates obtained from the LDpred2 and SBayesR software respectively, using summary statistic data of 8,430,446 SNPs obtained from the REGENIE software. GMRM denotes estimates obtained from a Bayesian mixture model at 2,174,071 SNP markers ("GMRM 2M"). gVAMP denotes estimates obtained from an adaptive EM Bayesian mixture model within a vector approximate message passing (VAMP) framework, using either 887,060 ("gVAMP 880k"), 2,174,071 ("gVAMP 2M"), or 8,430,446 SNP markers ("gVAMP 8M").

Phenotype	LDpred2	SBayesR	GMRM 2.17M	gVAMP 880k	gVAMP 2.17M	gVAMP 8M
CHOL	0.147	0.149	**0.153**	0.140	0.152	**0.153**
EOSI	0.107	0.112	0.122	0.114	0.120	**0.124**
HbA1c	0.087	0.090	0.092	0.085	0.092	**0.095**
HDL	0.199	0.208	0.213	0.192	0.209	**0.219**
MCH	0.178	0.215	0.221	0.203	0.218	**0.223**
MCV	0.196	0.234	**0.244**	0.222	0.240	**0.244**
RBC	0.186	0.191	**0.199**	0.182	0.195	*0.198*
BMI	0.100	0.118	0.133	0.107	0.132	**0.141**
DBP	0.065	0.067	**0.071**	0.058	0.065	**0.071**
FVC	0.098	0.103	0.111	0.097	0.109	**0.112**
BMD	0.188	0.194	0.201	0.183	0.198	**0.204**
HT	0.231	0.362	0.450	0.419	0.449	**0.457**
SBP	0.068	0.071	**0.073**	0.061	0.072	**0.073**

As gVAMP is the only algorithm that scales to tens of millions of WGS variants, we can only validate and benchmark gVAMP against state-of-the-art approaches in a number of alternative datasets. We find that *(i)* gVAMP outperforms summary statistic approaches [6, 7] for polygenic risk score prediction, *(ii)* it outperforms REGENIE [1] for mixed-linear model association testing, and *(iii)* it has similar performance to MCMC approaches [8], but in a fraction of the compute time, which allows analyses at far larger scale that then result in improved performance (Figure 4).

Specifically, we compare the prediction accuracy of gVAMP to the widely used summary statistic methods LDpred2 [6] and SBayesR [7], and to the individual-level method GMRM [8] for imputed SNP data in the UK Biobank across 13 traits (training data sample size and trait codes given in Table S1). gVAMP outperforms all methods for most phenotypes and, in comparison to published estimates, we obtain the highest out-of-sample prediction accuracy yet reported to date for most traits. Specifically, for human height, we obtain an accuracy of 45.7%, which is a 97.8% relative increase over LDpred2 (Table 1 and Figure 4) and higher than the accuracy obtained from the latest height GWAS study of 3.5M people of 44.7% [24], despite our sample size of only 414,055. However, we caution that modelling WGS data does not improve the out-of-sample prediction obtained as

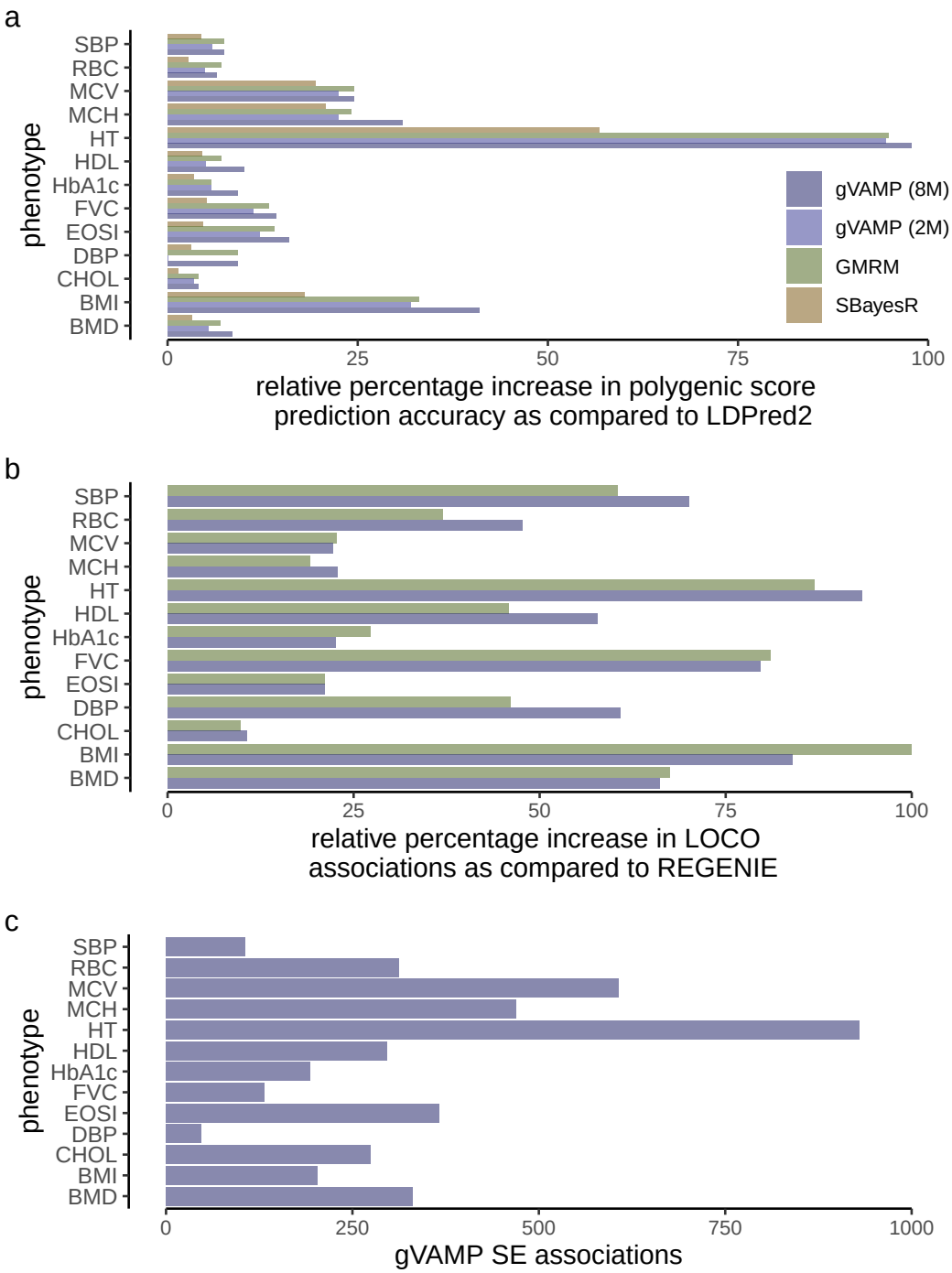

Figure 4. Validating gVAMP through polygenic risk score accuracy and association testing benchmarks in the UK Biobank within imputed SNP data. *(a)* Relative prediction accuracy of gVAMP in a hold-out set of the UK Biobank across 13 traits as compared to other approaches. *(b)* Relative number of leave-one-chromosome-out (LOCO) testing of gVAMP across 13 UK Biobank traits as compared to other approaches at 8,430,446 markers. *(c)* Number of genome-wide fine-mapped associations obtained via gVAMP SE association testing for 13 UK Biobank traits at a p-value threshold of less than $5 \cdot 10^{-8}$ for all 8,430,446 SNP markers.

compared to the 8.4M imputed SNP results for human height, and this is again likely because of the reduction in the MAF of the markers included within the WGS model.

Generally, gVAMP performs similarly to GMRM, an MCMC sampling algorithm, improving over

Table 2. Genome-wide significant associations for 13 UK Biobank traits from GMRM, gVAMP and REGENIE at 8,430,446 genetic variants. REGENIE denotes results obtained from leave-one-chromosome-out (LOCO) testing using the REGENIE software, with 882,727 SNP markers used for step 1 and 8,430,446 markers used for the LOCO testing of step 2. GMRM refers to LOCO testing at 8,430,446 SNPs, using a Bayesian MCMC mixture model in step 1, with either 882,727 ("GMRM 880k") or 2,174,071 SNP markers ("GMRM 2M"). gVAMP refers to LOCO testing at 8,430,446 SNPs, using the framework presented here, where in step 1 either 882,727 ("gVAMP 880k"), 2,174,071 ("gVAMP 2M"), or 8,430,446 SNP markers ("gVAMP 8M") were used. We also present leave-one-out ("gVAMP 8M LOO", see Methods) and state-evolution (SE) p-value testing for 8,430,446 SNP markers ("gVAMP 8M SE", see Methods). For LOCO testing, the values give the number of genome-wide significant linkage disequilibrium independent associations selected based upon a p-value threshold of less than $5 \cdot 10^{-8}$ and R^2 between SNPs in a 5 Mb genomic segment of less than 1%. For LOO and SE testing, values give the number of genome-wide significant associations selected based upon a p-value threshold of less than $5 \cdot 10^{-8}$.

Phenotype	REGENIE	GMRM 2M	gVAMP 880k	gVAMP 2M	gVAMP 8M	gVAMP 8M LOO	gVAMP 8M SE
CHOL	571	627	567	603	**632**	379	274
EOSI	572	**693**	607	630	**693**	568	367
HbA1c	337	**429**	365	385	*413*	229	193
HDL	692	1009	759	812	**1092**	488	297
MCH	746	889	773	810	**916**	994	470
MCV	970	**1190**	997	1062	*1185*	875	607
RBC	897	1229	982	1079	**1325**	764	312
BMI	688	**1376**	852	1175	*1266*	220	203
DBP	291	425	343	419	**468**	108	47
FVC	549	**994**	664	721	*986*	323	132
BMD	522	**874**	615	668	*867*	561	331
HT	2712	5070	3615	4452	**5242**	2553	930
SBP	311	499	351	388	**529**	160	106

it when analysing the full set of 8,430,446 imputed SNPs (Table 1 and Figure 4). gVAMP estimates the h^2_{SNP} of each of the 13 traits at an average of 3.4% less than GMRM when using 2,174,071 SNP markers, but at an average of 3.9% greater than GMRM when using 8,430,446 SNP markers. This implies that more of the phenotypic variance is captured by the SNPs when the full imputed SNP data are used (Figure S3). We note that GMRM already takes several days to analyze 2,174,071 SNPs and would take many weeks to analyze 8,430,446 SNPs, making it entirely infeasible to run at that scale. In contrast, gVAMP yields estimates on 8,430,446 SNPs in under a day (Supplementary Note 1, Figure S5c).

Second, we highlight that gVAMP can also be used for standard leave-one-chromosome-out (LOCO) statistical testing. We compare gVAMP to REGENIE and to GMRM for association testing of the 13 traits within a MLMA framework. gVAMP performs similarly in LOCO testing conducted using a predictor from GMRM, with the use of the full 8,430,446 imputed SNP markers generally improving performance (Table 2 and Figure 4b). REGENIE yields far fewer associations than either GMRM or gVAMP for all traits (Table 2 and Figure 4b), consistent with simulation study results presented in Supplementary Note 1. Using the gVAMP SE association testing framework, we find hundreds of marker associations for each trait that can be localised to the single locus level, conditional on all other SNPs genome-wide (Table 2, Figure 4c), with the obvious caveat that these results are for imputed SNP data and re-analysis of WGS data may yield different results as we show above for height. For all 13 traits, we find that the SE association estimates we obtain converge in number and location after iteration 20 (Figure S4).

In addition, we conduct a series of simulation studies showing that gVAMP is the only approach to generate genetic predictors and association test statistics in a single step, without additional computations, with accuracy similar to individual-level MCMC methods achieved in a fraction of the compute time (Supplementary Note 1). As compared to REGENIE, gVAMP completes in 2/3 of the time given the same data and compute resources, and it is dramatically faster (12.5× speed-up) than GMRM (Supplementary Note 1, Figure S5c).

Discussion

Our results reveal that a different genetic architecture for human height is inferred in WGS data, as compared to imputed SNP data, which shows that fine-mapping results can be miscalibrated by missing rare variants. Although large sample sizes of WGS data will be needed to pinpoint the variants responsible for the heritability of traits, our results show that the prioritization of relevant genes and gene sets is feasible at smaller sample sizes in imputed data. We highlight that gVAMP is not restricted to the analysis of WGS data, and it also provides a general approach to obtain genetic predictors and MLMA association test statistics in a single step, with accuracy similar to individual-level MCMC methods, but in a fraction of the compute time. We demonstrate this both in an extensive simulation study and in the analysis of 13 UK Biobank traits. Importantly, we provide a different association testing approach where the effects of each locus or burden score can be estimated conditional on all other DNA variation genome-wide. This allows associations to be localised to the single-locus, or single-gene level, refining associations by testing each of them against a full genetic background of millions of DNA variants.

There are a number of remaining limitations. Our results suggest that the detection and accurate estimation of the effects of height-associated variants is expected to be difficult even with millions of WGS samples. There are a very large number of rare variants within the human population that are missing from our WGS analysis of 16,854,878 variants, and we believe it is quite likely that the true underlying genetic architecture of human height is even rarer than we present here. Two solutions could be to *(i)* apply gVAMP region-by-region, or *(ii)* make slight modifications in the implementation, so that gVAMP streams data, at the cost of increased run time. Additionally, while our approach can be applied within any sub-grouping of data (by age, genetic sex, ethnicity, etc.), this is not within the scope of the present work. Combining inference across different groups is of great importance [28], and previous work suggests that better modelling within a single large biobank can facilitate improved association testing in other global biobanks [29]. Here, while our approach can be used in the same way, maximising association and prediction across the human population requires a model that is capable of accounting for differences in the design matrix (minor allele frequency and linkage disequilibrium patterns) across different datasets. Our ongoing work now aims at expanding the gVAMP framework to make inference across a diverse range of human groups, to model different outcome distributions (binary outcomes, time-to-event, count data, etc.), to allow for different effect size relationships across allele frequency and LD groups, to model multiple outcomes jointly, and to do all of this using summary statistic as well as individual-level data across different biobanks. This is key to obtaining the sample sizes that are likely required to fully explore the genetic basis of complex traits.

In summary, gVAMP is a different way to create genetic predictors and to conduct association testing. With increasing sample sizes reducing standard errors, a vast number of genomic regions are being identified as significantly associated with trait outcomes by one-SNP-at-a-time association testing. Such large numbers of findings will make it increasingly difficult to determine the relative importance of a given mutation, especially in whole genome sequence data with dense, highly correlated variants. Thus, it is crucial to develop statistical approaches fitting all variants jointly and asking whether, given the LD structure of the data, there is evidence for an effect at each locus, conditional on all others.

Methods

gVAMP algorithm

Approximate message passing (AMP) was originally proposed for linear regression [11, 14, 30] assuming a Gaussian design matrix $\boldsymbol{X}$. To accommodate a wider class of structured design matrices, vector approximate message passing (VAMP) was introduced in [12]. The performance of VAMP can be precisely characterized via a deterministic, low-dimensional *state evolution* recursion, for any right-orthogonally invariant design matrix. We recall that a matrix is right-orthogonally invariant if its right singular vectors are distributed according to the Haar measure, i.e., they are uniform in the group of orthogonal matrices.

gVAMP extends EM-VAMP, introduced in [12,22,23], in which the prior parameters are adaptively learnt from the data via EM, and it is an iterative procedure consisting of two steps: *(i)* denoising, and *(ii)* linear minimum mean square error estimation (LMMSE). The denoising step accounts for the prior structure given a noisy estimate of the signal $\boldsymbol{\beta}$, while the LMMSE step utilizes phenotype values to further refine the estimate by accounting for the LD structure of the data.

A key feature of the algorithm is the so called *Onsager correction*: this is added to ensure the asymptotic normality of the noise corrupting the estimates of $\boldsymbol{\beta}$ at every iteration. Here, in contrast to MCMC or other iterative approaches, the normality is guaranteed under mild assumptions on the normalized genotype matrix. This property allows a precise performance analysis via state evolution and, consequently, the optimization of the method.

In particular, the quantity $\gamma_{1,t}$ in line 7 of Algorithm 1 is the state evolution parameter tracking the error incurred by $\boldsymbol{r}_{1,t}$ in estimating $\boldsymbol{\beta}$ at iteration t. The state evolution result gives that $\boldsymbol{r}_{1,t}$ is asymptotically Gaussian, i.e., for sufficiently large N and P, $\boldsymbol{r}_{1,t}$ is approximately distributed as $\mathcal{N}(\boldsymbol{\beta}, \gamma_{1,t}^{-1}\boldsymbol{I})$. Here, $\boldsymbol{\beta}$ represents the signal to be estimated, with the prior learned via EM steps at iteration t:

$$\beta_i \sim (1-\lambda_t)\cdot\delta_0(\cdot) + \lambda_t \cdot \sum_{l=1}^{L} \pi_{t,l}\cdot\mathcal{N}(\cdot, 0, \sigma_{t,l}^2), \quad \forall\, i = 1, \ldots, P.$$

Compared to Equation (2), the subscript t in $\lambda_t, \pi_{t,l}, \sigma_{t,l}$ indicates that these parameters change through iterations, as they are adaptively learned by the algorithm. Similarly, $\boldsymbol{r}_{2,t}$ is approximately distributed as $\mathcal{N}(\boldsymbol{\beta}, \gamma_{2,t}^{-1}\boldsymbol{I})$. The Gaussianity of $\boldsymbol{r}_{1,t}, \boldsymbol{r}_{2,t}$ is enforced by the presence of the Onsager coefficients $\alpha_{1,t}$ and $\alpha_{2,t}$, see lines 17 and 22 of Algorithm 1, respectively. We also note that $\alpha_{1,t}$

Algorithm 1 gVAMP

Input: preprocessed normalized genotype matrix $\boldsymbol{X} \in \mathbb{R}^{N \times P}$, max number of iterations N_{it}, initial estimate of effect sizes $\boldsymbol{r}_{1,0} = \boldsymbol{0}_P \in \mathbb{R}^P$, initial estimate of effect sizes precision $\gamma_{1,0} = 10^{-6} > 0$, initial estimate of noise precision $\gamma_{\epsilon,0} = 2$, initial set of parameters defining the prior distribution $\Theta_0 = \{\lambda, (\pi_i^{(0)})_{i=1}^L, (\sigma_i^{(0)})_{i=1}^L\}$, max number of variance auto-tuning steps $N_{\text{var_tune}} = 5 \in \mathbb{N}$, threshold for stopping criterion $\varepsilon = 10^{-4} > 0$, damping factor $\rho \in (0, 1)$.
for $t = 0, 1, \ldots, N_{\text{it}}$ **do**
  Denoising step
  for $k = 0, 1, \ldots, N_{\text{var_tune}}$ **do**
    $\hat{\boldsymbol{\beta}}_{1,t} = \mathbb{E}[\boldsymbol{\beta} | \boldsymbol{r}_{1,t} = \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I}), \gamma_{1,t}, \Theta_t]$
    if $t > 0$ **then**
      Variance auto-tuning step of estimation error for $\boldsymbol{\beta}$ in the denoising step, called $\gamma_{1,t}$
      EM update of the prior distribution parameters Θ, called Θ_t
      if $|\gamma_{1,t} - \gamma_{1,t}^{(\text{previous})}| < 10^{-3}$ **then**
        break
      end if
    end if
  end for
  if $t \geq 0$ **then**
    $\hat{\boldsymbol{\beta}}_{1,t} = \rho \cdot \hat{\boldsymbol{\beta}}_{1,t} + (1 - \rho) \cdot \hat{\boldsymbol{\beta}}_{1,t-1}$
  end if
  $\alpha_{1,t} = \gamma_{1,t} \cdot \langle \text{Var}[\boldsymbol{\beta} | \boldsymbol{r}_{1,t} = \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I}), \gamma_{1,t}, \Theta_t] \rangle$
  $\gamma_{2,t} = \gamma_{1,t} \cdot (1 - \alpha_{1,t})/\alpha_{1,t}$
  $\boldsymbol{r}_{2,t} = (\hat{\boldsymbol{\beta}}_{1,t} - \alpha_{1,t}\boldsymbol{r}_{1,t})/(1 - \alpha_{1,t})$

  LMMSE step
  $\hat{\boldsymbol{\beta}}_{2,t} = (\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{y} + \gamma_{2,t}\boldsymbol{r}_{2,t})$
  $\alpha_{2,t} = \gamma_{2,t} \cdot \text{Tr}[(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}]/P$
  $\gamma_{1,t+1} = \gamma_{2,t} \cdot (1 - \alpha_{2,t})/\alpha_{2,t}$
  if $t > 1$ **then**
    Variance auto-tuning step of estimation error for $\boldsymbol{\beta}$ in the LMMSE step, called $\gamma_{2,t}$
  end if
  $\boldsymbol{r}_{1,t+1} = (\hat{\boldsymbol{\beta}}_{2,t} - \alpha_{2,t}\boldsymbol{r}_{2,t})/(1 - \alpha_{2,t})$
  EM update of the estimate of γ_ϵ, called $\gamma_{\epsilon,t}$

  if $t \geq 1$ and $||\hat{\boldsymbol{\beta}}_{1,t} - \hat{\boldsymbol{\beta}}_{1,t-1}||_2/||\hat{\boldsymbol{\beta}}_{1,t-1}||_2 < \varepsilon$ **then**
    break
  end if
end for
return $\hat{\boldsymbol{\beta}}_{1,t}$

(resp. $\alpha_{2,t}$) is the state evolution parameter linked to the error incurred by $\hat{\boldsymbol{\beta}}_{1,t}$ (resp. $\hat{\boldsymbol{\beta}}_{2,t}$).

The vectors $\boldsymbol{r}_{1,t}, \boldsymbol{r}_{2,t}$ are obtained after the LMMSE step, and they are further improved via the denoising step, which respectively gives $\hat{\boldsymbol{\beta}}_{1,t}, \hat{\boldsymbol{\beta}}_{2,t}$. In the denoising step, we exploit our estimate of the approximated posterior by computing the conditional expectation of $\boldsymbol{\beta}$ with respect to $\boldsymbol{r}_{1,t}, \boldsymbol{r}_{2,t}$ in order to minimize the mean square error of the estimated effects. For example, let us focus on

the pair $(\boldsymbol{r}_{1,t}, \hat{\boldsymbol{\beta}}_{1,t})$ (analogous considerations hold for $(\boldsymbol{r}_{2,t}, \hat{\boldsymbol{\beta}}_{2,t})$). Then, we have that

$$\hat{\boldsymbol{\beta}}_{1,t} = f_t(\boldsymbol{r}_{1,t}) = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{r}_{1,t} = \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I}), \lambda_t, \{\pi_{t,l}\}_{l=1}^L, \{\sigma_{t,l}^2\}_{l=1}^L]. \tag{3}$$

Here, $f_t : \mathbb{R} \to \mathbb{R}$ denotes the denoiser at iteration t and the notation $f_t(\boldsymbol{r}_{1,t})$ assumes that the denoiser f_t is applied component-wise to elements of $\boldsymbol{r}_{1,t}$. Note that, in line 15 of Algorithm 1, we take this approach one step further by performing an additional step of damping, see "Algorithm stability" below.

From Bayes theorem, one can calculate the posterior distribution (which here has the form of a spike-and-slab mixture of Gaussians) and obtain its expectation. Hence, by denoting a generic component of $\boldsymbol{r}_{1,t}$ as r_1, it follows that

$$\begin{aligned} f_t(r_1) &= \frac{\lambda_t \cdot \sum_{l=1}^L \pi_{t,l} \cdot \frac{r_1 \cdot \sigma_{t,l}^2}{\gamma_{1,t}^{-1} + \sigma_{t,l}^2} \cdot \mathcal{N}(r_1; 0, \gamma_{1,t}^{-1} + \sigma_{t,l}^2)}{(1-\lambda_t) \cdot \mathcal{N}(r_1; 0, \gamma_{1,t}^{-1}) + \lambda_t \sum_{l=1}^L \pi_{t,l} \cdot \mathcal{N}(r_1; 0, \gamma_{1,t}^{-1} + \sigma_{t,l}^2)} \\ &= \frac{\lambda_t \cdot \sum_{l=1}^L \pi_{t,l} \cdot \frac{r_1 \cdot \sigma_{t,l}^2}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)^{3/2}} \cdot \mathrm{EXP}(\sigma_{t,l}^2)}{(1-\lambda_t) \cdot \gamma_1^{1/2} \cdot \mathrm{EXP}(0) + \lambda_t \cdot \sum_{l=1}^L \pi_{t,l} \cdot \frac{1}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)^{1/2}} \cdot \mathrm{EXP}(\sigma_{t,l}^2)}, \end{aligned} \tag{4}$$

where $\mathcal{N}(r_1; 0, \gamma_{1,t}^{-1} + \sigma_{t,l}^2)$ denotes the probability density function of a Gaussian with mean 0 and variance $\gamma_{1,t}^{-1} + \sigma_{t,l}^2$ evaluated at r_1. Furthermore, we set

$$\mathrm{EXP}(\sigma^2) = \exp\left(-\frac{r_1^2}{2} \cdot \frac{\sigma_{t,*}^2 - \sigma^2}{(\gamma_{1,t}^{-1} + \sigma^2)(\gamma_{1,t}^{-1} + \sigma_{t,*}^2)} \right),$$

with $\sigma_{t,*}^2 := \max_l(\sigma_{t,l}^2)$. This form of the denoiser is particularly convenient, as we typically deal with very sparse distributions when estimating genetic associations. We also note that the calculation of the Onsager coefficient in line 17 of Algorithm 1 requires the evaluation of a conditional variance, which is computed as the ratio of the derivative of the denoiser over the error in the estimation of the signal, i.e.,

$$\mathrm{Var}[\beta_i|(\boldsymbol{r}_{1,t})_i = \beta_i + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I}), \lambda_t, \{\pi_{t,l}\}_{l=1}^L, \{\sigma_{t,l}^2\}_{l=1}^L] = f_t'((\boldsymbol{r}_{1,t})_i)/\gamma_{1,t}. \tag{5}$$

The calculation of the derivative of f_t is detailed in Supplementary Note 2.

If one has access to the singular value decomposition (SVD) of the data matrix $\boldsymbol{X}$, the per-iteration complexity is of order $\mathcal{O}(NP)$. However, at biobank scales, performing the SVD is computationally infeasible. Thus, the linear system $(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{y} + \gamma_{2,t}\boldsymbol{r}_{2,t})$ (see line 21 of Algorithm 1) needs to be solved using an iterative method, in contrast to having an analytic solution in terms of the elements of the singular value decomposition of $\boldsymbol{X}$. In the next section, we provide details on how we overcome this issue.

Scaling up using warm-start conjugate gradients

We approximate the solution of the linear system $(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{y} + \gamma_{2,t}\boldsymbol{r}_{2,t})$ with a symmetric and positive-definite matrix via the *conjugate gradient method* (CG), see Algorithm

2 in Supplementary Note 2, which is included for completeness. If κ is the condition number of $\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I}$, the method requires $\mathcal{O}(\sqrt{\kappa})$ iterations to return a reliable approximation.

Additionally, inspired by [31], we initialize the CG iteration with an estimate of the signal from the previous iteration of gVAMP. This warm-starting technique leads to a reduced number of CG steps that need to be performed and, therefore, to a computational speed-up. However, this comes at the expense of potentially introducing spurious correlations between the signal estimate and the Gaussian error from the state evolution. Such spurious correlations may lead to algorithm instability when run for a large number of iterations (also extensively discussed below). This effect is prevented by simply stopping the algorithm as soon as the R^2 measure on the training data or the number of SE associations starts decreasing.

In order to calculate the Onsager correction in the LMMSE step of gVAMP (see line 22 of Algorithm 1), we use the Hutchinson estimator [32] to estimate the quantity $\mathrm{Tr}[(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}]/P$. We recall that this estimator is unbiased, in the sense that, if $\boldsymbol{u}$ has i.i.d. entries equal to -1 and $+1$ with the same probability, then

$$\mathbb{E}[\boldsymbol{u}^T(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}\boldsymbol{u}/P] = \mathrm{Tr}[(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}]/P.$$

Furthermore, in order to perform an EM update for the noise precision γ_ϵ one has to calculate the trace of a matrix which is closely connected to the one we have seen in the previous paragraph. In order to do so efficiently, i.e., to avoiding solving another large-dimensional linear system, we store the inverted vector $(\gamma_{\epsilon,t}\boldsymbol{X}^T\boldsymbol{X} + \gamma_{2,t}\boldsymbol{I})^{-1}\boldsymbol{u}$ and reuse it again in the EM update step (see the subparagraph on EM updates).

Algorithm stability

We find that the application of existing EM-VAMP algorithms to the UK Biobank dataset leads to diverging estimates of the signal. This is due to the fact that the data matrix (the SNP data) might not conform to the properties required in [12], especially that of right-rotational invariance. Furthermore, incorrect estimation of the noise precision in line 28 of Algorithm 1 may also lead to instability of the algorithm, as previous applications of EM-VAMP generally do not leave many hyperparameters to estimate.

To mitigate these issues, different approaches have been proposed including row or/and column normalization, damping (i.e., doing convex combinations of new and previous estimates) [33], and variance auto-tuning [23]. In particular, to prevent EM-VAMP from diverging and ensure it follows its state evolution, we empirically observe that the combination of the following techniques is crucial.

1. We perform *damping* in the space of denoised signals. Thus, line 15 of Algorithm 1 reads as

$$\hat{\boldsymbol{\beta}}_{1,t} = \rho \cdot \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{r}_{1,t}, \Theta_t] + (1-\rho) \cdot \hat{\boldsymbol{\beta}}_{1,t-1},$$

   in place of $\hat{\boldsymbol{\beta}}_{1,t} = \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{r}_{1,t}, \Theta_t]$. Here, $\rho \in (0,1)$ denotes the damping factor. This ensures that the algorithm is making smaller steps when updating a signal estimate.

2. We perform *auto-tuning* of $\gamma_{1,t}$ via the approach from [23]. Namely, in the auto-tuning step, one refines the estimate of $\gamma_{1,t}$ and the prior distribution of the effect size vector $\boldsymbol{\beta}$ by jointly

re-estimating them. If we denote the previous estimates of $\gamma_{1,t}$ and Θ with $\gamma_{1,t}^{(\text{previous})}$ and $\Theta^{(\text{previous})}$, then this is achieved by setting up an expectation-maximization procedure whose aim is to maximize

$$\mathbb{E}\big[\log p(\boldsymbol{\beta}, \boldsymbol{r}_{1,t}|\gamma_{1,t}, \Theta)|\boldsymbol{r}_{1,t}, \gamma_{1,t}^{(\text{previous})}, \Theta^{(\text{previous})}\big]$$

with respect to $\gamma_{1,t}$ and Θ.

3. We *filter* the design matrix for first-degree relatives to reduce the correlation between rows, which has the additional advantage of avoiding potential confounding of shared-environmental effects among relatives.

Estimation of the prior and noise precision via EM

The VAMP approach in [12] assumes exact knowledge of the prior on the signal $\boldsymbol{\beta}$, which deviates from the setting in which genome-wide association studies are performed. Hence, we adaptively learn the signal prior from the data using expectation-maximization (EM) steps, see lines 8 and 28 of Algorithm 1. This leverages the variational characterization of EM-VAMP [22], and its rigorous theoretical analysis presented in [23]. In this subsection, we summarize the hyperparameter estimation results derived based upon [34] in the context of our model. We find that the final update formulas for our hyperparameter estimates are as follows.

- Sparsity rate λ: We define $\{\zeta_j\}_{j=1}^P$ as:

$$\zeta_j := \frac{\lambda_t \cdot \sum_{i=1}^L \pi_{i,t} \cdot \mathcal{N}((\boldsymbol{r}_{1,t})_j; 0, \sigma_{i,t}^2 + \gamma_{1,t}^{-1})}{\lambda_t \cdot \sum_{i=1}^L \pi_{i,t} \cdot \mathcal{N}((\boldsymbol{r}_{1,t})_j; 0, \sigma_{i,t}^2 + \gamma_{1,t}^{-1}) + (1-\lambda_t) \cdot \mathcal{N}((\boldsymbol{r}_{1,t})_j; 0, \gamma_{1,t}^{-1})}, \quad \forall j = 1, \ldots, P.$$

  The intuition behind $\{\zeta_j\}_{j=1}^P$ is that each ζ_j tells what fraction of posterior probability mass was assigned to the event that it has a non-zero effect. Then, the update formula for the sparsity rate λ_{t+1} reads as

$$\lambda_{t+1} = \frac{1}{P} \sum_{j=1}^P \zeta_j.$$

- Probabilities of mixture components in the slab part $\{\pi_i\}_{i=1}^L$: We define $\{\xi_{j,i}\}_{i=1,j=1}^{L,P}$ as

$$\xi_{j,i} = \frac{\pi_{i,t} \cdot \mathcal{N}((\boldsymbol{r}_{1,t})_j; 0, \sigma_i^2 + \gamma_{1,t}^{-1})}{\sum_{l=1}^L \cdot \pi_{l,t} \cdot \mathcal{N}((\boldsymbol{r}_{1,t})_j; 0, \sigma_l^2 + \gamma_{1,t}^{-1})}, \quad \forall i = 1, \ldots, L, \ \forall j = 1, \ldots, P.$$

  The intuition behind $\{\xi_{j,i}\}_{i=1,j=1}^{L,P}$ is that each $\xi_{j,i}$ approximates the posterior probability that a marker j belongs to a mixture i conditional on the fact that it has non-zero effect. Thus, the update formula for $\pi_{i,t+1}$ reads as

$$\pi_{i,t+1} = \frac{\sum_{j=1}^P \zeta_j \xi_{j,i}}{\sum_{j=1}^P \zeta_j}, \quad \forall i = 1, \ldots, L.$$

- Variances of mixture components in the slab part $\{\sigma_i^2\}_{i=1}^L$: Using the same notation, the update formula reads as

$$\sigma_{i,t+1}^2 = \frac{\sum_{j=1}^P \zeta_j \cdot \xi_{j,i} \cdot \left[\left(\frac{(\boldsymbol{r}_{1,t})_j \cdot \gamma_{1,t}}{\gamma_{1,t} + \sigma_{i,t}^{-2}} \right)^2 + \frac{1}{\gamma_{1,t} + \sigma_{i,t}^{-2}} \right]}{\sum_{j=1}^P \zeta_j \cdot \xi_{j,i}}, \quad \forall i = 1, \ldots, L.$$

  Here we also introduce a mixture merging step, i.e., if the two mixtures are represented by variances that are close to each other in relative terms, then we merge those mixtures together. Thus, we adaptively learn the mixture number.

- Precision of the error γ_ϵ : We define $\boldsymbol{\Sigma}_t := (\gamma_{\epsilon,t} \boldsymbol{X}^T \boldsymbol{X} + \gamma_{2,t} \boldsymbol{I})^{-1}$. Then, the update formula for the estimator of γ_ϵ reads as

$$\gamma_{\epsilon,t+1} = \frac{1}{\frac{||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{2,t}||^2}{N} + \frac{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{\Sigma}_t\boldsymbol{X}^T)}{N}}.$$

  In the formula above, the term $||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{2,t}||^2/N$ takes into account the quality of the fit of the model, while the term $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{\Sigma}_t\boldsymbol{X}^T)/N$ prevents overfitting by accounting for the structure of the prior distribution of the effect sizes via the regularization term $\gamma_{2,t}$. We note that the naive evaluation of this term would require an inversion of a matrix of size $P \times P$. We again use the Hutchinson estimator for the trace to approximate this object, i.e., $\mathrm{Tr}(\boldsymbol{X}\boldsymbol{\Sigma}_t\boldsymbol{X}^T) = \mathrm{Tr}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}_t) \approx \boldsymbol{u}^T(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}_t)\boldsymbol{u}$, where $\boldsymbol{u}$ has i.i.d. entries equal to -1 and $+1$ with the same probability. Furthermore, instead of solving a linear system $\boldsymbol{\Sigma}_t\boldsymbol{u}$ with a newly generated $\boldsymbol{u}$, we re-use the $\boldsymbol{u}$ sampled when constructing the Onsager coefficient, thus saving the time needed to construct the object $\boldsymbol{\Sigma}_t\boldsymbol{u}$.

C++ code optimization

Our open-source gVAMP software (`https://github.com/medical-genomics-group/gVAMP`) is implemented in C++, and it incorporates parallelization using the OpenMP and MPI libraries. MPI parallelization is implemented in a way that the columns of the normalized genotype matrix are approximately equally split between the workers. OpenMP parallelization is done on top of that and used to further boost performance within each worker by simultaneously performing operations such as summations within matrix vector product calculations. Moreover, data streaming is employed using a lookup table, enabling byte-by-byte processing of the genotype matrix stored in PLINK format with entries encoded to a set $\{0, 1, 2\}$:

$$\left(\boxed{0 \quad 1} \boxed{0 \quad 0} \boxed{1 \quad 1} \boxed{1 \quad 0} \right) \mapsto \left(\boxed{\text{NaN}} \boxed{2} \boxed{0} \boxed{1} \right)$$

The lookup table enables streaming in the data in bytes, where every byte (8 bits) encodes the information of 4 individuals. This reduces the amount of memory needed to load the genotype matrix. In addition, given a suitable computer architecture, our implementation supports SIMD instructions which allow handling four consecutive entries of the genotype matrix simultaneously. To make the comparisons between different methods fair, the results presented in the paper do not assume usage of SIMD instructions. Additionally, we emphasize that all calculations take

un-standardized values of the genotype matrix in the form of standard PLINK binary files, but are conducted in a manner that yields the parameter estimates one would obtain if each column of the genotype matrix was standardized.

UK Biobank data

Participant inclusion

UK Biobank has approval from the North-West Multicenter Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants (`https://www.ukbiobank.ac.uk/ethics/`), and these ethical regulations cover the work in this study. Written informed consent was obtained from all participants.

Our objective is to use the UK Biobank to provide proof of principle of our approach and to compare to state-of-the-art methods in applications to biobank data. We first restrict our analysis to a sample of European-ancestry UK Biobank individuals to provide a large sample size and more uniform genetic background with which to compare methods. To infer ancestry, we use both self-reported ethnic background (UK Biobank field 21000-0), selecting coding 1, and genetic ethnicity (UK Biobank field 22006-0), selecting coding 1. We project the 488,377 genotyped participants onto the first two genotypic principal components (PC) calculated from 2,504 individuals of the 1,000 Genomes project. Using the obtained PC loadings, we then assign each participant to the closest 1,000 Genomes project population, selecting individuals with PC1 projection $\leq$ absolute value 4 and PC2 projection $\leq$ absolute value 3. We apply this ancestry restriction as we wish to provide the first application of our approach, and to replicate our results, within a sample that is as genetically homogeneous as possible. Our approach can be applied within different human groups (by age, genetic sex, ethnicity, etc.). However, combining inference across different human groups requires a model that is capable of accounting for differences in minor allele frequency and linkage disequilibrium patterns across human populations. Here, the focus is to first demonstrate that our approach provides an optimal choice for biobank analyses, and ongoing work focuses on exploring differences in inference across a diverse range of human populations. Secondly, samples are also excluded based on UK Biobank quality control procedures with individuals removed of *(i)* extreme heterozygosity and missing genotype outliers; *(ii)* a genetically inferred gender that did not match the self-reported gender; *(iii)* putative sex chromosome aneuploidy; *(iv)* exclusion from kinship inference; *(v)* withdrawn consent.

Whole genome sequence data

We process the population-level WGS variants, recently released on the UK Biobank DNAnexus platform. We use BCF tools to process thousands of pVCF files storing the chunks of DNA sequences, applying elementary filters on genotype quality (GQ ≤ 10), local allele depth (smpl_sum LAD < 8), missing genotype (F_MISSING > 0.1), and minor allele frequency (MAF < 0.0001). We select this MAF threshold as it means that on average about 80 people will have a genotype that is non-zero, which was the lowest frequency for which we felt that there was adequate power in the data to detect the variants. While we accept that it is quite possible to include additional rare variants, we wished for a conservative threshold that was at least an order of magnitude lower than the threshold

we used for the imputed SNP data described below to facilitate a comparison among the analysis of the different data types.

Simultaneously, we normalize the indels to the most recent reference, removing redundant data fields to reduce the size of the files. For all chromosomes separately, we then concatenate all the pre-processed VCF files and convert them into PLINK format. The compute nodes on the DNAnexus system are quite RAM limited, and it is not possible to analyse the WGS data outside of this system, which restricts the number of variants that can be analysed jointly. To reduce the number of variants to the scale which can be fit in the largest computational instance available on the DNAnexus platform, we rank variants by minor allele frequency and remove the variants in high LD with the most common variants using the PLINK clumping approach, setting a 1000 kb radius, and R^2 threshold to 0.36. This selects a focal common variant from a group of other common variants with correlation ≥ 0.6, which serves to capture the common variant signal into groups, whilst keeping all rare variation within the data. Finally, we remove the variants sharing the same base pair position, not keeping any of these duplicates, and merge all the chromosomes into a large data instance, including the final 16,854,878 WGS variants.

Imputed SNP data

We use genotype probabilities from version 3 of the imputed autosomal genotype data provided by the UK Biobank to hard-call the single nucleotide polymorphism (SNP) genotypes for variants with an imputation quality score above 0.3. The hard-call-threshold is 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p-value for the Hardy-Weinberg test larger than 10^{-6}, as determined in the set of unrelated Europeans) we select those with MAF ≥ 0.002 and rs identifier, in the set of European-ancestry participants, providing a dataset of 9,144,511 SNPs. From this, we took the overlap with the Estonian Genome Centre data as described in [8] to give a final set of 8,430,446 autosomal markers.

For our simulation study and UK Biobank analyses described below, we select two subsets of 8,430,446 autosomal markers. We do this by removing markers in very high LD using the "clumping" approach of PLINK, where we rank SNPs by minor allele frequency and then select the highest MAF SNPs from any set of markers with LD $R^2 \geq 0.8$ within a 1MB window to obtain 2,174,071 markers. We then further subset this with LD $R^2 \geq 0.5$ to obtain 882,727 SNP markers. This results in the selection of two subsets of "tagging variants", with only variants in very high LD with the tag SNPs removed. This allows us to compare analysis methods that are restricted in the number of SNPs that can be analysed, but still provide them a set of markers that are all correlated with the full set of imputed SNP variants, limiting the loss of association power by ensuring that the subset is correlated to those SNPs that are removed.

Whole exome sequence data burden scores

We then combine this data with the UK Biobank whole exome sequence data. The UK Biobank final release dataset of population level exome variant calls files is used (`https://doi.org/10.1101/572347`). Genomic data preparation and aggregation is conducted with custom pipeline (repo) on the UK Biobank Research Analysis Platform (RAP) with DXJupyterLab Spark Cluster App

(v. 2.1.1). Only biallelic sites and high quality variants are retained according to the following criteria: individual and variant missingness $< 10\%$, Hardy-Weinberg Equilibrium p-value $> 10^{-15}$, minimum read coverage depth of 7, at least one sample per site passing the allele balance threshold > 0.15. Genomic variants in canonical, protein coding transcripts (Ensembl VERSION) are annotated with the Ensembl Variant Effect Predictor (VEP) tool (docker image ensemblorg/ensembl-vep:release_110.1). High-confidence (HC) loss-of-function (LoF) variants are identified with the LOFTEE plugin (v1.0.4_GRCh38). For each gene, homozygous or multiple heterozygous individuals for LoF variants have received a score of 2, those with a single heterozygous LoF variant 1, and the rest 0. We chose to use the WES data to create the burden scores rather than the WGS data as existing well-tested pipelines were available.

Phenotypic records

Finally, we link these DNA data to the measurements, tests, and electronic health record data available in the UK Biobank [35] and, for the imputed SNP data, we select 7 blood based biomarkers and 6 quantitative measures which show $\geq 15\%$ SNP heritability and $\geq 5\%$ out-of-sample prediction accuracy [8]. Our focus is on selecting a group of phenotypes for which there is sufficient power to observe differences among approaches. We split the sample into training and testing sets for each phenotype, selecting 15,000 individuals that are unrelated (SNP marker relatedness < 0.05) to the training individuals to use as a testing set. This provides an independent sample of data with which to access prediction accuracy. We restrict our prediction analyses to this subset as predicting across other biobank data introduces issues of phenotypic concordance, minor allele frequency and linkage disequilibrium differences. In fact, our objective is to simply benchmark methods on as uniform a dataset as we can. As stated, combining inference across different human groups, requires a model that is capable of accounting for differences in minor allele frequency and linkage disequilibrium patterns across human populations and, while our algorithmic framework can provide the basis of new methods for this problem, the focus here is on benchmarking in the simpler linear model setting. Samples sizes and traits used in our analyses are given in Table S1.

Statistical analysis in the UK Biobank

gVAMP model parametes for WGS

We apply gVAMP to the WGS data to analyse human height using the largest computational instance currently available on the DNAnexus platform, employing 128 cores and 1921.4 GB total memory. Efficient C++ gVAMP implementation allows for parallel computing, utilizing OpenMP and MPI libraries. Here, we split the memory requirements and computational workload between 2 OpenMP threads and 64 MPI workers. For the prior initialization, we set an initial number of 22 non-zero mixtures, we let the variance of those mixtures follow a geometric progression to a maximum of $1/N$, with N the sample size, and we let the probabilities follow a geometric progression with factor $1/2$. The prior probability $1 - \lambda$ of SNP markers being assigned to the 0 mixture is initialized to 99.5%. The SNP marker effect sizes are initialised with 0. Based on the experiments in the UK Biobank imputed dataset, in WGS we set the initial damping factor ρ to 0.1, and adjust it to 0.05 for iteration 4 onward, stabilizing the algorithm. We then report the results corresponding

to the iterate having the largest number of SE associations. We also note that, after the first few iterations, the number of SE associations is typically rather stable (see Figure S4).

gVAMP model parameters for imputed SNP data

We run gVAMP on the 13 UK Biobank phenotypes on the full 8,430,446 SNP set, and on the 2,174,071 and 882,727 LD clumped SNP set. We find that setting the damping factor ρ to 0.1 performs well for all the 13 outcomes in the UK Biobank that we have considered. For the prior initialization, we set an initial number of 22 non-zero mixtures, we let the variance of those mixtures follow a geometric progression to a maximum of $1/N$, with N the sample size, and we let the probabilities follow a geometric progression with factor 1/2. The SNP marker effect sizes are initialised with 0. This configuration works well for all phenotypes. We also note that our inference of the number of mixtures, their probabilities, their variances and the SNP marker effects is not dependent upon specific starting parameters for the analyses of the 2,174,071 and 882,727 SNP datasets, and the algorithm is rather stable for a range of initialization choices. Similarly, the algorithm is stable for different choices of the damping ρ, as long as said value is not too large.

Generally, appropriate starting parameters are not known in advance and this is why we learn them from the data within the EM steps of our algorithm. However, it is known that EM can be sensitive to the starting values given and, thus, we recommend initialising a series of models at different values to check that this is not the case (similar to starting multiple Monte Carlo Markov chains in standard Bayesian methods). The feasibility of this recommendation is guaranteed by the significant speed-up of our algorithm compared to existing approaches, see Supplementary Note 1, Figure S5c.

For the sparsity parameter, we consider either initializing it to $50,000$ included signals ($\lambda_0 = 50,000/P$), or to further increase the probability of SNP markers being assigned to the 0 mixture to 97%, which results in a sparser initialised model. We also consider inflating the variances to a maximum of $10/N$ to allow for an underlying effect size distribution with longer tails. It is trivial to initialise a series of models and to monitor the training R^2, SNP heritability, and residual variance estimated within each iteration over the first 10 iterations. Given the same data, gVAMP yields estimates that more closely match GMRM when convergence in the training R^2, SNP heritability, residual variance, and out-of-sample test R^2 are smoothly monotonic within around 10-40 iterations. Following this, training R^2, SNP heritability, residual variance, and out-of-sample test R^2 may then begin to slightly decrease as the number of iterations becomes large. Thus, as a stopping criterion for the 2,174,071 and 882,727 SNP datasets, we choose the iteration that maximizes the training R^2, and in practice it is easy to optimise the algorithm to the data problem at hand.

We highlight the iterative nature of our method. Thus, improved computational speed and more rapid convergence is achieved by providing better starting values for the SNP marker effects. Specifically, when moving from 2,174,071 to 8,430,446 SNPs, only columns with correlation $R^2 \geq 0.8$ are being added back into the data. Thus, for the 8,430,446 SNP set, we initialise the model with the converged SNP marker and prior estimates obtained from the 2,174,071 SNP runs, setting to 0 the missing markers. Furthermore, we lower the value of the damping factor ρ, with typical values being 0.05 and 0.01. We experiment both with using the noise precision from the initial 2,174,071 SNP runs and with setting it to 2. We then choose the model that leads to a smoothly monotonic

curve in the training R^2. We observe that SNP heritability, residual variance, and out-of-sample test R^2 are also smoothly monotonic within 25 iterations. Thus, as a stopping criterion for the 8,430,446 SNP dataset, we choose the estimates obtained after 25 iterations for all the 13 traits. We follow the same process when extending the analyses to include the WES rare burden gene scores.

Polygenic risk scores and SNP heritability

gVAMP produces SNP effect estimates that can be directly used to create polygenic risk scores. The estimated effect sizes are on the scale of normalised SNP values, i.e., $(\boldsymbol{X}_j - \mu_{X_j})/SD(\boldsymbol{X}_j)$, with μ_{X_j} the column mean and $SD(\boldsymbol{X}_j)$ the standard deviation, and thus SNPs in the out-of-sample prediction data must also be normalized. We provide an option within the gVAMP software to do phenotypic prediction, returning the adjusted prediction R^2 value when given input data of a PLINK file and a corresponding file of phenotypic values. gVAMP estimates the SNP heritability as the phenotypic variance (equal to 1 due to normalization) minus 1 divided by the estimate of the noise precision, i.e., $h^2_{SNP} = 1 - 1/\gamma_\epsilon$.

We compare gVAMP to a MCMC sampler approach (GMRM) with a similar prior (the same number of starting mixtures) as presented in [8]. We select this comparison as the MCMC sampler was demonstrated to exhibit the highest genomic prediction accuracy up to date [8]. We run GMRM for 2000 iterations, taking the last 1800 iterations as the posterior. We calculate the posterior means for the SNP effects and the posterior inclusion probabilities of the SNPs belonging to the non-zero mixture group. GMRM estimates the SNP heritability in each iteration by sampling from an inverse χ^2 distribution using the sum of the squared regression coefficient estimates.

We also compare gVAMP to the summary statistics prediction methods LDpred2 [6] and SBayesR [7] run on the 2,174,071 SNP dataset. In fact, we find that running on the full 8,430,446 SNP set is either computationally infeasible or entirely unstable, and we note that neither approach has been applied to data of this scale to date. For SBayesR, following the recommendation on the software webpage (`https://cnsgenomics.com/software/gctb/#SummaryBayesianAlphabet`), after splitting the genomic data per chromosomes, we calculate the so-called *shrunk* LD matrix, which use the method proposed by [36] to shrink the off-diagonal entries of the sample LD matrix toward zero based on a provided genetic map. We make use of all the default values: `--genmap-n` 183, `--ne` 11400 and `--shrunk-cutoff` 10^{-5}. Following that, we run the SBayesR software using summary statistics generated via the REGENIE software (see "Mixed linear association testing" below) by grouping several chromosomes in one run. Namely, we run the inference jointly on the following groups of chromosomes: $\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}, \{7, 8\}, \{9, 10, 11\}, \{12, 13, 14\}$ and $\{15, 16, 17, 18, 19, 20, 21, 22\}$. This allows to have locally joint inference, while keeping the memory requirements reasonable. All the traits except for Blood cholesterol (CHOL) and Heel bone mineral density T-score (BMD) give non-negative R^2; CHOL and BMD are then re-run using the option to remove SNPs based on their GWAS p-values (threshold set to 0.4) and the option to filter SNPs based on LD R-Squared (threshold set to 0.64). For more details on why one would take such an approach, one can check `https://cnsgenomics.com/software/gctb/#FAQ`. As the obtained test R^2 values are still similar, as a final remedy, we run standard linear regression over the per-group predictors obtained from SBayesR on the training dataset. Following that, using the learned parameters, we make a linear combination of the per-group predictors in the test dataset to obtain

the prediction accuracy given in the table.

For LDpred2, following the software recommendations, we create per-chromosome banded LD matrices with the window size of 3cM. After the analysis of the genome-wide run of LDpred2, we establish that the chains do not converge even after tuning the shrinkage factor, disabling the sign jump option and disabling the usage of MLE (`use_MLE=FALSE` option). For this reason, we opt to run LDpred2 per chromosome, in which case the chains converge successfully. Twenty chains with different proportion of causal markers are run in the LDpred2 method, for each of the chromosomes independently. Then, a standard linear regression involving predictors from different chromosomes is performed to account for correlations between SNPs on different chromosomes, which achieved better test R^2 than the predictors obtained by stacking chromosomal predictors. In summary, for both LDpred2 and SBayesR we have tried to find the optimal solution to produce the highest possible out-of-sample prediction accuracy, contacting the study authors, if required, for guidance.

Mixed linear model association testing

We conduct mixed linear model association testing using a leave-one-chromosome-out (LOCO) estimation approach on the 8,430,446 and 2,174,071 imputed SNP markers. LOCO association testing approaches have become the field standard and they are two-stage: a subset of markers is selected for the first stage to create genetic predictors; then, statistical testing is conducted in the second stage for all markers one-at-a-time. We consider REGENIE [1], as it is a recent commonly applied approach. We also compare to GMRM [8], a Bayesian linear mixture of regressions model that has been shown to outperform REGENIE for LOCO testing. For the first stage of LOCO, REGENIE is given 887,060 markers to create the LOCO genetic predictors, even if it is recommended to use 0.5 million genetic markers. We compare the number of significant loci obtained from REGENIE to those obtained if one were to replace the LOCO predictors with: *(i)* those obtained from GMRM using the LD pruned sets of 2,174,071 and 887,060 markers; and *(ii)* those obtained from gVAMP at all 8,430,446 markers and the LD pruned sets of 2,174,071 and 887,060 markers. We note that obtaining predictors from GMRM at all 8,430,446 markers is computationally infeasible, as using the LD pruned set of 2,174,071 markers already takes GMRM several days. In contrast, gVAMP is able to use all 8,430,446 markers and still be faster than GMRM with the LD pruned set of 2,174,071 markers.

LOCO testing does not control for linkage disequilibrium within a chromosome. Thus, to facilitate a simple, fair comparison across methods, we clump the LOCO results obtained with the following PLINK commands: `--clump-kb` 5000 `--clump-r2` 0.01 `--clump-p1` 0.00000005. Therefore, within 5Mb windows of the DNA, we calculate the number of independent associations (squared correlation ≤ 0.01) identified by each approach that pass the genome-wide significance testing threshold of $5 \cdot 10^{-8}$. As LOCO can only detect regions of the DNA associated with the phenotype and not specific SNPs, given that it does not control for the surrounding linkage disequilibrium, a comparison of the number of uncorrelated genome-wide significance findings is conservative.

gVAMP SE association testing

We provide an alternative approach to association testing, which we call *state evolution p-value testing* (SE association testing), where the effects of each marker can be estimated conditional on all other genetic variants genome-wide. Relying on the properties of the EM-VAMP estimator, whose noise is asymptotically Gaussian due to the Onsager correction [12], we have $\boldsymbol{r}_{1,t} \approx \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I})$, where $\boldsymbol{\beta}$ is the ground-truth value of the underlying genetic effects vector. More precisely, one can show that $\frac{1}{N}\|\boldsymbol{r}_{1,t} - \boldsymbol{\beta} - \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I})\| \to 0$, as $N, P \to \infty$, with the ratio N/P being fixed. Therefore, for each marker with index j, a one-sided p-value for the hypothesis test $\mathrm{H0} : \beta_j = 0$ is given by $\Phi(-|(\boldsymbol{r}_{1,t})_i| \cdot \gamma_{1,t}^{1/2})$, where Φ is the CDF of a standard normal distribution and $(\boldsymbol{r}_{1,t})_i$ denotes the i-th component of the vector $\boldsymbol{r}_{1,t}$. We conduct this association testing for height in the WGS data and for the full 8,430,446 imputed SNP markers for the empirical UK Biobank analysis of 13 traits, using the estimates of $\boldsymbol{r}_{1,t}$. We remark that the testing results are generally stable after 20 iterations (Figure S4). To these, we apply a Bonferroni multiple testing correction to give a conservative comparison for presentation, but we note that the estimates made are joint, rather than marginal, and thus FDR control methods may also be an alternative.

Data availability

This project uses the UK Biobank data under project number 35520. UK Biobank genotypic and phenotypic data is available through a formal request at (`http://www.ukbiobank.ac.uk`). All summary statistic estimates are released publicly on Dryad: `https://doi.org/xx.xxxx/dryad.xxxxxxxxx`.

Code availability

The gVAMP code `https://github.com/medical-genomics-group/gVAMP` is fully open source. The scripts used to execute the model are available at `https://github.com/medical-genomics-group/gVAMP`. R version 4.2.1 is available at `https://www.r-project.org/`. PLINK version 1.9 is available at `https://www.cog-genomics.org/plink/1.9/`. REGENIE is available at `https://github.com/rgcgithub/regenie`. bignspr 1.12.4 package that contains LDpred2 is available at `https://privefl.github.io/bigsnpr/index.html`. SBayesR is available at `https://cnsgenomics.com/software/gctb/#Overview`.

Acknowledgements

We would like to thank Malgorzata Borczyk for creating the gene burden scores. We thank Robin Beaumont, Amedeo Roberto Esposito, Gareth Hawkes, Philip Schniter, Matthew Stephens, Pragya Sur, Peter Visscher, Michael Weedon and Harry Wright for providing valuable suggestions and comments on earlier versions of the work. We would like to acknowledge the participants and investigators of the UK Biobank study. This project was funded by a Lopez-Loreta Prize to MM, by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria. High-performance computing was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

Author contributions

MM and MRR conceived the study. AD, MM and MRR designed the study. AD derived the model and the algorithm, with input from MM and MRR. AD wrote the software, with input from MM and MRR. AD, JB, MM, and MRR conducted the analysis and wrote the paper. All authors approved the final manuscript prior to submission.

Ethical approval declaration

This project uses UK Biobank data under project 35520. UK Biobank genotypic and phenotypic data is available through a formal request at `http://www.ukbiobank.ac.uk`. The UK Biobank has ethics approval from the North West Multi-centre Research Ethics Committee (MREC). Methods were carried out in accordance with the relevant guidelines and regulations, with informed consent obtained from all participants.

References

1. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097–1103 (2021).

2. Loh, P.-R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

3. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).

4. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics* **53**, 1616–1621 (2021).

5. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273–1300 (2020).

6. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).

7. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications* **10**, 5086 (2019).

8. Orliac, E. J. *et al.* Improving gwas discovery and genomic prediction accuracy in biobank data. *Proceedings of the National Academy of Sciences* **119**, e2121279119 (2022).

9. Spence, J. P., Sinnott-Armstrong, N., Assimes, T. L. & Pritchard, J. K. A flexible modeling and inference framework for estimating variant effect sizes from gwas summary statistics. *bioRxiv* (2022).

10. Lawson, D. J. *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human Genetics* **139**, 23–41 (2020).

11. Donoho, D. L., Maleki, A. & Montanari, A. Message Passing Algorithms for Compressed Sensing. *Proceedings of the National Academy of Sciences* **106**, 18914–18919 (2009).

12. Rangan, S., Schniter, P. & Fletcher, A. K. Vector approximate message passing. *IEEE Transactions on Information Theory* **65**, 6664–6684 (2019).

13. Feng, O. Y., Venkataramanan, R., Rush, C., Samworth, R. J. *et al.* A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning* **15**, 335–536 (2022).

14. Bayati, M. & Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory* **57**, 764–785 (2011).

15. Montanari, A. & Venkataramanan, R. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics* **45**, 321–345 (2021).

16. Barbier, J., Krzakala, F., Macris, N., Miolane, L. & Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences* **116**, 5451–5460 (2019).

17. Barbier, J., Camilli, F., Mondelli, M. & Sáenz, M. Fundamental limits in structured principal component analysis and how to reach them. *Proceedings of the National Academy of Sciences* **120** (2023).

18. Jeon, C., Ghods, R., Maleki, A. & Studer, C. Optimality of large MIMO detection via approximate message passing. In *IEEE International Symposium on Information Theory*, 1227–1231 (2015).

19. Metzler, C. A., Maleki, A. & Baraniuk, R. G. From denoising to compressed sensing. *IEEE Trans. Information Theory* **62**, 5117–5144 (2016).

20. Eksioglu, E. M. & Tanc, A. K. Denoising AMP for MRI reconstruction: BM3D-AMP-MRI. *SIAM Journal on Imaging Sciences* **11**, 2090–2109 (2018).

21. Zhong, X., Su, C. & Fan, Z. Empirical bayes pca in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 853–878 (2022).

22. Fletcher, A. K. & Schniter, P. Learning and free energies for vector approximate message passing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4247–4251 (2017).

23. Fletcher, A. K., Sahraee-Ardakan, M., Rangan, S. & Schniter, P. Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

24. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

25. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* **54**, 263–273 (2022).

26. Robinson, M. R. *et al.* Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature Genetics* **49**, 1174–1181 (2017).

27. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature Communications* **12**, 1164 (2021).

28. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nature Reviews Genetics* **19**, 175–185 (2018).

29. Campos, A. I. *et al.* Boosting the power of genome-wide association studies within and across ancestries by using polygenic scores. *Nature Genetics* **55**, 1769–1776 (2023).

30. Krzakala, F., Mézard, M., Sausset, F., Sun, Y. & Zdeborová, L. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment* **2012**, P08009 (2012).

31. Skuratovs, N. & Davies, M. E. Warm-starting in message passing algorithms. In *2022 IEEE International Symposium on Information Theory (ISIT)*, 1187–1192 (2022).

32. Hutchinson, M. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation* **19**, 433–450 (1990).

33. Vila, J., Schniter, P., Rangan, S., Krzakala, F. & Zdeborová, L. Adaptive damping and mean removal for the generalized approximate message passing algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021–2025 (2015).

34. Vila, J. & Schniter, P. Expectation-maximization gaussian-mixture approximate message passing. *IEEE Transactions on Signal Processing* **61** (2012).

35. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

36. Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The Annals of Applied Statistics* **4**, 1158 – 1182 (2010).

37. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014).

38. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLOS ONE* **3**, 1–8 (2008).

39. Patxot, M. *et al.* Probabilistic inference of the genetic architecture underlying functional enrichment of complex traits. *Nature Communications* **12**, 6972 (2021).

Supplementary information

Joint modelling of whole genome sequence data for human height via approximate message passing

Al Depope, Jakub Bajzik, Marco Mondelli, Matthew R. Robinson

Supplementary Tables

Table S1. The 13 UK Biobank traits used within the study. Phenotypic names and their codes used in the study. The sample size, N, gives the number of individuals with training data measures.

Phenotype	Code	Sample size, N
Blood: cholesterol	CHOL	395,025
Blood: eosinophill count	EOSI	401,452
Blood: glycated haemoglobin	HbA1c	394,912
Blood: High density lipoprotein	HDL	360,286
Blood: mean corpuscular haemoglobin	MCH	402,201
Blood mean corpuscular volume	MCV	402,202
Red blood cell count	RBC	402,204
Body mass index	BMI	413,595
Diastolic blood pressure	DBP	377,358
Forced vital capacity	FVC	376,724
Heel bone mineral density	BMD	231,693
Standing height	HT	414,055
Systolic blood pressure	SBP	377,347

Supplementary Figures

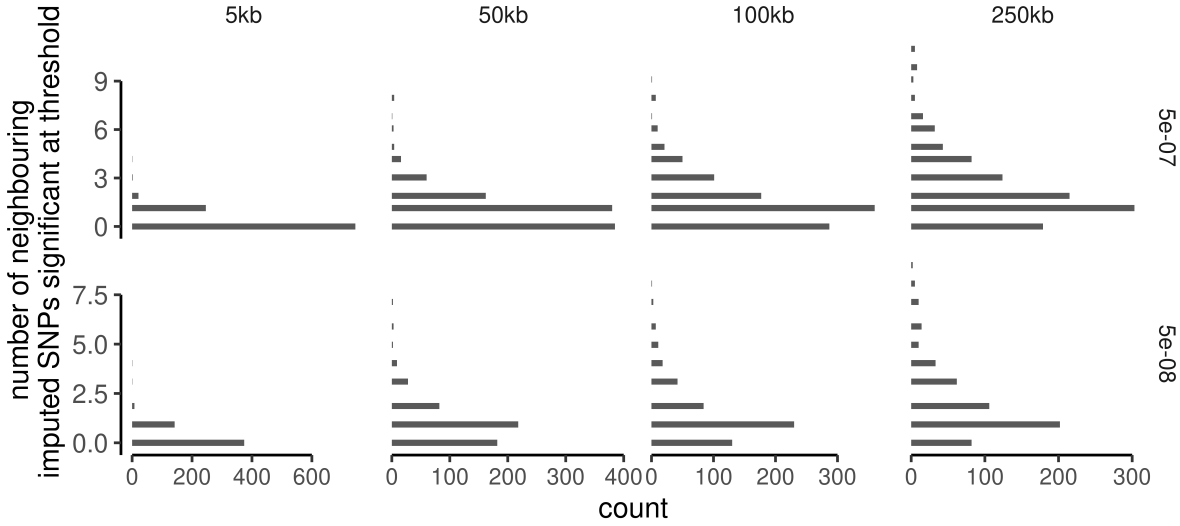

Figure S1. Whole genome sequence variants discovered at two different significance thresholds that are not discovered in imputed SNP data can have multiple neighbouring imputed SNPs that are discovered as significantly height associated. For each whole genome sequence variant discovered as height associated at $p \leq 5 \cdot 10^{-8}$ or $p \leq 5 \cdot 10^{-7}$, we determine the number of imputed SNPs determined to be significantly height associated at the same significance level, for a base-pair distance of either 5kb, 50kb, 100kb, or 250kb from the focal WGS variant. We observe that most WGS findings have 0 neighboring findings in close proximity, but can have multiple neighboring significant imputed variant findings at distance.

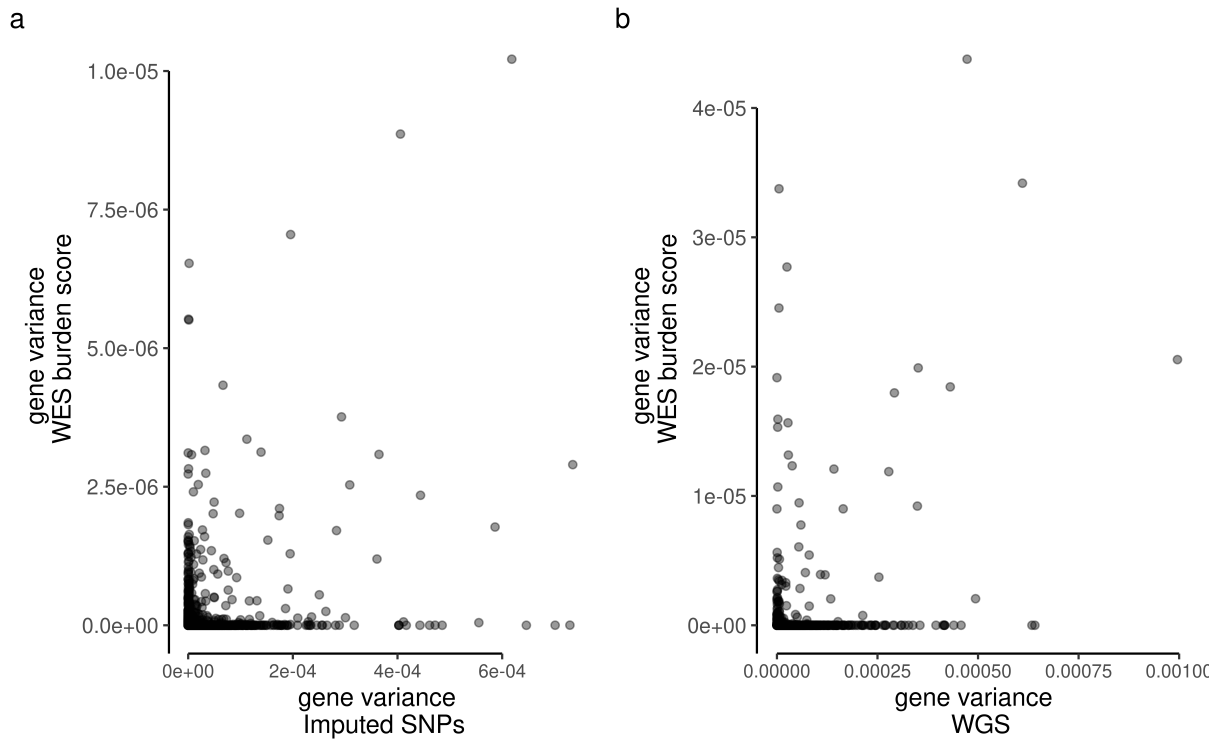

Figure S2. The variance attributable to gene burden scores calculated from whole exome sequence data (y-axis) shows no relationship to the variance attributable to DNA variants within and around the gene (x-axis) for either imputed SNPs or whole genome sequence variants.

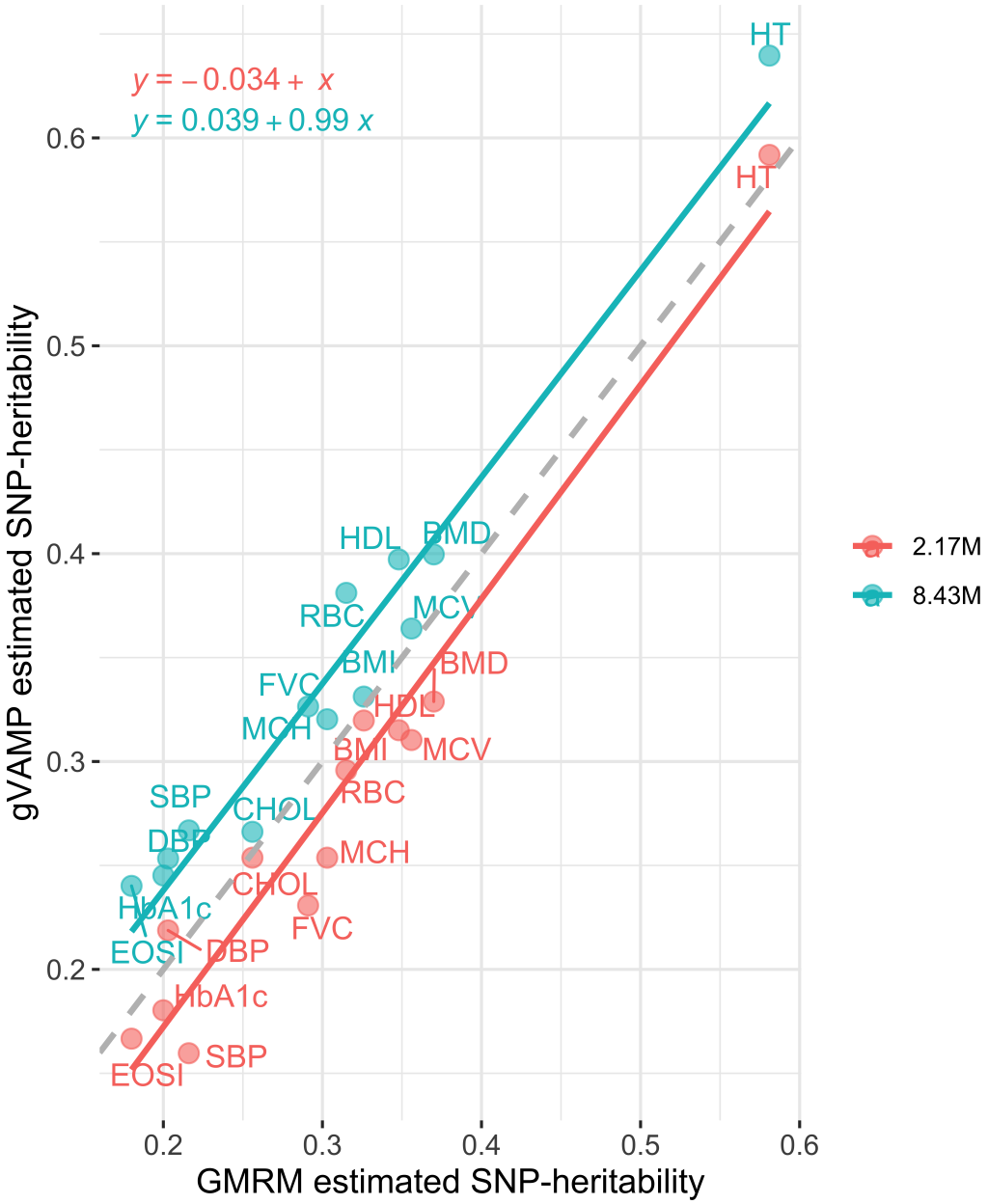

Figure S3. SNP heritability estimation of GMRM versus gVAMP with different numbers of SNP markers across 13 trait in the UK Biobank. Comparison of the proportion of phenotypic variation attributable to 2,174,071 autosomal SNP genetic markers (SNP heritability) estimated by GMRM (x-axis) to the SNP heritability estimated by gVAMP (y-axis) at either the same 2,174,071 SNPs (red) or 8,430,446 SNP markers (blue). The slope of the lines shows a 1-to-1 relationship of gVAMP to GMRM, but with an average of 3.4% lower estimate for gVAMP at 2.17M SNPs. Analysing 8.4M SNPs with gVAMP increases the heritability estimate over GMRM by 3.9%, which is consistent with an increase in phenotypic variance captured by the full imputed sequence data, as opposed to a selected subset of SNP markers. The dashed grey line gives $y = x$.

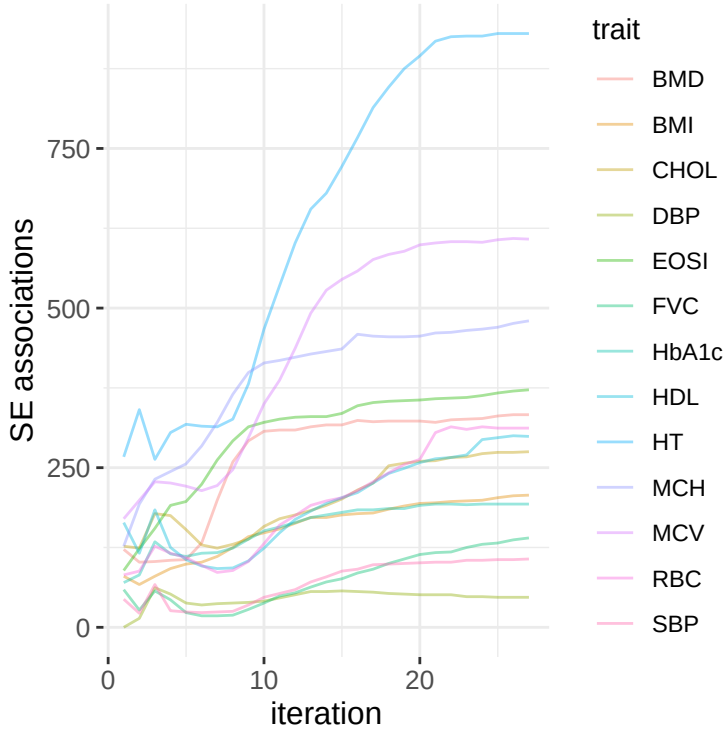

Figure S4. Convergence of SE p-value testing with increasing number of iterations for 13 UK Biobank traits. AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers. We show this SE p-value testing approach for each iteration of our iterative algorithm, where we calculate the number of genome-wide fine-mapped associations for 13 UK Biobank traits at a p-value threshold of less than $5 \cdot 10^{-8}$ for all 8,430,446 SNP markers.

Supplementary Note 1

Simulation study methods

To support our empirical analyses we conduct a simulation study using the 8,430,446 UK Biobank genetic marker data with 414,055 individuals. We randomly sample 40,000 causal variants genome-wide to give a highly polygenic genetic basis. To these, we allocate effect sizes from a Gaussian with mean zero and variance 0.5/40000, where 0.5 is the proportion of variance attributable to the SNP markers (SNP heritability). Multiplying the simulated SNP effects by normalized values of the 40,000 causal markers, gives a vector of genetic values of length $N = 414055$ with variance 0.5. To this we add a vector of noise, drawn from a Gaussian with mean zero and variance 0.5, to produce a response variable of length N, with zero mean and unit variance.

We analyse the simulated response variable with gVAMP, using either 8,430,446, 2,174,071 or 887,060 SNP markers with identical initialisation to that described in the Methods for the empirical UK Biobank study. We also analyse the data with GMRM using 2,174,071 or 887,060 SNP markers (as this completes within reasonable compute time and resource use), running for 2,500 iterations with 500 iteration burn in. Finally, we run REGENIE using 887,060 SNP markers for the first stage and 8,430,446 SNP markers for the second stage LOCO testing.

We begin by comparing the LOCO association testing results obtained by REGENIE to those obtained by replacing the REGENIE predictors with predictors obtained from GMRM using 2,174,071 markers and gVAMP using either 8,430,446, 2,174,071 or 887,060 SNP markers within the gVAMP software.

To facilitate a simple, fair comparison of the true positive rate (TPR) and false discovery rate (FDR) across methods, we clump the LOCO results obtained with the following PLINK commands: `--clump-kb` 5000 `--clump-r2` 0.01 `--clump-p1` 0.00000005. Therefore, within 5Mb windows of the DNA, we calculate the number of independent associations (squared correlation ≤ 0.01) identified by each approach that pass the genome-wide significance testing threshold of $5 \cdot 10^{-8}$. This is the same procedure performed for MLMA testing (see "Mixed linear model association testing" in the Methods). For each identified genome-wide significant association, we then ask if it is correlated (squared correlation ≥ 0.01) to a causal variant: if so, we classify it as a true positive; otherwise, we classify it as a false positive. The true positive rate is calculated as the number of true positives divided by the total number of simulated causal variants, and it is also known as the recall, or sensitivity, reflecting the power of a statistical test. The false discovery rate is calculated as the number of false positives divided by the number of genome-wide significant associations, and it is a measure of the proportion of discoveries that are false. As genome-wide association studies aim to detect regions of the DNA associated with the phenotype, the definition of a false discovery as the detection of a variant at genome-wide significance when that variant has squared correlation ≤ 0.01 with a causal variant within 5Mb is a very conservative one. We present these results in Figure S5a.

We then compare the out-of-sample prediction accuracy and the SNP heritability estimated by GMRM with that obtained by gVAMP, following the same procedures outlined in the Methods for the empirical UK Biobank analysis. For the out-of-sample prediction, we use a hold-out set of 15,000 individuals that are unrelated (SNP marker relatedness < 0.05) to the training individuals. We present these results in Figure S5b.

We conduct five simulation replicates, as we find that this is sufficient to contrast methods, with GMRM and gVAMP giving very consistent estimates across replicates, and REGENIE being highly variable. We compare the run time for the first stage analysis of REGENIE to the total run times of gVAMP and GMRM across different marker sets using 50 CPU from a single AMD compute node. We present these run time results in Figure S5c.

Additionally, we compare the SE p-value testing results of gVAMP on the 8,430,446 and 2,174,071 SNP datasets to the posterior inclusion probabilities calculated for each SNP using GMRM. The theoretical expectation is that both methods should yield broadly similar results, but in practice p-value association testing and posterior inclusion probability testing are not easily comparable. Thus, we simply present TPR and FDR calculations for these models at different significance thresholds in Figure S8. A true positive is defined as an SNP that *(i)* has a test statistic passing the threshold, and *(ii)* is a true causal variant. This reflects power to localise marker effects to the single-locus level. A false discovery is classified as a SNP that *(i)* has a test statistic passing the threshold, and *(ii)* is not the exact true causal variant. Our objective here is to simply explore the power and FDR of the SE testing across a range of thresholds. We avoid prescribing specific significance thresholds, leaving this as a choice for practitioners.

To support our findings further, we repeat our simulation again but we randomly select 40,000 causal variants from the 887,060 markers. Our objective is to compare REGENIE and gVAMP in the scenario where the causal variants are present in the data used to create the predictors for the first step of LOCO. This ensures that our findings are not just driven by only having SNPs correlated with the causal variants in step 1. Additionally, as well as simulating the causal marker effects from a Gaussian, we also simulate them from a mixture of Gaussians. Specifically, we simulate effect sizes for the 40,000 causal variants from a mixture of three Gaussian distributions with probabilities $1/2, 1/3, 1/6$ and variances 0.5/40,000, 5/40,000, 50/40,000. Multiplying the simulated SNP effects by the normalized values of the 40,000 causal markers gives a vector of genetic values of length $N = 414,055$ with variance 0.5. To this we add a vector of noise, drawn from a Gaussian with mean zero and variance 0.5, to produce a response variable of length N, with zero mean and unit variance. We conduct five simulation replicates for the Gaussian effect size setting and five for the mixture setting, because we again find that this is sufficient to contrast methods, with gVAMP giving very consistent estimates across replicates and REGENIE being highly variable. We present these results to compare the TPR and FDR of REGENIE with that of gVAMP in Figure S6.

Finally, we repeat our simulation once more but we randomly select 40,000 causal variants from the 2,174,071 SNP data. Our objective is to compare GMRM and gVAMP to empirically assess the Bayes optimality of gVAMP when applied to genomic data. We simulate the causal marker effects from both a Gaussian and a mixture of Gaussians, and compare SNP heritability of the two methods under these different effect size distributions. We present these results in Figure S7.

Simulation study results

We start by discussing the LOCO testing results for the setting in which 40,000 SNP markers are randomly selected from the full set of 8,430,446 SNPs. We find that gVAMP performs similarly to the individual-level Bayesian approach of GMRM in true positive rate (TPR), whilst controlling the false discovery rate (FDR) below the 5% level (Figure S5a). Both approaches outperform the

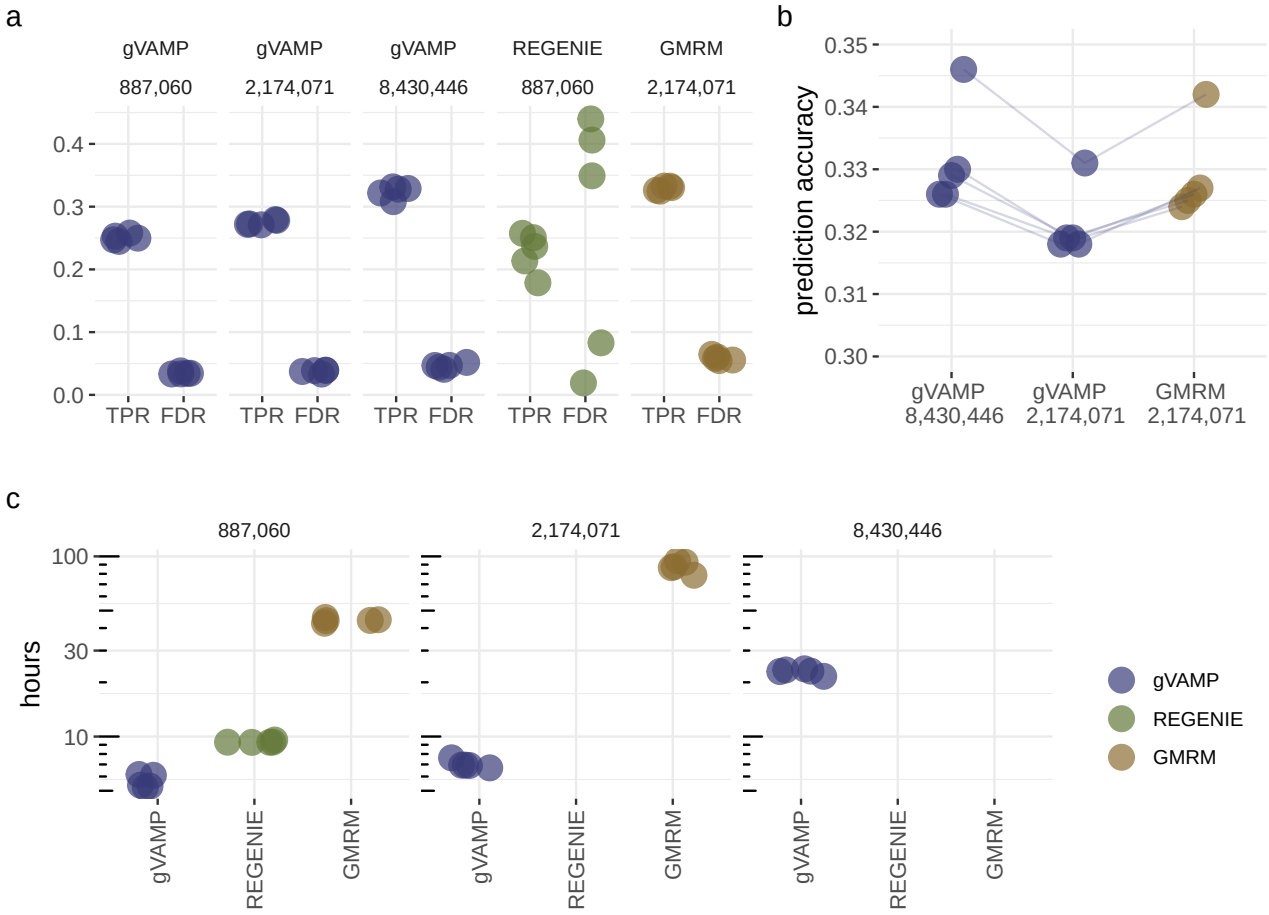

Figure S5. Simulation study of association testing power and run time using UK Biobank genotype data. We consider 8,430,446 SNP markers, randomly select 40,000 as causal and use these to simulate a phenotype. Standard leave-one-chromosome-out (LOCO) association testing approaches are two-stage, with a subset of markers selected for the first stage. Here we select either all markers, 2,174,071 markers, or 887,060 markers for the first stage and then use all markers for the second stage LOCO testing. In *(a)*, we apply gVAMP, REGENIE, or GMRM to these data and calculate the true positive rate (TPR) and the false discovery rate (FDR). In the first stage, we set REGENIE to utilize only 887,060 markers, despite only 500,000 being recommended (see `https://rgcgithub.github.io/regenie/faq/`), GMRM up to 2,174,071 markers, whilst gVAMP can utilise the full range. The FDR is well controlled at 5% or less for both gVAMP and GMRM, but not for REGENIE. Power (TPR) is higher for gVAMP and GMRM as compared to REGENIE. For *(b)*, we compare out-of-sample prediction accuracy for polygenic risk scores created at different sets of markers from gVAMP (8,430,446 and 2,174,071) and GMRM (2,174,071). *(c)* gives the run time in hours for the first stage analysis of gVAMP, REGENIE, and GMRM, across different marker sets using 50 CPU from a single compute node. gVAMP takes 2/3 of the time of a single-trait analysis in REGENIE using 887,060 markers, remains faster then REGENIE using 2,174,071 markers, and is the only approach capable of analysing 8,430,446 markers jointly within 24 hours.

commonly used REGENIE software in both TPR and FDR, which does not always control the FDR below 5% (Figure S5a). We repeat our simulation by selecting 40,000 causal SNPs from the 887,060 marker subset so that the causal variants are within the set used for the first step of all methods, finding that the results remain the same across two different effect size distributions (Figure S6). Thus, power and accuracy are higher for gVAMP and GMRM as compared to REGENIE for two reasons: *(i)* given the same data, the models show improved performance (Figure S6), and *(ii)*

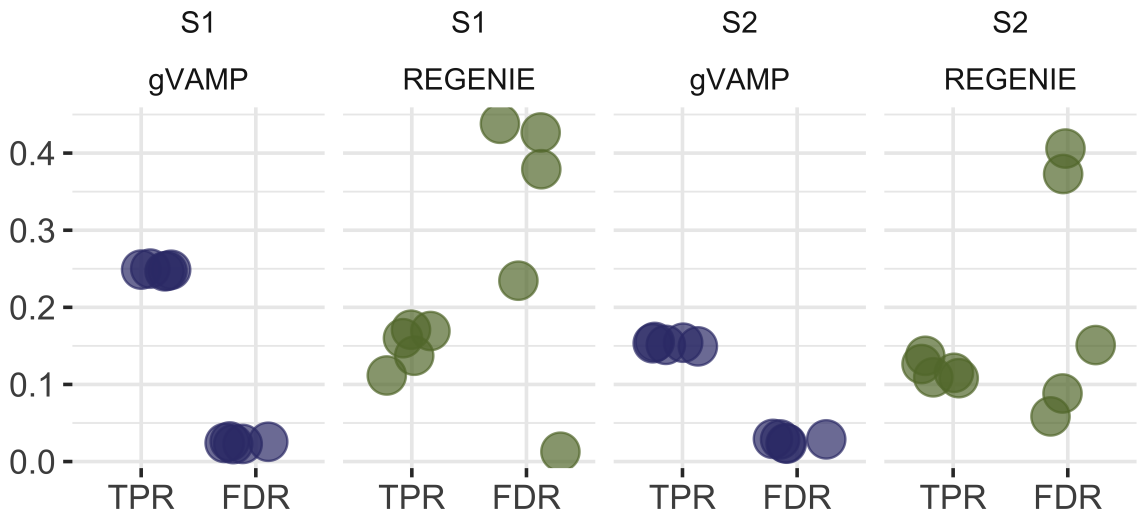

Figure S6. Comparison of gVAMP and REGENIE association testing within identical data. True positive rate (TPR) and false discovery rate (FDR) for leave-one-chromosome-out (LOCO) testing where 887,060 markers are used for both the first step of REGENIE and for gVAMP and where all simulated causal variants are contained within this set. LOCO testing is then conducted over the full set of 8,430,446 SNP markers. "S1" refers to causal variant effects simulated from a Gaussian distribution; "S2" refers to causal variant effects whose distribution is a mixture of Gaussians. We perform five simulation replicates.

more SNP markers can be utilised to create the predictors, with the benefit of controlling for all genome-wide effects rather than a subset, which in turn controls the FDR (Figure S5a and S6).

For MLMA, association testing power (TPR) depends upon the sample size and the out-of-sample prediction accuracy of the predictors obtained from the first step [37]. For gVAMP to have Bayes-optimal empirical performance, polygenic risk score prediction accuracy should match that of GMRM. When simulating data by selecting 40,000 causal markers from 8,430,446 imputed SNP markers and then only using a subset of 2,174,071 markers for analysis, we find that gVAMP loses only 0.5% to 1% accuracy over GMRM. However, we highlight that, by analysing all 8,430,446 imputed SNP markers, gVAMP improves over GMRM (Figure S5b). We note that analysing all 8,430,446 SNPs is computationally infeasible for GMRM.

A key feature of gVAMP is its computational efficiency, which allows for joint processing of the full set of 8,430,446 markers. gVAMP completes in 2/3 of the time of REGENIE given the same data and compute resources, and it is dramatically faster (12.5× speed-up) than the MCMC sampling algorithm GMRM; even with 8,430,446 imputed SNP markers, the model yields estimates in under a day (Figure S5c).

Polygenic risk score prediction accuracy depends upon h^2_{SNP}, the number of true underlying causal variants and the sample size [38], which are fixed in our simulation. When simulating effects over 40,000 SNPs randomly selected from 8,430,446 markers and then using only a subset of 2,174,071 markers to estimate h^2_{SNP}, both gVAMP and GMRM give estimates that are lower than the simulated value, which is expected as all causal variants are not given to the model (Figure S7). gVAMP gives correct estimates when given the full 8,430,446 markers and when we repeat

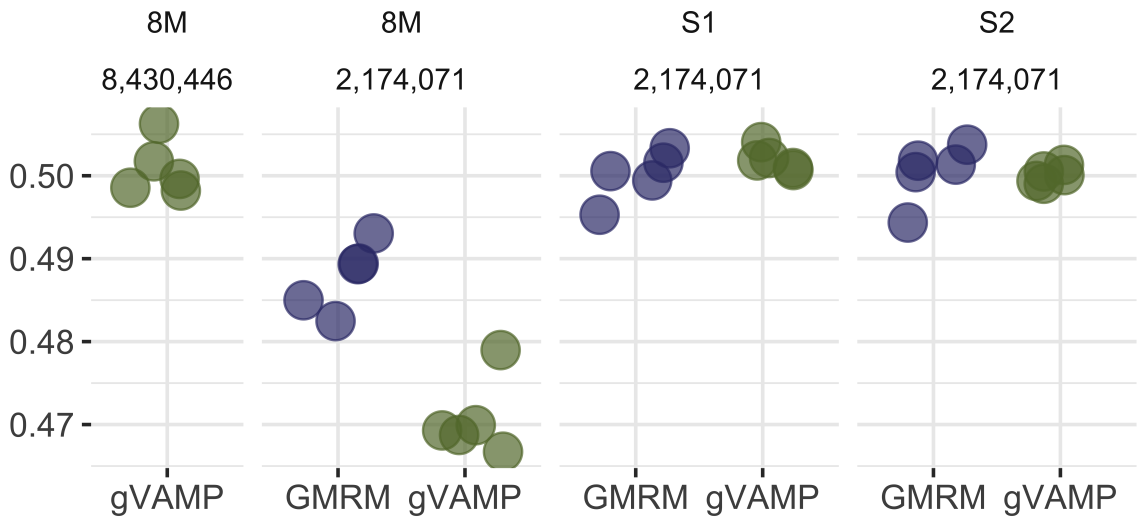

Figure S7. SNP heritability estimation of GMRM versus gVAMP with different numbers of SNP markers in the simulation. Comparison of the proportion of phenotypic variation attributable to either 8,430,446 or a subset of 2,174,071 autosomal single nucleotide polymorphism (SNP) genetic markers (SNP heritability) estimated by GMRM and gVAMP. We consider three simulation scenarios: "8M" represents the scenario of 40,000 causal SNP markers randomly selected from 8,430,446 total SNPs with effects sampled from a Gaussian distribution and total SNP heritability of 0.5; "S1" represents the scenario of 40,000 causal SNPs randomly selected from 2,174,071 total SNPs with effects sampled from a Gaussian distribution and total SNP heritability of 0.5; and finally "S2" represents the scenario of 40,000 causal SNPs randomly selected from 2,174,071 total SNPs with effects sampled from a mixture of Gaussians and total SNP heritability of 0.5. Points give the posterior means for GMRM and the convergence of gVAMP from five simulation replicates. Analysing 8,430,446 SNPs with gVAMP increases the heritability estimate over GMRM. This is consistent with an increase in phenotypic variance captured by the full imputed sequence data, as opposed to analyzing a selected subset of SNP markers, in which case gVAMP estimates are lower than those obtained from GMRM. Given the same data containing all the causal variants, the algorithms perform similarly irrespective of the underlying effect size distributions ("S1" and "S2").

the simulations selecting 40,000 causal variants from 2,174,071 markers, gVAMP and GMRM give identical inference under both Gaussian and a mixture of Gaussian effect size distributions (Figure S7).

gVAMP provides an alternative approach to association testing where the effects of each marker can be estimated conditional on all other genetic variants genome-wide (see "gVAMP SE association testing" in the Methods). The expectation is that SE association testing should yield broadly similar results to posterior inclusion probability testing from Bayesian fine-mapping approaches. Fine-mapping approaches have been developed to overcome the issue that individual-level Bayesian methods cannot be applied to full sequence data and have similar priors to GMRM, thus we restrict our comparison to this method. GMRM has previously been shown to outperform other Bayesian fine-mapping approaches [39].

Comparing gVAMP to GMRM at 2,174,071 SNP markers, as GMRM cannot analyse more than this within reasonable time frames, we find that, for significance thresholds of $p \leq 0.005$, the FDR is controlled at $\leq 5\%$, with greater power than GMRM posterior inclusion probabilities (Figure S8). For 8,430,446 imputed SNP markers, stronger linkage disequilibrium limits the assignment of

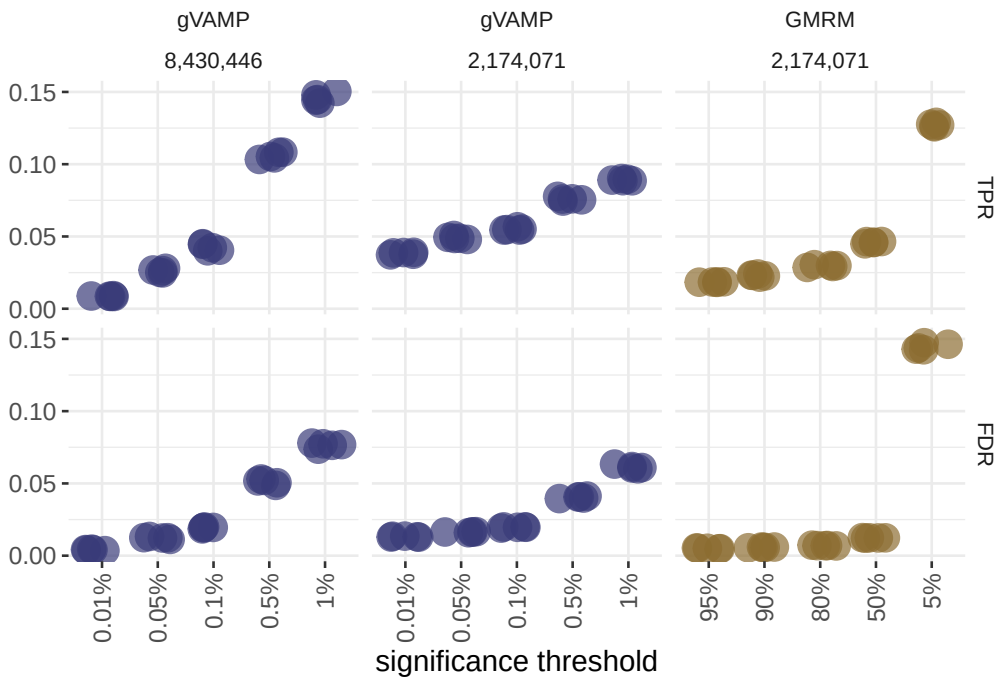

Figure S8. Whole genome fine-mapping of gVAMP in a simulation study using UK Biobank genotype data. True positive rate (TPR) and false discovery rate (FDR) of SE association testing at 2,174,071 and 8,430,446 markers for different significance thresholds. We then compare this to the TPR and FDR of genome-wide fine-mapping using the posterior inclusion probability of each SNP generated by GMRM. For significance thresholds of $p \leq 0.005$, the FDR is controlled at $\leq 5\%$, with greater power than GMRM posterior inclusion probabilities.

significance to the single-marker resolution, reducing the TPR, but FDR improves as effects are resolved to the correct single-marker level when all causal variants are within the data (Figure S8), supporting our main results. Thus, our algorithm facilitates individual-level (and summary-level) Bayesian methods to be applied to all variants jointly at scale, so that genetic variant effects can be localised to single-locus resolution conditional on all other genetic variants within a cohort.

Supplementary Note 2

Onsager correction calculation

In order to ensure Gaussianity of residuals, gVAMP calculates the so-called Onsager correction based on (5). For such calculation, the derivative of the denoising function f_t defined in (3) is required. Let us denote the numerator and denominator of (4) with $\mathrm{Num}(r_1)$ and $\mathrm{Den}(r_1)$, respectively. Then,

$$\frac{\partial \mathrm{Num}(r_1)}{\partial r_1} = \lambda_t \cdot \sum_{l=1}^{L} \pi_{t,l} \cdot \frac{\sigma_{t,l}^2}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)^{3/2}} \cdot \mathrm{EXP}(\sigma_{t,l}^2) \cdot \left[1 - r_1^2 \cdot \frac{(\sigma_{t,*}^2 - \sigma_{t,l}^2)}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)(\gamma_{1,t}^{-1} + \sigma_{t,*}^2)}\right],$$

$$\frac{\partial \mathrm{Den}(r_1)}{\partial r_1} = -r_1 \cdot \left[\lambda_t \cdot \sum_{l=1}^{L} \frac{\pi_{t,l}}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)} \cdot \frac{\sigma_{t,*}^2 - \sigma_{t,l}^2}{(\gamma_{1,t}^{-1} + \sigma_{t,l}^2)(\gamma_{1,t}^{-1} + \sigma_{t,*}^2)} \cdot \mathrm{EXP}(\sigma_{t,l}^2) + (1 - \lambda_t) \cdot \frac{\gamma_{1,t}^{3/2} \cdot \sigma_{t,*}^2}{(\gamma_{1,t}^{-1} + \sigma_{t,*}^2)} \cdot \mathrm{EXP}(0)\right].$$

Thus, the Onsager correction reads

$$\frac{\partial}{\partial r_1}\left(\frac{\mathrm{Num}(r_1)}{\mathrm{Den}(r_1)}\right) = \frac{\frac{\partial}{\partial r_1}\mathrm{Num}(r_1)}{\mathrm{Den}(r_1)} - \frac{\mathrm{Num}(r_1) \cdot \frac{\partial}{\partial r_1}\mathrm{Den}(r_1)}{(\mathrm{Den}(r_1))^2}.$$

Conjugate gradient algorithm for solving linear systems

Algorithm 2 Conjugate gradient method for solving a symmetric linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Input: Initial estimate of the solution $\boldsymbol{x}_0$, initial residual $\boldsymbol{r}_0 = \boldsymbol{b}$, initial search direction $\boldsymbol{p}_0 = \boldsymbol{r}_0$, linear system matrix $\boldsymbol{A}$, right-hand side vector $\boldsymbol{b}$, stopping error threshold $\varepsilon > 0$.
for $n = 1, 2, 3, \ldots$ **do**
  $\alpha_n = \frac{\boldsymbol{r}_{n-1}^T \boldsymbol{r}_{n-1}}{\boldsymbol{p}_{n-1}^T \boldsymbol{A} \boldsymbol{p}_{n-1}}$
  $\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + \alpha_n \boldsymbol{p}_{n-1}$
  $\boldsymbol{r}_n = \boldsymbol{r}_{n-1} - \alpha_n \boldsymbol{A} \boldsymbol{p}_{n-1}$
  $\beta_n = \frac{\boldsymbol{r}_n^T \boldsymbol{r}_n}{\boldsymbol{r}_{n-1}^T \boldsymbol{r}_{n-1}}$
  $\boldsymbol{p}_n = \boldsymbol{r}_n + \beta_n \boldsymbol{p}_{n-1}$
  If $n \geq 1$ and $||\boldsymbol{x}_n - \boldsymbol{x}_{n-1}||_2 / ||\boldsymbol{x}_n||_2 < \varepsilon$, then **break**
end for
return $\boldsymbol{x}_n$