# Light-speed whole genome association testing and prediction via Approximate Message Passing

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

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

*corresponding author
† indicates joint supervision

## Abstract

Efficient utilization of large-scale biobank data is crucial for inferring the genetic basis of disease and predicting health outcomes from the DNA. Yet we lack accurate statistical models to estimate the effect of each locus, conditional on all other genetic variants, controlling for both local and long-range linkage disequilibrium. Additionally, we lack algorithms which scale to data where health records are linked to whole genome sequence information. To address these issues, we develop a new algorithmic paradigm, based on Approximate Message Passing (AMP), specifically tailored for both genomic prediction and association testing. Our gVAMP (genomic Vector AMP) approach requires less than a day to jointly estimate the effects of 8.4 million imputed genetic variants in over 400,000 UK Biobank participants, and it provides an association testing framework capable of directly fine-mapping each genetic variant, or gene burden score, conditional on all other measured DNA variation genome-wide. Across 13 traits, we find 8,222 genome-wide significant autosomal associations that are localised to the single-locus level, conditional on all other imputed loci. Extending the model to jointly estimate the effects of rare variant gene burden scores from sequencing data and imputed X chromosome variants, conditional on all other genes and all 8.8 million variants, we find 60 genes where rare coding mutations significantly influence phenotype, and 76 associations localised to the single-locus level on chromosome X, for five traits. In comparison to existing state-of-the-art methods, both in simulations and in application to the UK Biobank, gVAMP yields out-of-sample prediction accuracy comparable to individual-level Bayesian methods, outperforms summary statistic Bayesian methods, and outperforms REGENIE for standard association testing, in a fraction of the compute time. This truly large-scale development of the AMP framework establishes the foundations for a far wider range of statistical analyses for hundreds of millions of variables measured on millions of people.

Keywords: approximate message passing; genome-wide association testing; genomic prediction; polygenic risk scores

# Introduction

Association testing [1–4], fine-mapping [5] and polygenic risk score methods [6, 7] are all based on forms of adaptive penalized regression models, with parameters estimated by restricted maximum likelihood (REML), Markov Chain Monte Carlo (MCMC), expectation maximisation (EM), or variational inference (VI). Characteristically, REML and MCMC are computationally intensive and slow; EM and VI are faster, but trade speed for accuracy, with few theoretical guarantees. Current software implementations of these algorithms limit either the number of markers or individuals that can be analysed, and focus on only one specific type of analysis, despite improved estimation and prediction going hand-in-hand [8]. As a result, mixed linear model association (MLMA) approaches are restricted to using less than one million single nucleotide polymorphisms (SNPs) to control for the polygenic background when conducting association and burden testing [1, 2]. This results in a loss of testing power, as compared to a model fitting all markers jointly [7], and may be inadequate to control for fine-scale confounding factors, especially in whole genome sequence (WGS) data [9]. Polygenic risk score algorithms are limited to a few million SNPs, or lose power by modelling only blocks of genetic markers [10, 11], meaning that the limits to prediction accuracy with increasing marker density remain unexplored. Likewise, fine-mapping methods are generally limited to focal segments of the DNA [5], being incapable of fitting all genome-wide SNPs together. With increasing availability of whole-genome, whole-exome, and imputed sequence data for hundreds of thousands of individuals, it would be optimal to estimate variant effects conditional on the full polygenic background by fitting all DNA variants jointly, facilitating greater biological insight.

In order to address these issues, we propose a new framework for inference in genome-wide association studies (GWAS), based on Approximate Message Passing (AMP). AMP refers to a family of iterative algorithms that has been applied to a range of statistical estimation problems, including linear regression [12–14], generalised linear models [15–18], and low-rank matrix estimation [19–21]. The popularity of AMP stems from its unique features: *(i)* AMP can be tailored to take advantage of structural information available about the signal (i.e., a wide range of Bayesian priors can be used); *(ii)* under mild model assumptions, the AMP performance in the high-dimensional limit is precisely characterized by a succinct deterministic recursion called state evolution [13]; and *(iii)* using state evolution, it has been proved that AMP achieves Bayes-optimal performance in a number of settings [20, 22]. These features hold in a regime in which the number of covariates to be estimated and the number of observations are both large, and in theory, this gives AMP the *potential* to jointly estimate millions of DNA variant effects for millions of people controlling for linkage disequilibrium (LD), thus reaching the fundamental inference limits of genotype-phenotype data.

Here, we turn this potential into a practical, light-speed method, which we call genomic Vector AMP (gVAMP), an AMP algorithm tailored to genomic data. Previously proposed AMP algorithms (e.g., for wireless communications [23], compressive imaging [24], MRI [25], or PCA in genetics [26]) cannot be transferred to biobank analyses as: *(i)* they are entirely infeasible at scale, requiring expensive singular value decompositions; and *(ii)* they simply fail when applied to genome-wide DNA measures, due to the structure of the data being analyzed. Thus, we first outline the extensive algorithmic improvements that we made. We then use the UK Biobank data to show that gVAMP gives DNA variant effect sizes that can be used to create polygenic risk scores with comparable

prediction accuracy as the closest related model implemented in MCMC, but with dramatically improved speed and the scalability to utilise full sequence data. We demonstrate that, if gVAMP were used to conduct MLMA testing, then power, false discovery rate, and speed are greatly improved as compared to REGENIE [1]. Finally, we show how gVAMP provides an alternative to MLMA where the effects of each marker can be estimated conditional on all other genetic variants genome-wide, allowing effects to be localised to the single-locus level, and providing joint testing of rare sequence variant burden scores against a full genetic background of millions of DNA variants.

# Results

## Overview of gVAMP

In a study involving $N$ individuals and $P$ genetic variants, we assume data are stored in a genotype matrix the entries of which are encoded to a set $\{0, 1, 2\}$ depending on homozygosity and allele frequencies of each locus (column). Its column normalized version $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ such that each column mean is zero and variance is one, we call the normalized genotype matrix. We assume a general form of Bayesian linear regression, common to GWAS [7, 11], estimating the effects vector $\boldsymbol{\beta} \in \mathbb{R}^P$ from a vector of phenotype measurements $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ that satisfies the following linear relationship:

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

Here, $\boldsymbol{x}_i$ is a row vector 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 Euclidean inner product between normalized genotype associated to the $i$-th individual and the vector of effects, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is an unknown noise vector that is assumed to follow the multivariate normal distribution $\mathcal{N}(0, \gamma_\epsilon^{-1} \cdot \boldsymbol{I})$ with unknown noise precision parameter $\gamma_\epsilon^{-1}$. In words, the error terms among different individuals are assumed to be independent and identically distributed Gaussians with precision $\gamma_\epsilon^{-1}$.

To allow for a range of genetic effects, we select the prior on the signal $\boldsymbol{\beta}$ to be of an adaptive spike-and-slab form, and learn its parameters from the data. However, we note that a wide class of priors is possible within our framework. Specifically, for every $i = 1, \ldots, P$, we model $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ as

$$\beta_i \sim (1 - \lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{i=1}^{L} \pi_i \cdot \mathcal{N}(\cdot, 0, \sigma_i^2), \tag{2}$$

where $\lambda \in [0, 1]$ is the DNA variant inclusion rate; $L$ is a positive integer, $(\pi_i)_{i=1}^L$ and $(\sigma_i^2)_{i=1}^L$ denote, respectively, the number of Gaussian mixtures, the mixture probabilities and the variances for the slab component. We learn all of these parameters from the data in an adaptive Bayes expectation-maximisation (EM) framework. This avoids expensive cross-validation and yields biologically informative inference of the phenotypic variance attributable to the genomic data (the SNP heritability, $h^2_{SNP}$).

We find that the application of existing expectation-maximisation vector AMP (EM-VAMP) algorithms [27–29] to either simulated genomic data or the UK Biobank data set are entirely ineffective, leading to diverging estimates of the signal. We therefore combine a number of principled approaches to create gVAMP, an algorithm tailored to the analysis of genomic data. Algorithm 1

**Algorithm 1** gVAMP

**Input:** preprocessed normalized genotype matrix $\boldsymbol{X} \in \mathbb{R}^{N \times P}$, max number of iterations $N_{\text{it}}$, 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 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 $\Theta$, called $\Theta_t$
      **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 $\gamma_\epsilon$, 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** $\hat{\boldsymbol{\beta}}_{1,t}$

provides a succinct description of our proposed approach with full details given in the Methods. gVAMP is a form of EM-VAMP, first introduced in [27–29], which 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. As we show below, in both simulated settings and in applications to the UK Biobank data, gVAMP optimally infers genetic marker effects that can be directly used in polygenic

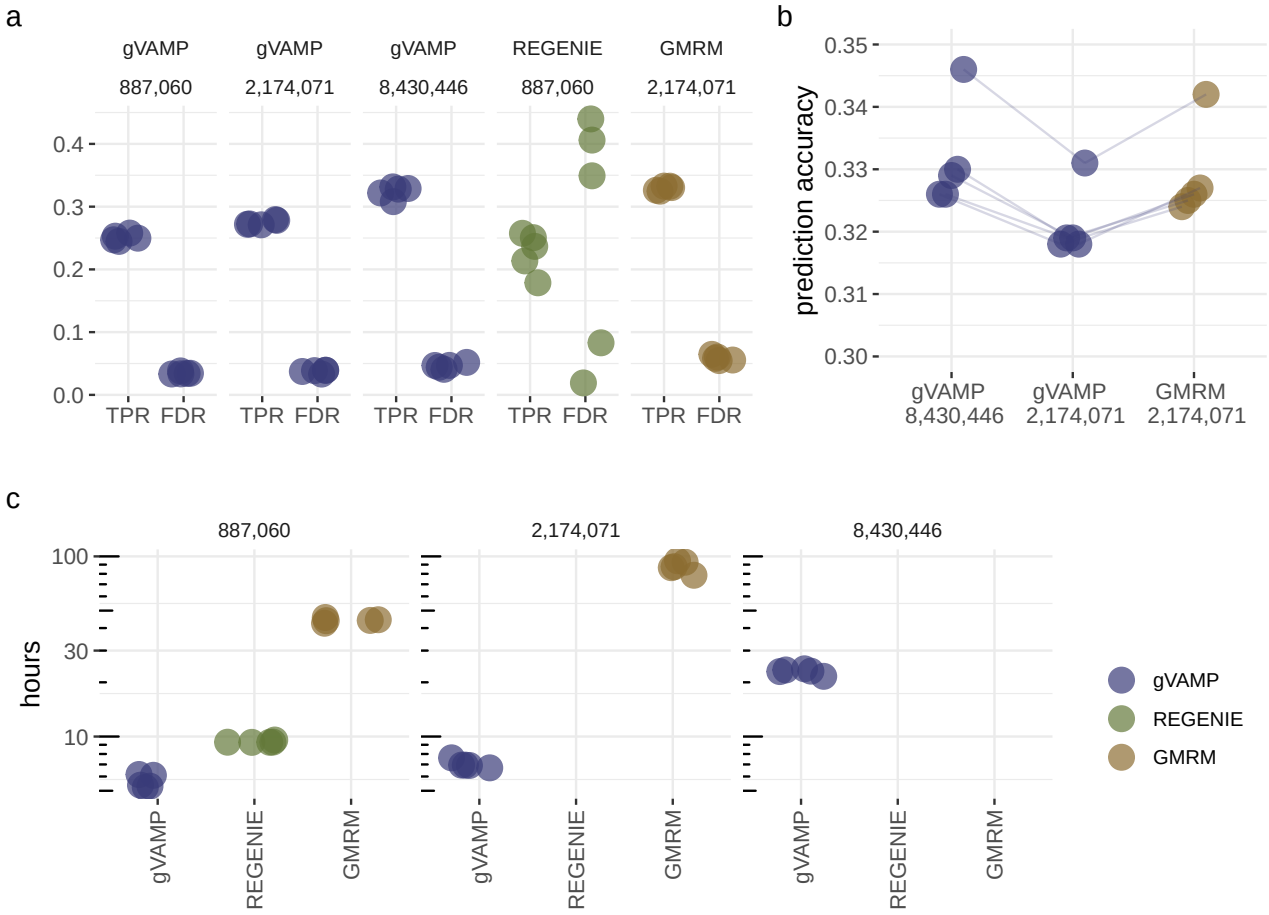


**Figure 1. Simulation study of association testing power and run time using UK Biobank genotype data**. We begin with 8,430,446 SNP markers, we randomly select 40,000 as causal and use these to simulate a phenotype (see Methods). Standard leave-one-chromosome-out (LOCO) association testing approaches 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 is half the speed 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.

risk score prediction and, at the same time, it estimates the significance of each locus, conditional on all other loci genome-wide (joint association testing controlling for both local and long-range linkage disequilibrium).

## gVAMP association testing power and prediction accuracy in simulations

There are three key features of our gVAMP algorithm. The *first* is the so called Onsager correction: this is added to ensure the asymptotic normality of the noise corrupting the estimates of $\boldsymbol{\beta}$ given

by the algorithm at every iteration. Here, in contrast to MCMC or other iterative algorithms 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 denoisers (for more details, see Methods). By taking the conditional expectation with respect to the posterior marginal density, given the estimates on the signal and the noise precision, Bayes-optimal performance can be obtained in some settings [20, 22]. Specifically, [27] provides evidence that VAMP achieves the information-theoretically optimal reconstruction error, computed via the replica method from statistical physics [30, 31].

This first feature has empirical importance for genomic studies, as it implies that our model should perform similarly to individual-level Bayesian analyses, unlike other variational inference approaches which trade accuracy for speed. Recent work has shown that Bayesian linear regression models provide highly accurate polygenic risk scores, maximise power in marginal association testing, and allow fine-mapping of marginal associations to obtain joint estimates [7]. Thus, we begin by comparing our model to state-of-the-art approaches in a simulation study using the UK Biobank data.

We compare gVAMP to widely used REGENIE [1] and GMRM [7] for association testing. Leave-one-chromosome-out (LOCO) association testing approaches have become the field standard and are two-stage, with a subset of markers selected for the first stage to create genetic predictors, and then statistical testing is conducted in the second stage for all markers one-at-a-time. From 8,430,446 imputed SNP markers, we randomly select 40,000 as causal and use these to simulate a phenotype. For the first stage of LOCO, we select either all 8,430,446 markers, or LD pruned sets of 2,174,071 and 887,060 markers. REGENIE is given 887,060 markers (it is recommended to use less than 1 million), GMRM up to 2,174,071 markers (as this completes within reasonable compute time and resource use), whilst gVAMP can utilise the full range. For the second stage, all methods then use all markers for the one-by-one LOCO testing.

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 1a). Both approaches outperform the commonly used REGENIE software in both TPR and FDR, which does not always control the FDR below 5% (Figure 1a). 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 S1). 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 S1), and *(ii)* 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 1a and S1).

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 [8]. 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

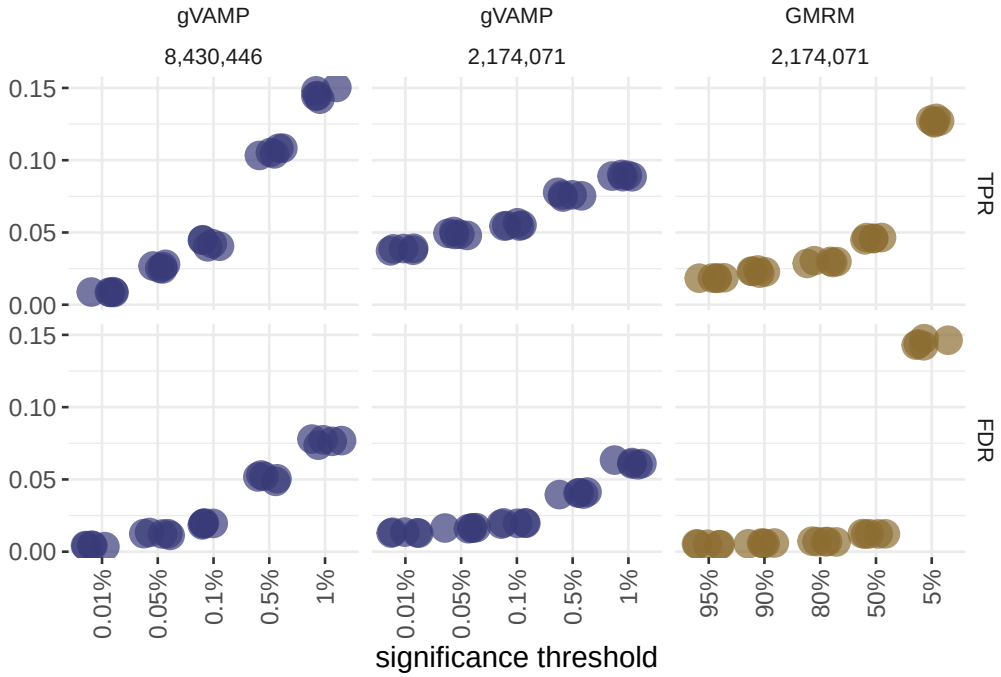


**Figure 2. Whole genome fine-mapping of gVAMP in a simulation study using UK Biobank genotype data**. Leave-one-chromosome-out testing detects regions of the DNA associated with outcomes, whereas fine-mapping aims to localise marker effects to a single SNP conditional on all other markers. AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers, removing the need for follow-up fine-mapping. We call this approach SE $p$-value testing, and we calculate the true positive rate (TPR) and false discovery rate (FDR) of this approach 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 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.

imputed SNP markers, gVAMP improves over GMRM (Figure 1b). We note that analysing all 8,430,446 SNPs is computationally infeasible for GMRM (see also the paragraph below on the speed of the algorithms).

In turn, polygenic risk score prediction accuracy depends upon the $h^2_{SNP}$, the number of true underlying causal variants and the sample size [32], 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 S2). gVAMP gives correct estimates when given the full 8,430,446 markers (Figure S2) and when we repeat 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 in these settings where the data contain all causal variants (Figure S2).

The *second key feature* of gVAMP is its computational efficiency, which makes it run in light-speed, allowing for joint processing of the full set of 8,430,446 markers. As compared to REGENIE, gVAMP completes in a fraction of the time given the same data and compute resources (Figure 1c). It is dramatically faster than the Gibbs sampling algorithm GMRM (Figure 1c). Even with 8,430,446 imputed SNP markers, the model yields estimates under a day (Figure 1c).

The *third key feature* of gVAMP is that it provides an alternative approach to association testing where the effects of each marker can be estimated conditional on all other genetic variants genome-wide. Our method relies on the properties of the gVAMP estimator, whose noise is asymptotically Gaussian due to the Onsager correction [29]. Namely, $\boldsymbol{r}_{1,t} \approx \boldsymbol{\beta} + \mathcal{N}(0, \gamma_{1,t}^{-1}\boldsymbol{I})$, where $\boldsymbol{\beta}$ is the ground-truth value of the underlying genetic effects vector. Hence, given a marker index $i$, a one-sided $p$-value for the hypothesis test H0 : $\beta_i = 0$ can be calculated as $\Phi(-|(\boldsymbol{r}_{1,t})_i| \cdot \gamma_{1,t}^{1/2})$, where $\Phi$ 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 refer to this association testing as *state evolution p-value testing* (SE association testing), and the expectation is that it 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. If they could, then genetic effects could be localised to single-locus resolution conditional on all other genetic variants within a cohort.

We compare gVAMP to GMRM, which has previously been shown to outperform other Bayesian fine-mapping approaches [33]. 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 2). For 8,430,446 imputed SNP markers, stronger linkage disequilibrium limits the assignment of 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 2).

Thus, taken together, we show 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, in a fraction of the compute time as compared to state-of-the-art alternatives. This makes gVAMP the optimal variational inference algorithmic choice for analyses of biobank data. In fact, in the next section, we demonstrate that gVAMP outperforms a wide variety of different approaches in the analysis of 13 UK Biobank traits and can scale to accommodate a wide range of input data including whole genome sequence information.

## Application of gVAMP to UK Biobank data

We first compare the polygenic risk score prediction accuracy of gVAMP to the widely used summary statistic methods LDPred2 [10] and SBayesR [11], and to the individual-level method GMRM [7] (Table 1, Figure 3a). Training data sample size and trait codes are given in Table S1 for each trait. gVAMP outperforms all methods for most phenotypes, often obtaining the highest out-of-sample prediction accuracy yet reported to date. Summary-statistic approaches based on marginal effects obtain from the second step of REGENIE have lower performance, given the same data, as compared to individual-level models (Table 1, Figure 3a). Generally, gVAMP performs similarly to GMRM, and it is able to improve over it when analysing the full set of 8,430,446 SNPs. This is expected as the models have similar priors, our algorithm is expected to achieve near-Bayes optimal performance, and the set of 2,174,071 SNP markers is a weakly LD pruned subset of the 8,430,446 set. 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

**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** |

markers. This is consistent with our simulation study, and it implies that more of the phenotypic variance is captured by the SNPs when the full imputed SNP data are used (Figure S3).

We then compare gVAMP to REGENIE and to GMRM for association testing of the 13 traits within a MLMA framework. gVAMP performs similarly in leave-one-chromosome-out (LOCO) testing conducted using a predictor from GMRM, with the use of the full 8,430,446 imputed SNP markers generally improving performance. REGENIE yields far fewer associations than either GMRM or gVAMP for all traits, consistent with our simulation results. Furthermore, gVAMP also provides leave-one-out association testing, where the effect of each SNP marker is assessed conditional on a genetic predictor created from all other SNPs. These results are reported in Table 2 and Figure 3b.

We highlight that gVAMP is the first algorithm capable of joint testing of the full imputed SNP data at the UK Biobank scale. 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 3c). For all 13 traits, we find that the SE association estimates we obtain converge in number and location after iteration 20 (Figure S4).

Finally, to provide a demonstration of the wide-scale applicability of our approach, we combine 8,430,446 autosomal imputed SNP markers with 394,823 X chromosome SNPs and 17,852 whole exome sequencing (WES) gene burden scores, and re-analyse five phenotypes, estimating the effects jointly within the gVAMP SE testing framework. As compared to our previous analysis reported above that used only 8,430,446 autosomal imputed SNP markers, we find that the genome-wide significant associations for each trait become split, being allocated mostly to the imputed data, but

**Table 2. Genome-wide significant associations for 13 UK Biobank traits from GMRM, gVAMP and REGENIE at 8,430,446 genetic variants**. Training data sample size and trait codes are given in Table S1 for each trait. 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 (see Methods). 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 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, 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 |

partially to the WES gene burden scores and X-chromosome imputed SNPs (Figure 4). The full gene list of WES and X chromosome findings for all traits is given in Supplementary Data Table 1.

We find WES burden scores calculated for the genes CALCR, CEP350, HSPA9, MOXD1 and SLC26A8 are significantly associated with mean corpuscular haemoglobin (MCH). All genes replicate, with nearby SNPs being associated with MCH in the Open Targets catalogue (which contains estimates from the GWAS catalogue, FinnGen and UK Biobank, `genetics.opentargets.org`) at genome-wide significance, are enriched in the regulation of erythrocyte differentiation (GO:0045647, EnrichR [34] p-value 0.00124). For heel bone mineral density (BMD), we find genome-wide significant associations for EFNA3, GRK5, and SCG2. These do not replicate in the Open Targets catalogue; however, EFNA3 was previously shown to play a role in the regulation of angiogenesis and VEGF signaling and implicated in the regulation of postmenopausal osteoporosis [35], GRK5 is a G-protein-coupled receptor gene linked to bone formation [36], and SCG2 is an early familial GWAS association for bone mineral density [37]. For red blood cell count (RBC), we find associations for COL4A4 and TFRC, for which associated SNPs were only recently discovered in large-scale meta-analysis [38]. For height (HT), 45 of the 50 WES genes findings are for genes that have been previously linked with human height in Open Targets. Of those not listed are SHOX (a reported cause of short stature in individuals with isolated or familial short stature [39]), TRIM68 (coactivator of androgen receptor), TRAPPC2 (variants associated with rare hereditary childhood short stature [40]), APOB (apolipoprotein B protein), and AEBP1 (a regulator of collagen fibrillogenesis). Using EnrichR [34]

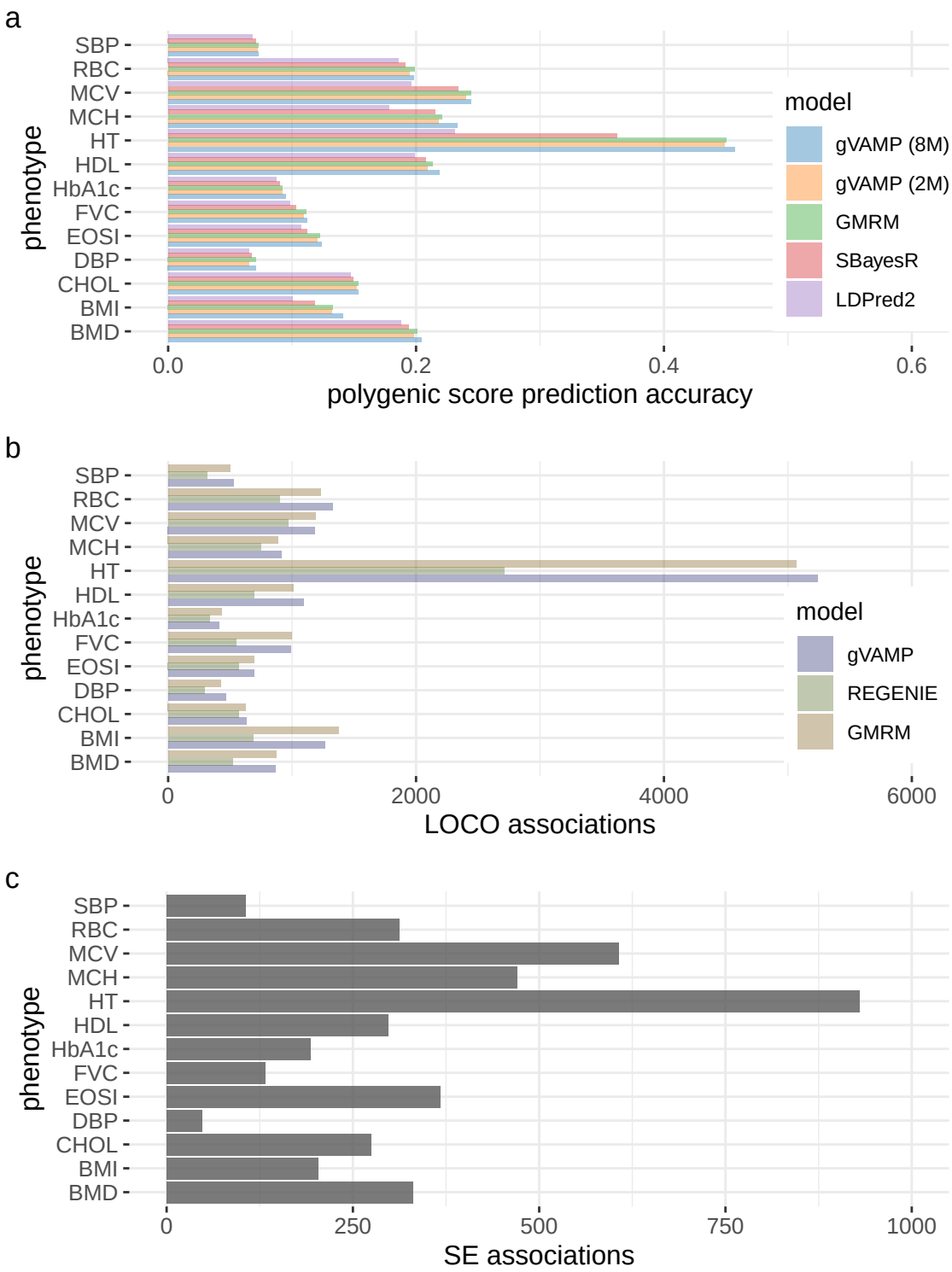


**Figure 3. gVAMP polygenic risk score accuracy and association testing in the UK Biobank.** (a) Prediction accuracy of gVAMP in a hold-out set of the UK Biobank across 13 traits as compared to other approaches. (b) Leave-one-chromosome-out (LOCO) testing of gVAMP across 13 UK Biobank traits as compared to other approaches at 8,430,446 markers. (c) AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers. We call this approach SE $p$-value testing, and 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.

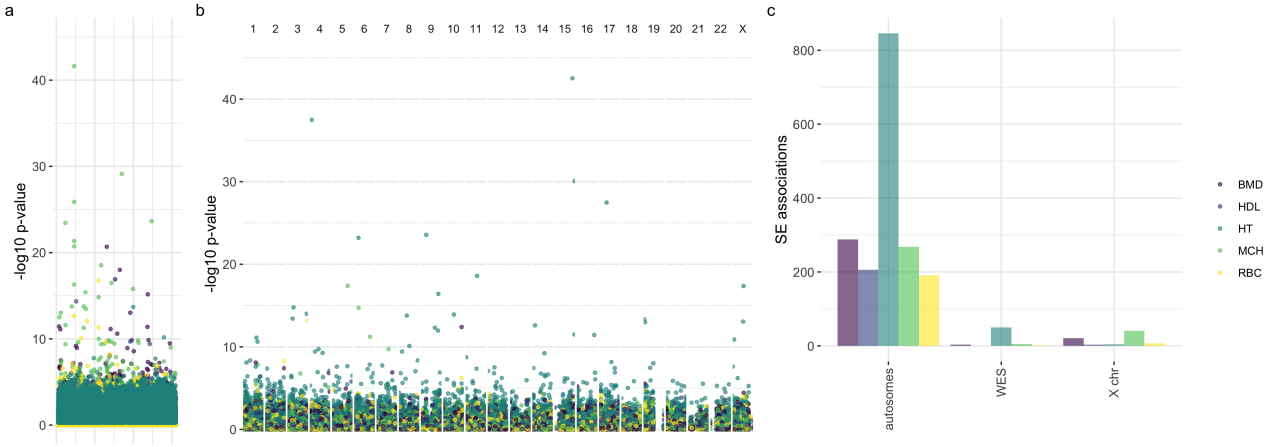


**Figure 4. gVAMP association testing combining imputed genetic markers and whole exome sequencing in the UK Biobank**. AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers. We combine 8,430,446 autosomal imputed SNP markers with 394,823 X chromosome SNPs and 17,852 whole exome sequencing gene burden scores, estimating the effects jointly within the gVAMP SE testing framework. We show the $-\log 10$ of the $p$-value for (a) X chromosome SNPs, and (b) the 17,852 whole exome sequencing gene burden scores. (c) gives the number of genome-wide associations at a $p$-value threshold of less than $5 \cdot 10^{-8}$.

we find that genes GHR, ACAN, GH1, SERPINH1, IGF1R and DDR2 are linked to endochondral ossification, with and without skeletal dysplasias (enrichment $p$-value $1.09 \cdot 10^{-8}$), COL27A1, COL24A1, SERPINH1 and ASPN are enriched in cartilage formation and assembly in both GO and Reactome databases ($6.31 \cdot 10^{-8}$) and SCUBE3, ACAN, COL27A1, COL24A1, SERPINH1, ASPN, LOXL2, and DDR2 are enriched in extracellular matrix organization in both GO and Reactome databases ($1.20 \cdot 10^{-8}$). Thus, alongside some novel findings, our results predominantly replicate across other studies, implicating the same genes. Here however, we are able to show that previous associations can be attributed to rare coding variants within these genes when effects are estimated conditional on all 8.8M imputed autosomal and X chromosome sequence variants.

Additionally of note, we find 21, 3, 41, 7, and 4 X chromosome associations that are conditional on all other X chromosome, autosomal, and WES burden scores for BMD, HDL, MCH, RBC, and HT respectively. The X chromosome is chronically understudied and thus replication in Open Targets is less likely. However, all HT associations are previously reported in Open Targets; of the 7 findings for RBC, 5 are novel and 2 (rs5913773 and rs5960376) replicate in a previous study [41]; and all HDL associations are novel and are located near the MAGEB gene family (supporting limited evidence in mouse [42]), DIAPH2 (a formin gene) and rs1339116 which is intragenic. All 21 findings for BMD are novel, except rs10126777 located near gene TBX22 which is linked to bone formation and cleft palate [43]. Likewise, all 41 findings for MCH are novel, with exceptions including rs189438396 upstream of PTCHD1 (p-value=0.0019, [41]). Taken together, these results suggest wide-spread X chromosome effects on haemoglobin levels and bone mineral density and demonstrate the utility of our approach for joint inference and association testing of genetic effects across data of any size.

## Discussion

We show that gVAMP generates genetic predictors and association test statistics in a single step, without additional computations, with accuracy similar to individual-level MCMC methods, in a fraction of the compute time as compared to state-of-the-art alternatives. gVAMP outperforms a wide variety of different approaches in the analysis of 13 UK Biobank traits when it is used for both polygenic risk scoring and MLMA association testing. Importantly, we provide an 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. 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 [44], and previous work has shown that better modelling within a single large biobank can facilitate improved association testing in other global biobanks [45]. Here, we have demonstrated that gVAMP provides an optimal choice for high-dimensional linear regression within genomic data and thus our approach can be used in the same way. However, 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.

While AMP algorithms have been proposed for generalised linear models [15–18,27] and distributed learning [46, 47], developing models for these tasks using genomic data requires a lot of further algorithm development and benchmarking. However, given the speed and accuracy of our approach, a wider range of statistical models may now be feasible for both large-scale sequence and multi-omics data. While we show the importance of fitting all markers jointly for association testing, there are differences across traits in out-of-sample prediction accuracy, where model choice matters more for some traits (HT, BMI) than others (CHOL, HbA1c). We may be reaching the limits of prediction of phenotype from the DNA from adaptive penalized regression approaches for some outcomes, and it remains to be seen whether additional gains can be made by utilising full whole genome sequence information.

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. This makes it crucial to develop statistical approaches that are capable of fitting all variants jointly, asking whether given the LD structure of the data, there is evidence for an effect at each locus, conditional on all others. The approach we present here has extensive

theoretical support and we demonstrate that it is an optimal variational inference framework for genomic data that is capable of scaling to accommodate a wide range of future input data, including whole genome sequence information.

# Methods

## gVAMP algorithm

Approximate message passing (AMP) was originally proposed for linear regression [12–14] assuming a Gaussian design matrix $\boldsymbol{X}$. To accommodate a wider class of structured design matrices, vector approximate message passing (VAMP) was introduced in [27]. 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. 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 $\alpha_{1,t}$ (resp. $\alpha_{2,t}$) is the state evolution parameter tracking 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 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. Namely,

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

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

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, conducting 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 symmetric and positive-definite matrix via the *conjugate gradient method* (CG), see Algorithm 2 in Supplementary Note, which is included for completeness. If $\kappa$ 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 [48], 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 eventual 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 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 [49] 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 $\gamma_\epsilon$ one has to calculate a 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 data set 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 [27], 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) [50], and variance auto-tuning [29]. 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 [29]. 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 $\Theta$ 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 $\Theta$.

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 [27] 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 [28], and its rigorous theoretical analysis presented in [29]. In this subsection, we summarize the hyperparameter estimation results derived based upon [51] in the context of our model. We find that the final update formulas for our hyperparameter estimates are:

- Sparsity rate $\lambda$: We define $\{\zeta_j\}_{j=1}^P$ as follows: $\forall j = 1, \ldots, P$,

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

  The intuition behind $(\zeta_j)_{j=1}^P$ is that each $\zeta_j$ 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 $\lambda_{t+1}$ reads as

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

- Probabilities of mixtures 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 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 $\gamma_\epsilon$ : 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 $\gamma_\epsilon$ 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 normalized genotype matrix are approximately equally spit 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 $\{0, 1, 2\}$:

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

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 in 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 comparions between different methods fair, the results presented in the rest of 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

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. 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 $\leq 0.9$ 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 minor allele frequency (MAF) $> 0.0002$ and rs identifier, in the set of European-ancestry participants, providing a data set of 9,144,511 SNPs. From this, we took the overlap with the Estonian Genome Centre data as described in [7] to give a final set of 8,430,446 autosomal markers and 444,055 individuals that were related at less than first degree relatives.

We then combine this data with X-chromosome data and the UK Biobank whole exome sequence data. For the X chromosome data, we follow the same QC steps outlined above obtaining 394,823 SNPs for 444,055 individuals. The UK Biobank final release data set 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.

Finally, we link these SNP data to the measurements, tests, and electronic health record data available in the UK Biobank data [52], and we select 7 blood based biomarkers and 6 quantitative measures which show $\geq 15\%$ SNP-heritability and $\geq 5\%$ out-of-sample prediction accuracy [7]. Again, our focus is on selecting a group of phenotypes for which there is sufficient power to observe differences among approaches.

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.

Additionally, 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 data set 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 first demonstrating and benchmarking in the simpler linear model setting.

## Statistical analysis in the UK Biobank

### gVAMP model parameters

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 $\rho$ 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 data sets, and the algorithm is rather stable for a range of initialization choices. Similarly, the algorithm is stable for different choices of the damping $\rho$, 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 Figure 1.

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 data sets, 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 $\rho$, 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 1/2. 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 data set, 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 X chromosome and 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 $\mu_{X_j}$ 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 minus 1 divided by $\gamma_w$, the estimate of the noise precision.

We compare gVAMP to a Gibbs sampler approach (GMRM) with a similar prior (the same number of starting mixtures) as presented in [7]. We select this comparison as the Gibbs sampler was demonstrated to exhibit the highest genomic prediction accuracy up to date [7]. 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 $\chi^2$ distribution using the sum of the squared regression coefficient estimates.

We also compare gVAMP to the summary statistics prediction methods LDPred2 [10] and SBayesR [11] run on the 2,174,071 SNP data set. 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 [53] 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 data set. Following that, using the learned parameters, we make a linear combination of the per-group predictors in the test data set 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. 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 [7], 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, as it is recommended to use less than 1 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 $\leq 0.01$) 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, as 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 [27], 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 $\Phi$ 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 one-by-one association testing on the full 8,430,446 SNP markers for the empirical UK Biobank analysis of 13 traits, using the estimates of $\boldsymbol{r}_{1,t}$ obtained at iteration $t = 25$, in accordance with the stopping criterion discussed above. We remark that the testing results are 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.

## Simulation study

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 $0.5/40,000$, where 0.5 is the proportion of variance attributable to the SNP markers (the SNP heritability). Multiplying the simulated SNP effects by normalized values of the 40,000 causal markers, gives a vector of genetic values of length $N = 414,055$ 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 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 above for the empirical UK Biobank study. We also analyse the data with GMRM using 2,174,071 or 887,060 SNP markers, 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 $\leq 0.01$) identified by each approach that pass the genome-wide significance testing threshold of $5 \cdot 10^{-8}$. For each identified genome-wide significant association, we then ask if it is correlated (squared correlation $\geq 0.01$) 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 $\leq 0.01$ with a causal variant within 5Mb is a very conservative one. We present these results in Figure 1a.

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 above 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 1b and Figure S2.

Additionally, we compare the SE $p$-value testing results of gVAMP on the 8,430,446 and 2,174,071 SNP data sets 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 2. 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.

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 1c.

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 S1.

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 S2.

## 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 and Amedeo Roberto Esposito, Philip Schniter, Matthew Stephens and Pragya Sur for providing valuable suggestions and comments on an early version 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, 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. 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).

7. 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).

8. 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).

9. 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).

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

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

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

13. 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).

14. 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).

15. Rangan, S. Generalized Approximate Message Passing for Estimation with Random Linear Mixing. In *IEEE International Symposium on Information Theory* (2011).

16. Sur, P. & Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**, 14516–14525 (2019).

17. Venkataramanan, R., Kögler, K. & Mondelli, M. Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning*, 22120–22144 (2022).

18. Schniter, P. & Rangan, S. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing* **63**, 1043–1055 (2014).

19. Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A. & Zdeborová, L. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory* **62**, 4228–4265 (2016).

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

21. 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).

22. Barbier, J. *et al.* Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences* **116**, 5451–5460 (2019).

23. 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).

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

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

26. 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).

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

28. 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).

29. Fletcher, A. K., Sahraee-Ardakan, M., Rangan, S. & Schniter, P. Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

30. Takeda, K., Uda, S. & Kabashima, Y. Analysis of CDMA systems that are characterized by eigenvalue spectrum. *Europhysics Letters* **76**, 1193 (2006).

31. Tulino, A. M., Caire, G., Verdú, S. & Shamai, S. Support recovery with sparsely sampled free random matrices. *IEEE Transactions on Information Theory* **59**, 4243–4271 (2013).

32. 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).

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

34. Xie, Z. *et al.* Gene set knowledge discovery with enrichr. *Current Protocols* **1**, e90 (2021).

35. Pandey, A., Shao, H., Marks, R. M., Polverini, P. J. & Dixit, V. M. Role of b61, the ligand for the eck receptor tyrosine kinase, in tnf-α-induced angiogenesis. *Science* **268**, 567–569 (1995).

36. Luo, J. *et al.* Regulation of bone formation and remodeling by G-protein-coupled receptor 48. *Development* **136**, 2747–2756 (2009).

37. Karasik, D. *et al.* Genome-wide pleiotropy of osteoporosis-related phenotypes: The framingham study. *Journal of Bone and Mineral Research* **25**, 1555–1563 (2010).

38. Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198—1213.e14 (2020).

39. Fukami, M., Seki, A. & Ogata, T. SHOX Haploinsufficiency as a Cause of Syndromic and Nonsyndromic Short Stature. *Molecular Syndromology* **7**, 3–11 (2016).

40. Zhang, C. *et al.* A novel deletion variant in trappc2 causes spondyloepiphyseal dysplasia tarda in a five-generation chinese family. *BMC Medical Genetics* **21**, 117 (2020).

41. Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214—1231.e11 (2020).

42. Liu, M. *et al.* Maged1 is a negative regulator of bone remodeling in mice. *The American Journal of Pathology* **185**, 2653–2667 (2015).

43. Andreou, A. M. *et al.* Tbx22 missense mutations found in patients with x-linked cleft palate affect dna binding, sumoylation, and transcriptional repression. *The American Journal of Human Genetics* **81**, 700–712 (2007).

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

45. 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).

46. Ma, Z. & Nandy, S. Community detection with contextual multilayer networks. *IEEE Transactions on Information Theory* **69**, 3203–3239 (2023).

47. Hayakawa, R., Nakai, A. & Hayashi, K. Distributed approximate message passing with summation propagation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4104–4108 (IEEE, 2018).

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

49. 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).

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

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

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

53. 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).

# Supplementary information

## Light-speed whole genome association testing and prediction via Approximate Message Passing

Al Depope, 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

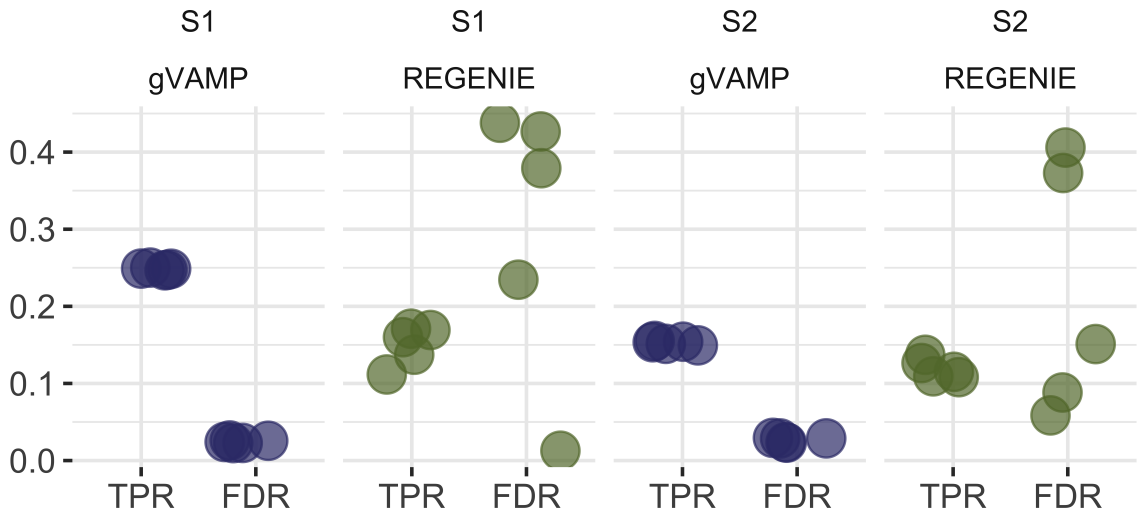


**Figure S1. Comparison of gVAMP and REGENIE association testing within identical data**. True positive rate (TPR) and false discovery rate (FDR) for leave-one-chromosome-out (LOCO) testing where 887,060 markers are used for both the first step of REGENIE and for gVAMP and where all simulated causal 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 (see Methods). We perform five simulation replicates.

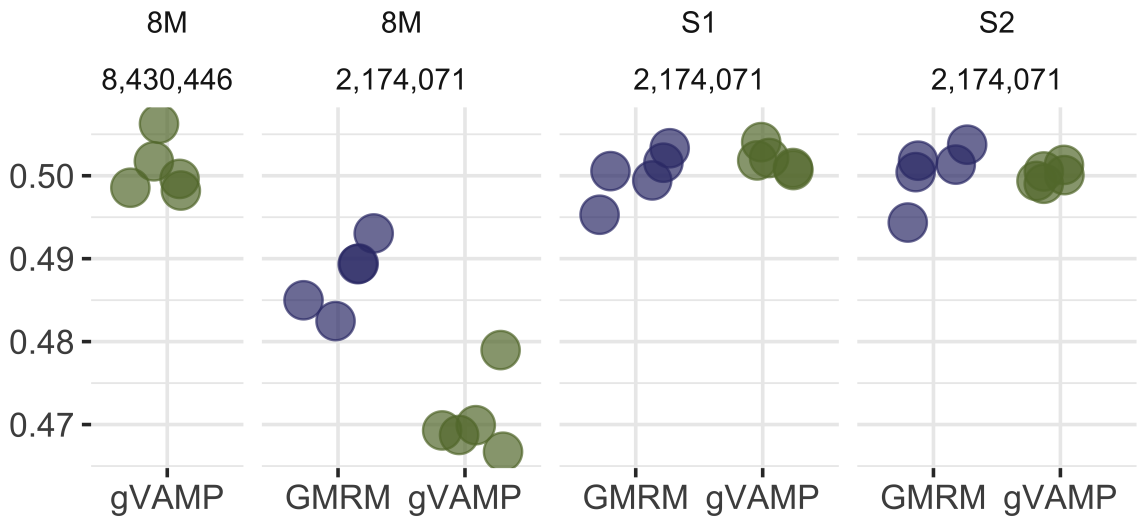


**Figure S2. SNP-heritability estimation of GMRM versus gVAMP with different numbers of SNP markers in the simulation**. Comparison of the proportion of phenotypic variation attributable to either 8,430,446 or a subset of 2,174,071 autosomal single nucleotide polymorphism (SNP) genetic markers (SNP-heritability) 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").

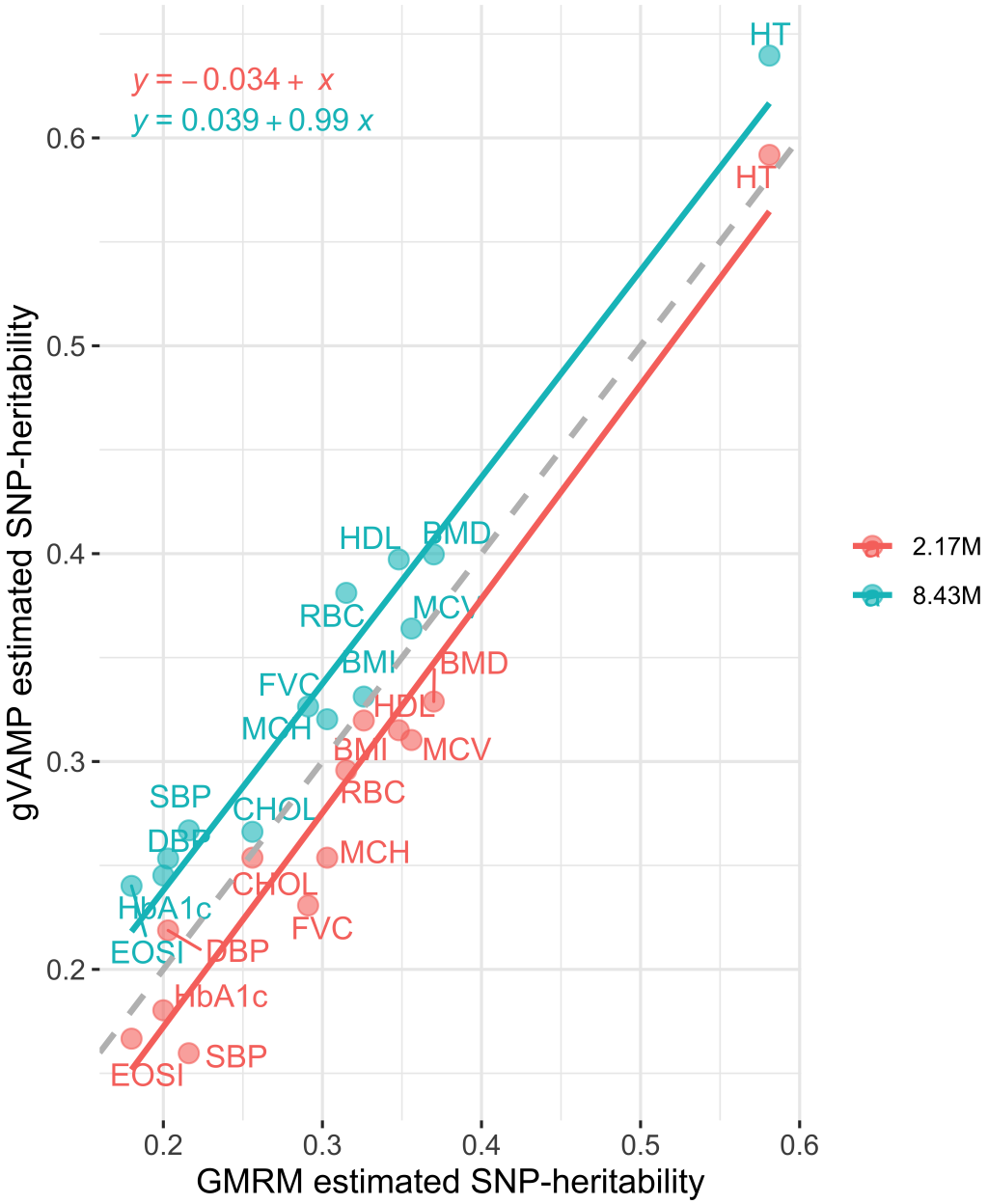


**Figure S3. SNP-heritability estimation of GMRM versus gVAMP with different numbers of SNP markers across 13 trait in the UK Biobank**. Comparison of the proportion of phenotypic variation attributable to 2,174,071 autosomal single nucleotide polymorphism (SNP) genetic markers (SNP-heritability) estimated 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 show 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 $y = x$.

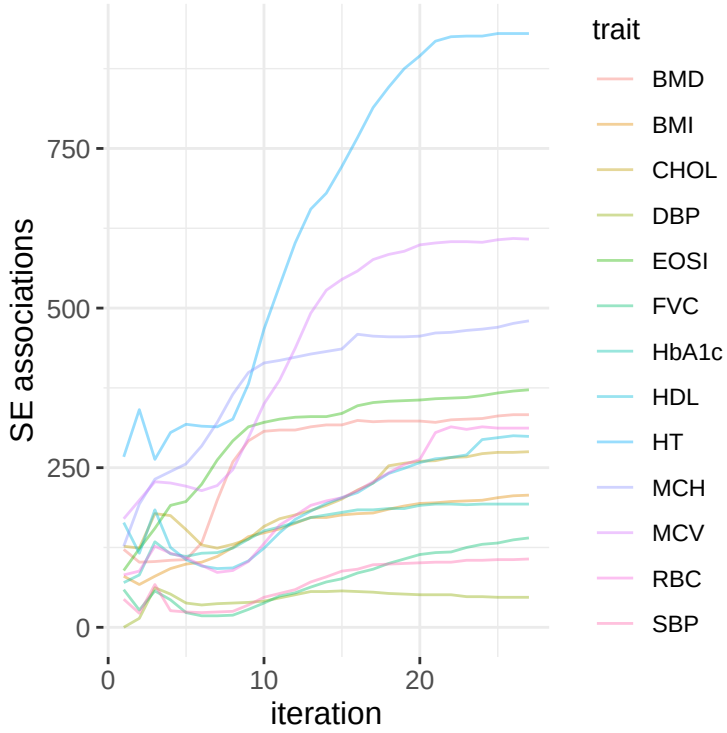


**Figure S4. Convergence of SE $p$-value testing with increasing number of iterations for 13 UK Biobank traits**. AMP theory provides a joint association testing framework, capable of estimating the effects of each genomic position conditional on all other SNP markers. We show this SE $p$-value testing 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

### 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** $x_n$