SparsePro: an efficient genome-wide fine-mapping method integrating summary statistics and functional annotations

Wenmin Zhang[1], Hamed Najafabadi[1,2,3], and Yue Li[1,4,*]

[1]Quantitative Life Science, McGill University, Montreal, Canada

[2]Department of Human Genetics, McGill University, Montreal, Canada

[3]McGill University and Genome Quebec Innovation Centre, Montreal, Canada

[4,*]School of Computer Science, McGill University, Montreal, Canada;

Correspondence to yueli@cs.mcgill.ca

Abstract

Identifying causal variants from genome-wide association studies (GWASs) is challenging due to widespread linkage disequilibrium (LD). Functional annotations of the genome may help prioritize variants that are biologically relevant and thus improve fine-mapping of GWAS results. However, classical fine-mapping methods have a high computational cost, particularly when the underlying genetic architecture and LD patterns are complex. Here, we propose a novel approach, SparsePro, to efficiently conduct functionally informed statistical fine-mapping. Our method enjoys two major innovations: First, by creating a sparse low-dimensional projection of the high-dimensional genotype, we enable a linear search of causal variants instead of an exponential search of causal configurations used in existing methods; Second, we adopt a probabilistic framework with a highly efficient variational

expectation-maximization algorithm to integrate statistical associations and functional priors. We evaluate SparsePro through extensive simulations using resources from the UK Biobank. Compared to state-of-the-art methods, SparsePro achieved more accurate and well-calibrated posterior inference with greatly reduced computation time. We demonstrate the utility of SparsePro by investigating the genetic architecture of five functional biomarkers of vital organs. We identify potential causal variants contributing to the genetically encoded coordination mechanisms between vital organs and pinpoint target genes with potential pleiotropic effects. In summary, we have developed an efficient genome-wide fine-mapping method with the ability to integrate functional annotations. Our method may have wide utility in understanding the genetics of complex traits as well as in increasing the yield of functional follow-up studies of GWASs.

1 Introduction

In recent years, establishment of large biobanks and advances in genotyping and sequencing technologies have enabled large-scale genome-wide association studies (GWASs) [1–3]. Although GWASs have revealed extensive associations between genetic variants and traits of interest, understanding the genetic architecture underlying these genetic associations remains challenging [4–6]. This is because GWASs typically rely on univariate regression models, which are not able to distinguish the causal variants from other variants in linkage disequilibrium (LD) [5, 7, 8].

Several statistical fine-mapping approaches have been proposed for identifying causal variants in GWASs while considering the underlying LD patterns. For instance, BIMBAM [9], CAVIAR [10] and CAVIARBF [11] estimate the posterior inclusion probabilities (PIPs) in a pre-defined locus by evaluating multivariate Gaussian likelihood enumerating all possible configurations. FINEMAP [12] accelerates the inference with a shotgun stochastic search focusing on the most likely subset of causal configurations. However, the number of causal configurations required to evaluate can grow exponentially as the number of causal variants increases, thus tremendously increasing the computational cost if multiple causal variants exist. SuSiE [13] introduces an it-

erative Bayesian stepwise selection algorithm for variable selection, which can also be applied to statistical fine-mapping with greatly improved computational efficiency.

Furthermore, it has been recognized that functional annotations of the genome may help prioritize variants that are biologically relevant, thus improving fine-mapping of GWAS results [8]. For example, PAINTOR [14] and RiVIERA [15] empirically estimate the impacts of functional annotations from statistical evidence, which improves the accuracy of fine-mapping but has a high computational cost, especially when multiple causal SNPs exist in the same locus. PolyFun [16] uses stratified LD score regression [17] to effectively partition trait heritability onto different annotations, and use these estimates of annotation-tagged heritability to specify functional priors for fine-mapping methods.

In this work, we propose a unified probabilistic framework called SparsePro for statistical fine-mapping with the capacity to incorporate functional annotations. Using a sparse projection accompanied with an efficient paired mean field variational expectation maximization inference algorithm [18], we demonstrate in both simulation studies and real data analyses that SparsePro achieves superior accuracy in identifying causal variants as well as computational efficiency compared to state-of-the-art approaches. We highlight its potential utility in enabling fine-mapping at a genome-wide scale and better profiling the genetic architecture for complex traits.

2 Results

2.1 SparsePro method overview

To fine-map causal SNPs, our method takes two lines of evidence (Figure 1). First, from estimated marginal associations between genetic variants and a complex trait of interest, accompanied by matched LD information, we can group correlated genetic variants together and assess their effects jointly. Then we infer the contribution of each SNP towards each group of causal effect separately to obtain PIPs. Second, optionally, if we have knowledge about any functional annotations where true causal SNPs may be enriched, we can estimate the relative

enrichment of these annotations, and prioritize SNPs with these annotations so that they are more likely to be considered causal variants a priori. Consequently, our model yields the PIP for each SNP with the capacity to assess the relative enrichment of candidate functional annotations.

2.2 Our contributions in the context of the existing methods

Our work is related to two existing methods, namely SuSiE [13] and PolyFun [16]. Inspired by the "sum of single effects" model in SuSiE, we introduced a sparse projection of the genotype in our model specification so that the identification of causal variants and estimation of causal effect sizes are separated. This sparse projection avoids exhaustive searching of all causal configurations. For statistical inference, SuSiE adopted an iterative Bayesian stepwise selection algorithm that operates on the Bayes Factors (BFs) [13]. In contrast, we developed a full variational inference algorithm [18] to jointly update the variational parameters for better approximation of the true posterior. Moreover, we have adapted our algorithm to GWAS summary statistics and provided appropriate estimates for the hyperparamters. In functionally informed statistical fine-mapping, PolyFun pioneered the use of genome-wide heritability estimates for setting functional priors [16]. Here, we aggregated the genome-wide statistical fine-mapping evidence with an expectation maximization scheme to prioritize relevant annotations and robustly derive genome-wide functional priors.

2.3 SparsePro demonstrates superior performance in simulation

In our genome-wide simulations, SparsePro consistently demonstrated superior accuracy in identifying true causal variants. SparsePro without annotation (SparsePro-) achieved an area under the precision-recall curve (AUPRC) of 0.3699, higher than the AUPRC of 0.2677 by FINEMAP and the AUPRC of 0.3573 by SuSiE (Figure 2A). Notably, SparsePro had a substantially higher recall (sensitivity) if the precision (1 - false discovery rate) was controlled to be high (Figure 2A), such as when precision > 0.75 (i.e. false discovery rate < 0.25 in Figure 2A). Moreover, SparsePro could incorporate functional priors with the effects of functional annotations accu-

rately estimated (Supplementary Table S1), and had further improved discriminative power. SparsePro+ achieved an AUPRC of 0.4636, outperforming both functionally informed FINEMAP (AUPRC = 0.3088) and functionally informed SuSiE (AUPRC = 0.4042) with functional priors derived by PolyFun. We also found that the performance of SparsePro was not sensitive to the pre-specified number of independent causal effects (Supplementary Table S2 and Supplementary Notes).

Compared to FINEMAP and SuSiE, the PIPs yielded by SparsePro appeared to be much more calibrated. Specifically, at a fixed precision (which corresponded to a fixed PIP threshold), the mean PIP of all SNPs considered to be causal variants by SparsePro was almost identical to the desired precision (Figure 2B). In contrast, the PIPs generated by FINEMAP and SuSiE appeared to be inflated (Figure 2B). For instance, if SNPs with a PIP > 0.8 were to be considered causal variants, SparsePro- and SparsePro+ would both have a median precision of 100% (Figure 2C). The selected SNPs by FINEMAP (median precision = 0.472) and SuSiE (median precision = 0.775) would include an excessive proportion of false positives, even with functional priors (median precision = 0.594 for FINEMAP and 0.775 for SuSiE; Figure 2C). The high precision by SparsePro was consistent for all frequently used PIP thresholds (Figure 2C), despite that FINEMAP and SuSiE possibly had a marginally higher recall.

Notably, SparsePro -/+ conferred not only higher fine-mapping precision, but also higher computational efficiency, as the computation time they required was 17% that of FINEMAP and 38% that of SuSiE (Figure 2D and Supplementary Table S3).

2.4 Fine-mapped SNPs by SparsePro are more enriched in tissue-specific eQTL and confer higher trait heritability

Genome-wide fine-mapping of five functional biomarkers based on the UK Biobank population identified multiple likely causal variants (Supplementary Table S4). As expected, we found that the fine-mapped SNPs were significantly enriched in tissue-specific eQTL for all five biomarkers, while results based on SparsePro showed the strongest enrichment (Figure 3A). For example, for total protein, the fine-mapped SNPs determined by SparsePro- were 4.00-fold (95%

CI: 3.25-4.92) more likely to be liver-specific eQTL than non-fine-mapped SNPs, compared to a 1.54-fold (95% CI: 1.35-1.75) enrichment based on fine-mapped SNPs by SuSiE. While SuSiE was substantially improved by functional priors derived from PolyFun with a 2.20-fold (95% CI: 1.97-2.45) enrichment, SparsePro+ identified fine-mapped SNPs demonstrating the highest biological relevance, being 4.06-fold (95% CI: 3.31-4.97) more likely to be liver-specific QTL.

Moreover, at most PIP thresholds, SNPs fine-mapped by SparsePro- explained a higher proportion of z-score variance than the same number of the most likely causal SNPs identified by SuSiE (Figure 3B and Supplementary Table S5). With functional annotations (Supplementary Table S5), though the PolyFun-informed SuSiE was able to locate SNPs that appeared to have a better predictive performance at some PIP thresholds for glucose and pulse rate, the fine-mapped SNPs by SparsePro+ consistently achieved a higher SNP heritability for estimated glomerular filtration rate, FEV1-FVC ratio, as well as total protein (Figure 3C and Supplementary Table S6).

2.5 Evidence of pleiotropic effects on functional biomarkers for five vital organs

Although the five biomarkers were considerably polygenic (Figure 4A), we found two SNPs which were potential causal variants for at least three of these biomarkers. Specifically, SNP rs1260326 (Figure 4B), a missense variant in the *GCKR* gene was fine-mapped for glomerular filtration rate (PIP = 1.000), blood glucose level (PIP = 0.998), pulse rate (PIP = 0.823) and total protein level (PIP = 1.000). Another SNP, rs5742915 (Figure 4C), a missense variant in the *PML* gene was fine-mapped for FEV1-FVC ratio (PIP = 0.858), pulse rate (PIP = 1.000) and total protein level (PIP = 0.987). These findings, along with other SNPs exhibiting pleiotropic effects (Supplementary Table S4), might point towards mechanisms of coordination between vital organs and might become viable targets for experimental validations.

3 Discussion

Accurately identifying trait-determining and disease-causing variants is fundamental in genetics and particularly important for appropriately interpreting GWAS results [5, 8]. In this work, we developed SparsePro, an efficient fine-mapping method to help prioritize causal variants for complex traits, possibly with prior functional information. Through genome-wide simulations, we showed that SparsePro was highly accurate and computationally efficient compared to existing methods. By fine-mapping genetic associations with five biomarkers for vital organ functions, we demonstrated that SparsePro identified candidate variants that were biologically relevant, including two variants with pleiotropic effects which might indicate genetically encoded coordinations between vital organs.

SparsePro has three important features. First of all, we implemented an efficient variational inference algorithm to approximate the posterior distribution of the causal variant indicators, instead of exhaustively searching through all possible causal configurations or performing stepwise regression. As a result, SparsePro can be significantly faster than classical fine-mapping methods, such as FINEMAP [12], and is more than twice faster than SuSiE [13], which has a similar variable selection framework but implements an iterative Bayesian stepwise selection procedure. The substantially enhanced computational efficiency thus permits statistical fine-mapping of large chunks of the genome instead of analyzing genetic associations on a per-locus basis as in most existing follow-up studies of GWASs. In our simulation studies, compared to locus-wise fine-mapping based on COJO lead SNPs, such a genome-wide fine-mapping requires neither a pre-specified p-value threshold (e.g. $\mathrm{p} < 5 \times 10^{-8}$) for determining candidate loci nor an arbitrary number of causal effects per locus. If functional annotations are available, the estimation of functional enrichment may also be more robust by including more variants.

Second, the PIPs yielded by SparsePro are well-calibrated. The properties of the paired mean field variational family ensure that our approximation matches closely with the true posterior distribution of the causal variant indicators [18]. Moreover, given GWAS summary statistics, we provide estimates for hyperparameters including τ_{y} and τ_{β} that are reasonable in the context of polygenic trait genetics. Consequently, at several commonly used PIP thresholds

for defining causal variants, SparsePro shows improved control of false positives, and demonstrates high accuracy in identifying causal variants, which, in real data examples, appear to be biologically meaningful.

Third, we propose and implement a probabilistic model that coherently integrate statistical evidence and functional prior information. The key difference between SparsePro+ and other methods that build functional priors, such as PolyFun [16] and PAINTOR [14], is that each annotation is tested for its relevance with the trait of interest before being used to derive the priors in our model. Therefore, functional annotations serve as complementary evidence when statistical evidence is not sufficient to discern causal variants. We posit that this property guarantees the prior information in functional annotations are utilized rigorously.

We note that SparsePro can further improved with the following future directions. First, SparsePro generally requires that the supplied LD reference panel matches well with that of the GWAS study population to guarantee convergence. We advocate that along with GWAS summary statistics, LD information should also be made publicly available, especially for studies involving populations of diverse genetic ancestries. Second, SparsePro now only supports binary annotations while compatibility with continuous annotations is also desirable. Last, the current variational expectation maximization scheme cannot jointly estimate the enrichment of highly correlated annotations accurately, thus selection of representative annotations is needed. Modifications that enable the inclusion of multiple correlated informative annotations, such as cell type-specific annotations can further improve the utility of SparsePro.

In summary, we have developed an efficient genome-wide fine-mapping method with the ability of integrate functional annotations. Our method may have wide utility in understanding the genetic architecture of complex traits, identifying target genes, and increasing the yield of functional follow-up studies of GWASs.

4 Method

4.1 SparsePro model specification

We assume the following data generative process (Figure 1) for a polygenic trait y:

1. Among G SNPs under investigation, the probability of the g-th SNP being causal is denoted as $\tilde{\pi}_{\mathrm{g}}$. We specify that

$$\tilde{\pi}_{\mathrm{g}} = \mathrm{softmax}(\mathbf{A}_{\mathrm{g}}\mathbf{w}) = \frac{\exp(\mathbf{A}_{\mathrm{g}}\mathbf{w})}{\sum_{\mathrm{g}'=1}^{\mathrm{G}} \exp(\mathbf{A}_{\mathrm{g}'}\mathbf{w})}$$

where $\mathbf{A}_{\mathrm{g},1\times\mathrm{M}}$ is the annotation vector of M candidate annotations for the g-th SNP; and $\mathbf{w}_{\mathrm{M}\times 1}$ is the vector of logarithm of relative enrichment. Here, we use the softmax function to ensure the prior probabilities sum up to 1. If no functional information is provided, the prior probability of being causal is considered equal for all SNPs, i.e. $\tilde{\pi}_{\mathrm{g}} = \frac{1}{\mathrm{G}}$.

2. We assume that there exist K independent causal effects, and that

$$\mathbf{s}_{\mathrm{k}} \sim \mathrm{Multinomial}(1, \tilde{\boldsymbol{\pi}})$$

where $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, ..., \tilde{\pi}_{\mathrm{G}})$ and $\mathbf{s}_{\mathrm{k}}$ is a 0/1 indicator vector of length G denoting which SNP is the causal SNP underlying the k-th causal effect, $\mathrm{k} \in \{1, ..., \mathrm{K}\}$.

3. The causal effect sizes are sampled from a normal distribution, i.e.

$$\beta_{\mathrm{k}} \sim \mathrm{N}(0, \tau_{\beta_{\mathrm{k}}}^{-1})$$

4. The continuous trait $\mathbf{y}_{\mathrm{N}\times 1}$ is generated from

$$\mathbf{y} = \mathrm{X}\sum_{\mathrm{k}} \mathbf{s}_{\mathrm{k}}\beta_{\mathrm{k}} + \boldsymbol{\epsilon}$$

or in matrix form:

$$\mathbf{y} = \mathrm{XS}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathrm{X}_{\mathrm{N}\times\mathrm{G}}$ is the full genotype matrix, $\mathrm{S}_{\mathrm{G}\times\mathrm{K}}$ is the sparse projection matrix $\boldsymbol{\beta}_{\mathrm{K}\times 1}$ is the causal effect vector and $\boldsymbol{\epsilon}_{\mathrm{N}\times 1}$ captures trait variance not attributable to the modelled genetic effects and $\epsilon_{\mathrm{i}} \sim \mathrm{N}(0, \tau_{\mathrm{y}}^{-1})$.

4.2 A variational inference algorithm for fine-mapping

With this model specification (Figure 1), finding the causal variants is equivalent to inferring the sparse projections $\mathbf{s}_{\mathrm{k}}$ and the effect sizes β_{k} given $\mathbf{y}$ and X for $\mathrm{k} \in \{1, ..., \mathrm{K}\}$. Namely, we aim to infer

$$\mathrm{p}(\mathrm{S}, \boldsymbol{\beta}|\mathbf{y}, \mathrm{X}, \tilde{\boldsymbol{\pi}}, \tau_{\beta}, \tau_{\mathrm{y}}) = \frac{\mathrm{p}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X}, \tilde{\boldsymbol{\pi}}, \tau_{\beta}, \tau_{\mathrm{y}})}{\mathrm{p}(\mathbf{y}|\mathrm{X}, \tilde{\boldsymbol{\pi}}, \tau_{\beta}, \tau_{\mathrm{y}})}$$

Here, we do not have a closed-form solution for this posterior by explicitly calculating the denominator, since the number of possible causal configurations grows exponentially with G. Unlike many fine-mapping approaches using sampling-based methods to search through a subset of possible causal configurations [12] [19], we adopt a paired mean field factorization of variational family to approximate the posterior [18]. It has been shown that the paired mean-field variational family has similar mode and shape as the desired posterior distribution, and that such inference can achieve high accuracy with substantially improved computational efficiency [18].

Specifically, the variational distribution used to approximate the above posterior distribution is proposed as

$$\mathrm{q}(\mathrm{S}, \boldsymbol{\beta}) = \prod_{\mathrm{k}} \mathrm{q}(\mathbf{s}_{\mathrm{k}}, \beta_{\mathrm{k}}) = \prod_{\mathrm{k}} \mathrm{q}(\mathbf{s}_{\mathrm{k}})\mathrm{q}(\beta_{\mathrm{k}}|\mathbf{s}_{\mathrm{k}})$$

This variational distribution preserves the dependency between $\mathbf{s}_{\mathrm{k}}$ and β_{k} given $\mathbf{y}$ observed. To find the best approximation, we minimize the Kullback-Leibler (KL) divergence between the posterior distribution and the proposed variational distribution, which is equivalent to maximizing the evidence lower bound (ELBO) [20]:

$$\mathrm{ELBO} = \mathrm{E_q}[\mathrm{logp}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta} | \mathrm{X}, \tilde{\boldsymbol{\pi}}, \tau_\beta, \tau_\mathrm{y})] - \mathrm{E_q}[\mathrm{logq}(\mathrm{S}, \boldsymbol{\beta})]$$

Based on the mean field assumptions [18], this optimization can be conducted iteratively for each causal effect k and SNP g with the following closed-form updates until convergence (details available in Supplementary Notes):

- We update posterior effect size for the g-th SNPs in the k-th causal effect

$$\mu^*_{\mathrm{kg}} = \frac{\tau_\mathrm{y}}{\tau^*_{\mathrm{kg}}}(\mathbf{X}_\mathrm{g}^\top \mathbf{y} - \mathbf{X}_\mathrm{g}^\top \mathrm{X} \sum_{\mathrm{k}' \neq \mathrm{k}} \boldsymbol{\gamma}^*_{\mathrm{k}'} \circ \boldsymbol{\mu}^*_{\mathrm{k}'}) \tag{1}$$

  with

$$\tau^*_{\mathrm{kg}} = \mathbf{X}_\mathrm{g}^\top \mathbf{X}_\mathrm{g} \tau_\mathrm{y} + \tau_{\beta_\mathrm{k}} \tag{2}$$

  where $\circ$ represents element-wise multiplication of vectors.

- We update the posterior probability of the g-th SNP being causal in the k-th causal effect

$$\gamma^*_{\mathrm{kg}} = \mathrm{softmax}(\log \tilde{\pi}_\mathrm{g} - \frac{1}{2} \log \frac{\tau^*_{\mathrm{kg}}}{2\pi} + \frac{\tau^*_{\mathrm{kg}} \mu^{*2}_{\mathrm{kg}}}{2}) \tag{3}$$

  We take the maximum of these K probabilities as the PIP for the g-th SNP.

4.3 Adaptation to GWAS summary statistics

The above variational inference algorithm requires access to large datasets containing both individual-level genotype X and phenotype data $\mathbf{y}$. Since a growing number of GWASs have released publicly available summary statistics (i.e., marginal effect size estimate $\hat{\beta}_\mathrm{g}$ and its standard error $\mathrm{se_g}$ for the g-th SNP), we have also adapted SparsePro to directly operate on these summary statistics with additional information from an LD reference panel (i.e. estimates of pairwise SNP-SNP correlation).

Specifically, if we have reasonable surrogates for $\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}}$, $\mathrm{X}^{\top}\mathrm{X}$ and $\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{y}$, we can plug them into Equations (1),(2), and (3). We include two forms of reformulation with respect to whether the genotypes are standardized to have zero mean and unit variance in the GWAS.

1. If the genotypes are standardized, we have

$$\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}} = \mathrm{N}$$

$$\mathrm{X}^{\top}\mathrm{X} = \mathrm{N} * \mathrm{LD}$$

$$\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{y} = \mathrm{N}\hat{\beta}_{\mathrm{g}}$$

   where N is the sample size.

2. If the genotypes are not standardized, we have

$$\hat{\beta}_{\mathrm{g}} = (\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}})^{-1}\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{y}$$

$$\mathrm{se}_{\mathrm{g}} = \sqrt{\mathrm{var}(\mathrm{y})(\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}})^{-1}}$$

   Therefore,

$$\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}} = \frac{\mathrm{var}(\mathrm{y})}{(\mathrm{se}_{\mathrm{g}}^{2})}$$

$$\mathrm{X}^{\top}\mathrm{X} = \mathrm{LD} * (\mathbf{se}^{\top}\mathbf{se})$$

$$\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{y} = \mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}} * \hat{\beta}_{\mathrm{g}}$$

   Notably, if $\mathbf{y}$ has been standardized to have unit variance prior to a GWAS, we naturally supply $\mathrm{var}(\mathrm{y}) = 1$. Otherwise, it can be estimated as $\mathrm{var}(\mathrm{y}) = 2\mathrm{Np}(1-\mathrm{p})\mathrm{se}^{2}$ where N (the study sample size), p (minor allele frequencies), and se (standard errors of effect size estimates) are usually available in GWAS summary statistics.

4.4 Variational expectation maximization for integrating functional annotations

To estimate the relative enrichment of functional annotations and further prioritize variants, we adopt an variational expectation maximization scheme to maximize ELBO with respect to the logarithm of relative enrichment ($\mathbf{w}$) first and then use the estimate $\hat{\mathbf{w}}$ to calculate $\tilde{\pi}_{\mathrm{g}}$ (prior probability of being causal) for each SNP.

Suppose we have M candidate annotations and $\mathrm{A_{gm}}$ ($\mathrm{m} \in \{1, ..., \mathrm{M}\}$) is a 0/1 indicator denoting whether the g-th SNP has the m-th annotation. By setting the derivative of ELBO with respect to $\mathrm{w_m}$ as 0, we have the following estimate for the logarithm of relevant enrichment (detailed in Supplementary Notes),

$$\mathrm{w_m} = \log(\frac{\mathrm{r_1}/\mathrm{r_0}}{\mathrm{k_1}/\mathrm{k_0}})$$

where

$$\mathrm{k_1} = \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 1] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})$$

$$\mathrm{k_0} = \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 0] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})$$

$$\mathrm{r_1} = \sum_{\mathrm{k,g}} [\mathrm{A_{gm}} = 1] \gamma^*_{\mathrm{kg}}$$

$$\mathrm{r_0} = \sum_{\mathrm{k,g}} [\mathrm{A_{gm}} = 0] \gamma^*_{\mathrm{kg}}$$

We note that this metric is equivalent to the logarithm of a relative risk, thus its standard error can be calculated as

$$\mathrm{se}(\mathrm{w_m}) = \sqrt{\frac{1}{\mathrm{r_1}} + \frac{1}{\mathrm{r_0}} - \frac{1}{\mathrm{k_1}} - \frac{1}{\mathrm{k_0}}}$$

We evaluate the significance of annotation enrichment with the log likelihood ratio test (G-test) [21]. Only annotations which demonstrate statistical significance are included in our model

to update the prior probability of being causal for each SNP. Specifically,

$$\tilde{\pi}_{\mathrm{g}} = \mathrm{softmax}(\sum_{\mathrm{m}} \mathrm{A}_{\mathrm{gm}} \hat{\mathrm{w}}_{\mathrm{m}})$$

This functionally informed prior may help prioritize causal SNPs in addition to statistical evidence.

4.5 Hyperparameter settings

We have three hyperparameters: number of causal effect K, inverse of unexplained variance τ_{y} and inverse variance of causal effect sizes $\tau_{\beta_{\mathrm{k}}}$ in our model. We show in Supplementary Notes that our model is not sensitive to the setting of K as long as K is larger than the actual number of independent effects, except that increasing K marginally increases the computation time.

We set τ_{y} as

$$\tau_{\mathrm{y}} = \frac{1}{\mathrm{var(y)} * (1 - \mathrm{h}^2)}$$

where h^2 is the local SNP heritability that can be estimated by a modified Heritability Estimation from Summary Statistics (HESS) [22] based on GWAS summary statistics (Supplementary Notes)

We set τ_{β} as

$$\tau_{\beta} = \frac{\mathrm{k}}{\mathrm{var(y)} * \mathrm{h}^2}$$

for each of the independent causal effects, $\mathrm{k} \in \{1, ..., \mathrm{K}\}$ to account for different effect sizes and to improve model identifiability.

4.6 Simulation studies

We conducted simulations to showcase the efficiency and utility of our method. We leveraged resources from the UK Biobank [1]. Specifically, we first retained 353,606 White British ances-

try participants by excluding one individual from each pair of closely related individuals (who had a 3rd degree or closer relationship). We then retrieved the genotypes of these individuals based on 271,699 SNPs which had a minor allele frequency ≥ 0.001 and an imputation quality score ≥ 0.6 on chromosome 22. Next, we sampled 50 causal SNPs with a two-fold relative enrichment amongst SNPs that were annotated as "conserved sequences" [23], "DNase I hypersensitive sites" (DHS) [24], "non-synonymous" [25], or overlapped with histone marks H3K27ac [26] or H3K4me3 [24]. We used the GCTA GWAS simulation pipeline [27] to simulate a continuous trait with a per-chromosome heritability of 0.01. We tested the association between each SNP and this simulated trait, and obtained GWAS summary statistics using the fastGWA software [28]. This process was replicated 22 times to simulate a GWAS. We obtained LD information calculated using the UK Biobank participants from `https://alkesgroup.broadinstitute.org/UKBB_LD/` [16]. These LD matrices were generated for genome-wide SNPs binned into sliding windows of 3 Mb where two neighboring windows had a 2-Mb overlap.

We applied SparsePro to the GWAS summary statistics with the above LD information, and iterated over all sliding windows, first without any functional annotation information. We denoted the fine-mapping results as "SparsePro-". Next, we aggregated the results from all 22 replications to estimate the relative enrichment for ten binary functional annotations. In addition to the five annotations simulated to be enriched of causal SNPs, we also included "actively transcribed regions" [29], "transcription start sites" [29], "promoter regions" [30], "5'-untranslated regions" [25], and "3'-untranslated regions" [25].

Annotations with a G-test p-value $< 1 \times 10^{-6}$ were selected to conduct functionally informed fine-mapping, and the results were denoted as "SparsePro+". τ_β and τ_{y} were set according to aforementioned empirical estimates. PIPs for SNPs in the 1-Mb centre of each 3-Mb sliding window were extracted.

We also performed fine-mapping with some of the state-of-the-art methods. To perform fine-mapping with conditional and joint (COJO) analyses [31] and FINEMAP [12], we first selected COJO lead SNPs based on GWAS summary statistics by performing stepwise model selection implemented in the GCTA-COJO software [27]. We then applied FINEMAP with shotgun stochastic search to SNPs in a 1-MB window centered at lead SNP to obtain PIPs. We wrote

an in-house script using the “susie_rss” function to perform genome-wide fine-mapping with SuSiE in the same sliding windows as SparsePro. We aggregated summary statistics from 22 replications and used PolyFun with the “baselineLF2.2.UKB” model [16] to calculate functional priors. The “baselineLF2.2.UKB” model contained all annotations used in SparsePro as well as additional pre-computed LD-related annotations for achieving optimal performance of PolyFun [16]. The estimated priors were provided to SuSiE via “prior_weights” and to FINEMAP via the –prior-snps function, respectively. The maximal number of causal SNPs in each locus was set to 5 for all methods.

We compared the performance of these methods in terms of precision (1 - false discovery rate), recall, calibration of PIPs, as well as computation time, all evaluated on a 2.1 GHz CPU node on Compute Canada.

4.7 Fine-mapping genetic determinants of functional biomarkers for vital organs

To investigate the genetically coordination mechanisms of vital organs, we performed GWAS in the UK Biobank [1] for five functional biomarkers: forced expiratory volume in one second to forced vital capacity (FEV1-FVC) ratio for lung function, estimated glomerular filtration rate for kidney function, pulse rate for heart function, total protein for liver function and blood glucose level for pancreatic islet function. For each trait, we first regressed out the effects of age, age^2, sex, genotyping array, recruitment centre, and the first 20 genetic principal components before inverse normal transforming the residuals to have zero mean and unit variance. We then performed GWAS analysis on the resulting z-scores with the fastGWA software [27, 28] to obtain summary statistics.

Using the summary statistics and the matched LD information [16], we performed genome-wide fine-mapping with SparsePro-, SparsePro+, SuSiE and PolyFun-informed SuSiE as described in simulation studies, except that the number of causal effects was set to 9 for each LD region to account for potential more causal variants.

To evaluate the biological relevance of SNPs fine-mapped by different methods, we assessed

their relative enrichment in tissue-specific expression quantitative loci (eQTL). Tissue-specific eQTL identifieid in the most recent release of the Genotype-Tissue Expression (GTEx) project [32, 33] were obtained from `https://gtexportal.org/home/datasets`. The eQTL information was not used by any functionally informed fine-mapping methods.

Additionally, we derived trait heritability conferred by fine-mapped SNPs, by SparsePro- and SparsePro+ respectively, at several commonly used PIP thresholds for determining causal variants: 0.50, 0.80, 0.90, 0.95, and 0.99. The adjusted R^2 obtained from multivariate linear regression of the biomarker z-scores against all fine-mapped SNPs was used as a surrogate of the SNP heritability. We compared these results to heritability captured by the same number of SNPs fine-mapped by SuSiE and PolyFun-informed SuSiE, separately at each PIP threshold. For instance, if SparsePro- identified J SNPs with a PIP > 0.5, we would select J SNPs with the highest PIP determined by SuSiE and compare the adjusted R^2. Notably, this analysis evaluates predictive associations instead of actual causality, hence adjusted R^2 is not a direct indicator of the validity of the fine-mapping results.

We selected SNPs with a PIP > 0.8 to explore possible pleiotropic effects using phenogram [34]. Loci with potential pleiotropic effects were visualized using LocusZoom [35].

5 Figure Legends

Figure 1. Graphical model representation of SparsePro. $\mathbf{w}$ denotes the logarithm of relative enrichment vector for functional annotations; $\mathbf{A_g}$ denotes the annotation vector for the g-th SNP; $\tilde{\pi}_{\mathrm{g}}$ denotes the functional prior probability of being the causal SNP for the g-th SNP; β_{k} denotes the causal effect size for the k-th effect, where $\beta_{\mathrm{k}} \sim \mathrm{N}(0, \tau_\beta)$; $\mathrm{s_k}$ denotes the sparse indicator for the k-th effect; $\mathbf{X_i}$ denotes the genotype vector for the i-th individual; and $\mathrm{y_i}$ denotes the phenotype for the i-th individual, where $\mathrm{y_i} \sim \mathrm{N}(\sum_{\mathrm{k}} \mathbf{X_i} \mathrm{s_k} \beta_{\mathrm{k}}, \tau_{\mathrm{y}})$.

Figure 2. SparsePro demonstrated improved accuracy and computational efficiency in genome-wide simulation results. (A) Precision-recall curves. Precision = 1 - false discovery rate. (B) Calibration of posterior inclusion probabilities. Mean posterior inclusion probability for all SNPs considered as causal variants at a fixed precision represents the expected precision. The black dashed line indicates optimal calibrations. (C) Precisions and recalls obtained at five frequently used posterior inclusion probability thresholds. Error bars indicate inter-quartile ranges. (D) Comparison of computational time. Boxes denote inter-quartile ranges and the line inside each box indicates the median running time.

Figure 3. Biological relevance of fine-mapped SNPs for five biomarkers in vital organs. (A) Relative enrichment of causal variants in tissue-specific eQTLs. Target traits and the corresponding organs are indicated. Estimates of relative enrichment with 95% confidence intervals are plotted on a logarithmic scale. (B) Comparison of the proportion of total trait variance explained by fine-mapped SNPs between SparsePro- and SuSiE.(C) Comparison of the proportion of total variance explained by fine-mapped SNPs between SparsePro+ and PolyFun informed SuSiE. Fine-mapped SNPs were identified at five posterior inclusion probability thresholds.

Figure 4. Fine-mapping of genetic associations for five functional biomarkers. (A) Illustration of genome-wide distribution of fine-mapped SNPs by 22 chromosomes. SNPs with a posterior inclusion probability > 0.80 are indicated. Two loci with potential pleiotropic effects are highlighted by red rectangles. Locus zoom plots are presented for these two loci: (B) locus with fine-mapped SNP rs1260326. and (C) locus with fine-mapped SNP rs5742915. SNPs in a $\pm$500 kb window are included, colored by r^2 with the corresponding fine-mapped SNP.

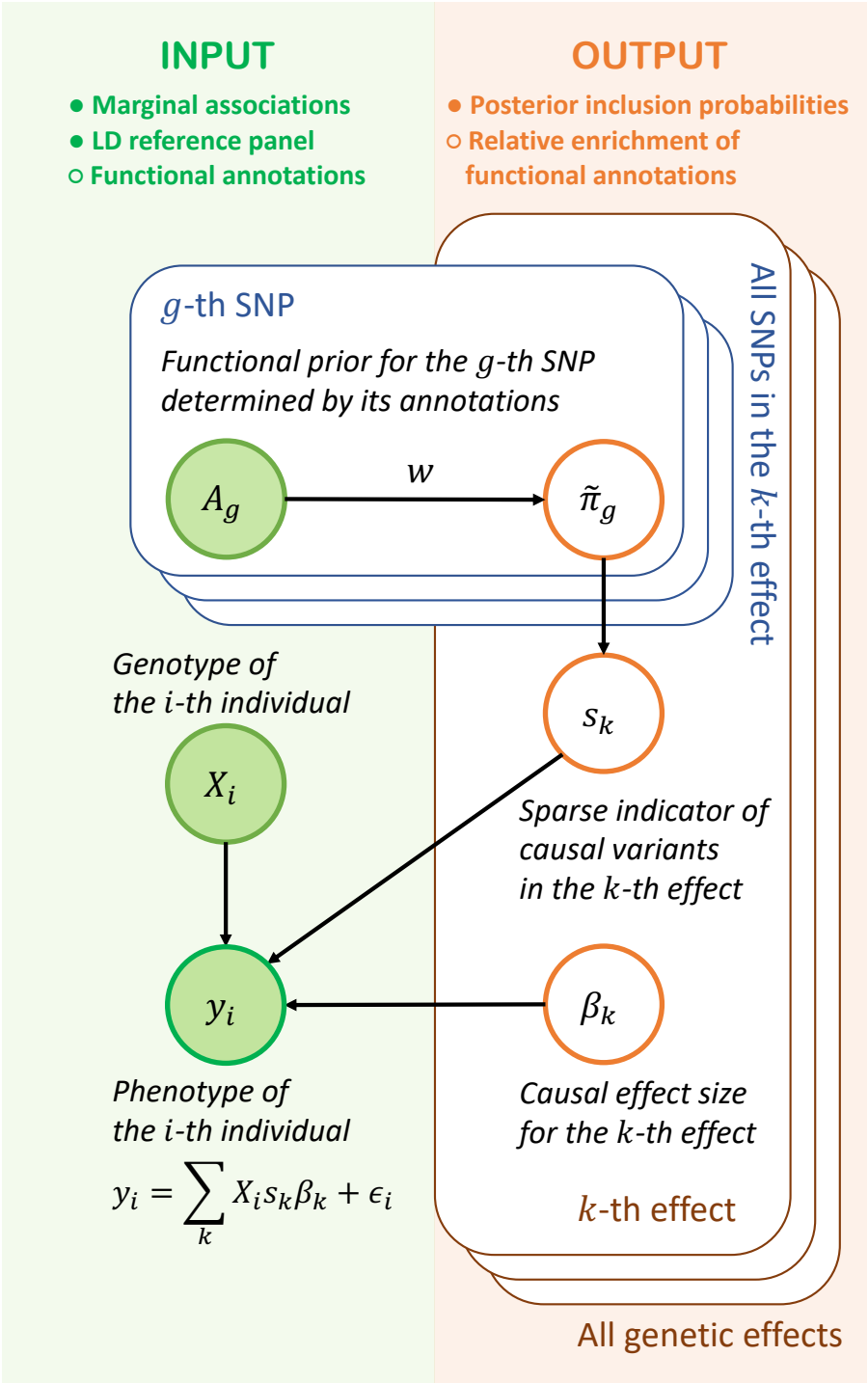

Figure 1: Graphical model representation of SparsePro. $\mathbf{w}$ denotes the logarithm of relative enrichment vector for functional annotations; $\mathbf{A_g}$ denotes the annotation vector for the g-th SNP; $\tilde{\pi}_\mathrm{g}$ denotes the functional prior probability of being the causal SNP for the g-th SNP; β_k denotes the causal effect size for the k-th effect, where $\beta_\mathrm{k} \sim \mathrm{N}(0, \tau_\beta)$; $\mathrm{s_k}$ denotes the sparse indicator for the k-th effect; $\mathbf{X_i}$ denotes the genotype vector for the i-th individual; and $\mathrm{y_i}$ denotes the phenotype for the i-th individual, where $\mathrm{y_i} \sim \mathrm{N}(\sum_\mathrm{k} \mathbf{X_i} \mathrm{s_k} \beta_\mathrm{k}, \tau_\mathrm{y})$.

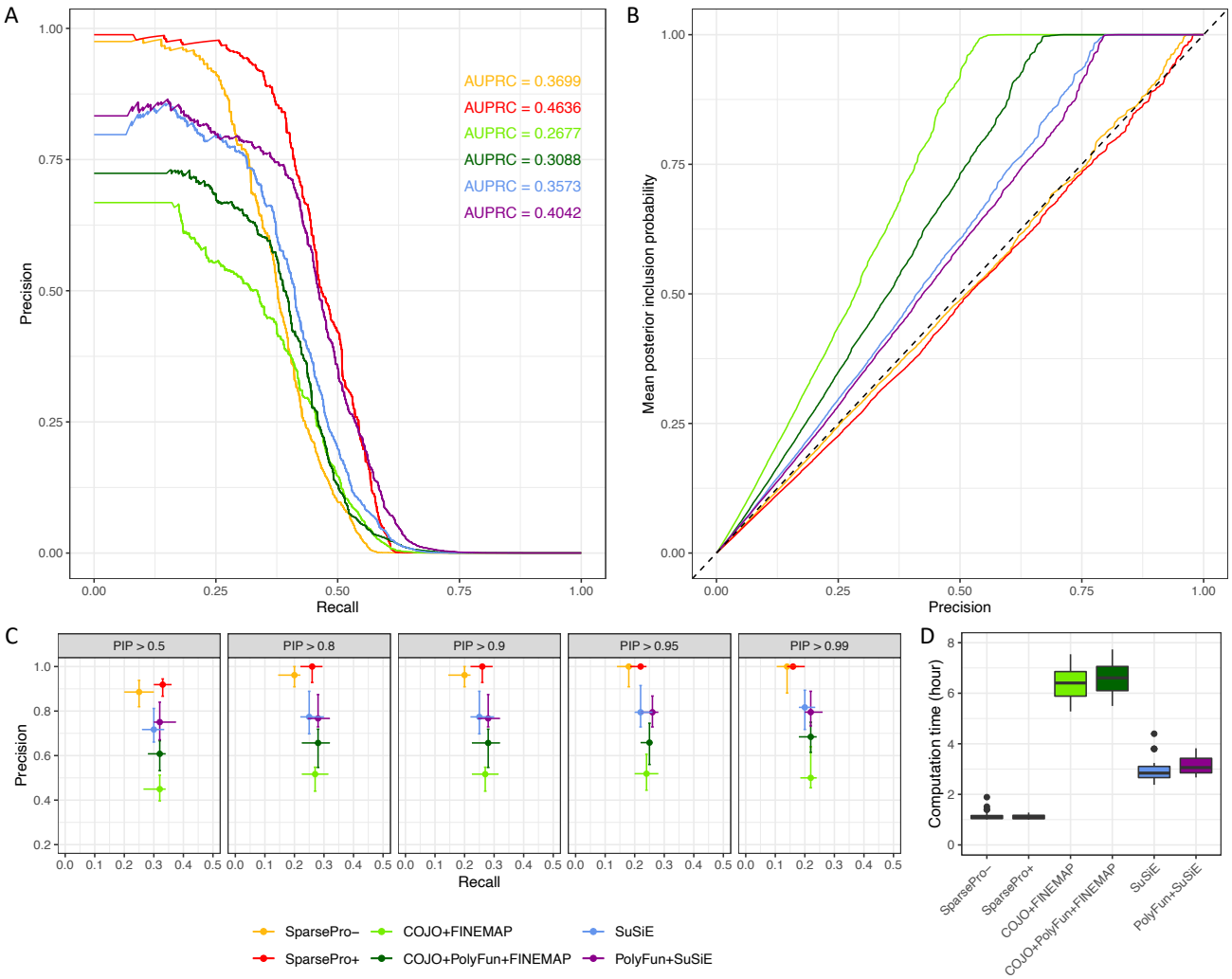

Figure 2: SparsePro demonstrated improved accuracy and computational efficiency in genome-wide simulation results. (A) Precision-recall curves. Precision = 1 - false discovery rate. (B) Calibration of posterior inclusion probabilities. Mean posterior inclusion probability for all SNPs considered as causal variants at a fixed precision represents the expected precision. The black dashed line indicates optimal calibrations. (C) Precisions and recalls obtained at five frequently used posterior inclusion probability thresholds. Error bars indicate inter-quartile ranges. (D) Comparison of computational time. Boxes denote inter-quartile ranges and the line inside each box indicates the median running time.

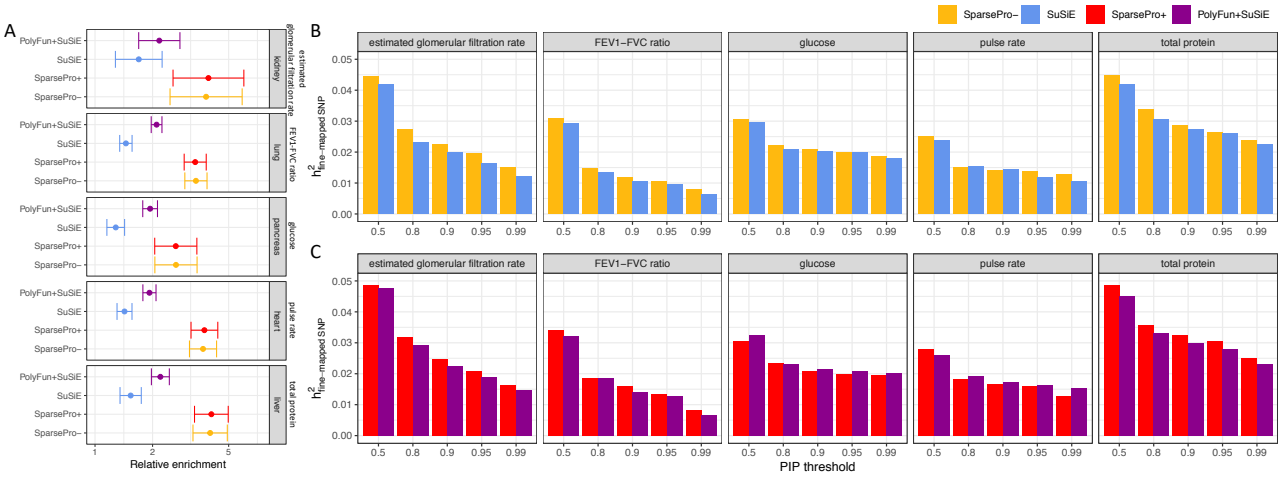

Figure 3: Biological relevance of fine-mapped SNPs for five biomarkers in vital organs. (A) Relative enrichment of causal variants in tissue-specific eQTLs. Target traits and the corresponding organs are indicated. Estimates of relative enrichment with 95% confidence intervals are plotted on a logarithmic scale. (B) Comparison of the proportion of total trait variance explained by fine-mapped SNPs between SparsePro- and SuSiE.(C) Comparison of the proportion of total variance explained by fine-mapped SNPs between SparsePro+ and PolyFun informed SuSiE. Fine-mapped SNPs were identified at five posterior inclusion probability thresholds.

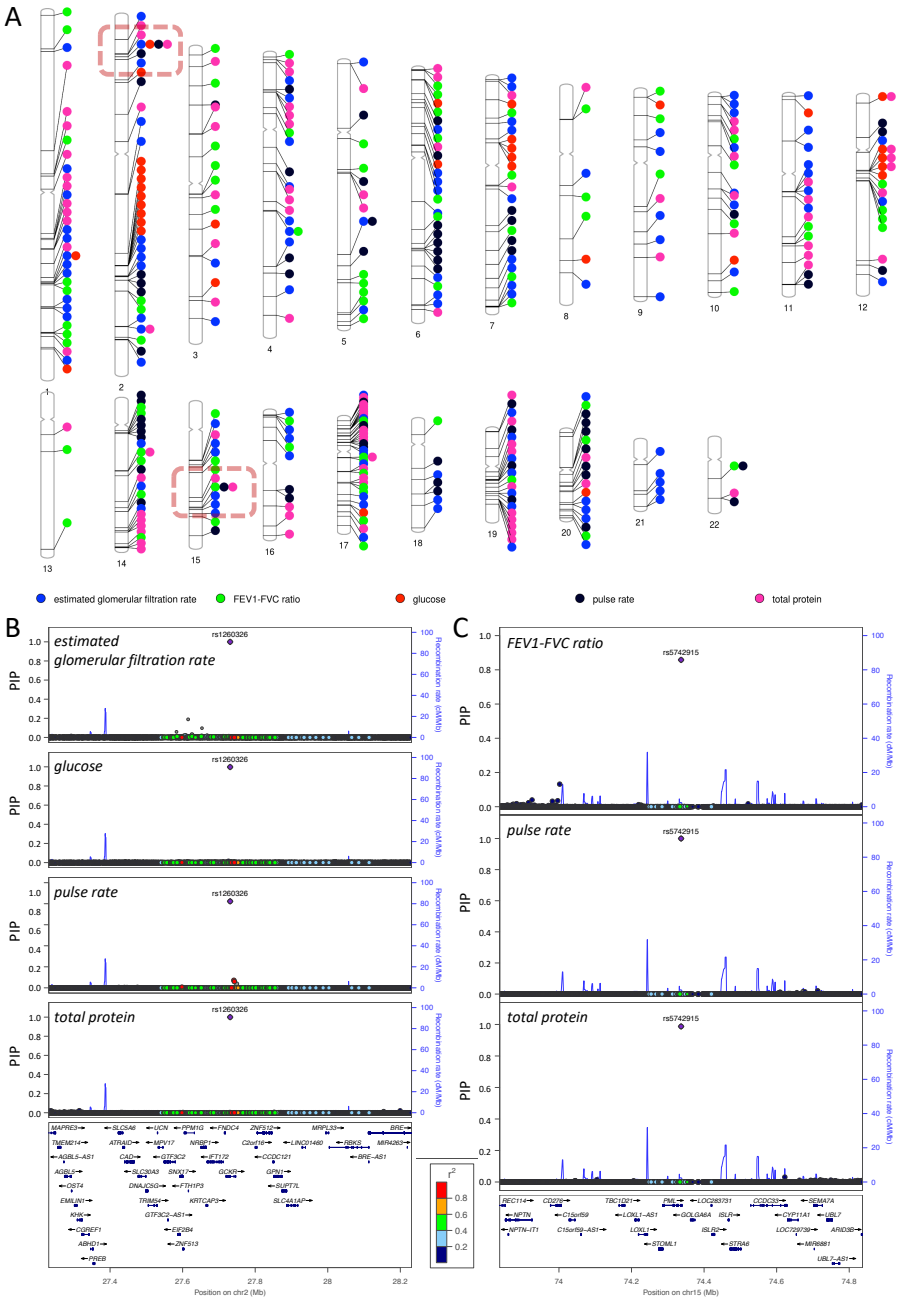

Figure 4: Fine-mapping of genetic associations for five functional biomarkers. (A) Illustration of genome-wide distribution of fine-mapped SNPs by 22 chromosomes. SNPs with a posterior inclusion probability > 0.80 are indicated. Two loci with potential pleiotropic effects are highlighted by red rectangles. Locus zoom plots are presented for these two loci: (B) locus with fine-mapped SNP rs1260326. and (C) locus with fine-mapped SNP rs5742915. SNPs in a $\pm$500 kb window are included, colored by r^2 with the corresponding fine-mapped SNP.

6 Table Legends

Table S1: relative enrichment of functional priors in simulation studies

Table S2: comparison of fine-mapping results based on different hyperparameter settings of K

Table S3: detailed composition of computation time

Table S4: fine-mapping results for five functional biomarkers based on the UK Biobank, including genetic variants with a PIP > 0.1

Table S5: relative enrichment of functional annotations for five functional biomarkers

Table S6: comparison of fine-mapped SNP heritability for five functional biomarkers

7 Acknowledgements

YL is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2019-0621), Fonds de recherche Nature et technologies (FRQNT) New Career (NC-268592), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative New Investigator start-up award (G249591). This study has been conducted using UK Biobank Resources under Application Number 45551. This study was enabled, in part, by support from Calcul Québec and Compute Canada. WZ has been supported by a doctoral training fellowship from the Healthy Brains, Healthy Lives Program, funded by the Canada First Research Excellence Fund (CFREF), Quebec's Ministère de l'Économie et de l'Innovation (MEI), and the Fonds de recherche du Québec (FRQS, FRQSC and FRQNT).

8 Author contributions

W.Z and Y.L have conceived the study and developed the methodology. W.Z created the computational software and ran the analyses. All authors interpreted the results. W.Z. drafted the initial manuscript. H.N and Y.L supervised this study and revised the manuscript critically.

9 Disclosures

The authors declare no conflict of interest.

10 Data and Software Availability

SparsePro is an open-access software and will be made publicly available after publication at `https://github.com/zhwm/SparsePro`. All simulation and plotting scripts to reproduce this study will be made publicly available at `https://github.com/zhwm/SparsePro_Paper`. Individual-level phenotype and genotype data from the UK Biobank are available upon successful application to its research committee. GCTA were downloaded from `https://cnsgenomics.com/software/gcta/bin/gcta_1.93.2beta.zip`. FINEAMP were downloaded from `http://www.christianbenner.com/finemap_v1.4_x86_64.tgz`. SuSiE (version 0.11.42) were installed from CRAN. PolyFun were installed from `https://github.com/omerwe/polyfun`. UK biobank LD information was downloaded from `https://alkesgroup.broadinstitute.org/UKBB_LD/`. Tissue-specific eQTL were obtained from `https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL_EUR.tar`.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).
2. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature genetics* **50,** 1593–1599 (2018).
3. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature genetics* **50,** 906–908 (2018).
4. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101,** 5–22 (2017).

5. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19,** 491–504 (2018).

6. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187,** 367–383 (2011).

7. Benner, C. *et al.* Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics* **101,** 539–551 (2017).

8. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Human molecular genetics* **24,** R111–R119 (2015).

9. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics* **3,** e114 (2007).

10. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198,** 497–508 (2014).

11. Chen, W. *et al.* Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200,** 719–736 (2015).

12. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32,** 1493–1501 (2016).

13. 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).

14. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics* **10,** e1004722 (2014).

15. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic acids research* **44,** e144–e144 (2016).

16. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics* **52,** 1355–1363 (2020).

17. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics* **50,** 621–629 (2018).

18. Titsias, M. & Lázaro-Gredilla, M. Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in neural information processing systems* **24,** 2339–2347 (2011).

19. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33,** 248–255 (2017).

20. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association* **112,** 859–877 (2017).

21. Woolf, B. The log likelihood ratio test (the G-test). *Annals of human genetics* **21,** 397–409 (1957).

22. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99,** 139–153 (2016).

23. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

24. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45,** 124–130 (2013).

25. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38,** e164–e164 (2010).

26. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).

27. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011).

28. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51,** 1749–1755 (2019).

29. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research* **41,** 827–841 (2013).

30. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).

31. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44,** 369–375 (2012).

32. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45,** 580–585 (2013).

33. Consortium, G. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (2020).

34. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData mining* **6,** 1–12 (2013).

35. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–2337 (2010).

SparsePro Supplementary Information

Wenmin Zhang[1], **Hamed Najafabadi**[1,2,3], **Yue Li**[1,4,*]

[1]Quantitative Life Science, McGill University, Montreal, Canada

[2]Department of Human Genetics, McGill University, Montreal, Canada

[3]McGill University and Genome Quebec Innovation Centre, Montreal, Canada

[4]School of Computer Science, McGill University

*Correspondence: `yueli@cs.mcgill.ca`

1 Supplementary Notes

1.1 SparsePro is not sensitive to hyperparameter K

The number of causal effect K is an important hyperparameter in statistical fine-mapping. In methods that exhaustively searching through causal configurations, the computation time increases exponentially with K since the number of candidate causal configurations also grows exponentially. In contrast, in SparsePro, the computation time increases linearly with K. In practice, most of the computation time has been spent on loading LD information, thus the computation time varies only slightly with $\mathrm{K} \in \{5, 7, 9, 11\}$. The output of SparsePro is not sensitive to the choice of K as long as K is greater than or equal to the actual number of causal effects. In our simulation studies, we found that with $\mathrm{K} = 7$, 9, or 11, the resulting PIPs were extremely highly correlated with those based on $\mathrm{K} = 5$, and the overal AUPRC metrics were also highly consistent (Supplementary Table S2).

1.2 Modified HESS estimates for hyperparameters τ_{y} and τ_{β}

Local heritability estimates are useful in setting hyperparameters for SparsePro. Shi et al. [22] provided an unbiased estimator for local heritability estimation based on summary statistics:

$$\hat{\mathrm{h}}_{\mathrm{g}} = \frac{\mathrm{N}\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathrm{R}^{-1}\hat{\boldsymbol{\beta}} - \mathrm{P}}{\mathrm{N} - \mathrm{P}}$$

where R is the LD matrix, $\hat{\boldsymbol{\beta}}$ is GWAS summary effect size, N is the sample size in the GWAS and P is the number of SNPs considered in a locus. However, this estimate requires that when generating summary statistics, both genotypes and phenotypes should be standardized to have zero mean and unit variance. Since summary statistics generated by some GWAS pipelines do not specifically standardize the genotypes and phenotypes, we modified the HESS estimator to account for the non-unit variance:

$$\hat{\mathrm{h}}_{\mathrm{g}} = \frac{(\hat{\boldsymbol{\beta}} \circ \mathbf{v})^{\mathrm{T}}(\mathrm{X}^{\top}\mathrm{X})^{-1}(\hat{\boldsymbol{\beta}} \circ \mathbf{v}) - \mathrm{var}(\mathrm{y})\mathrm{P}}{\mathrm{var}(\mathrm{y})(\mathrm{N} - \mathrm{P})}$$

where $\circ$ represents element-wise multiplication and v is a $\mathrm{P} \times 1$ vector: $\mathrm{v}_{\mathrm{p}} = \mathbf{X}_{\mathbf{p}}^{\top}\mathbf{X}_{\mathbf{p}}$ for the p-th SNP with genotype vector $\mathbf{X}_{\mathbf{p}}$. This estimate can be adapted to directly operate on summary statistics as explained in Methods.

1.3 Full derivation of variational EM algorithm:

As has been described in Methods, based on the data generative process, for the k-th causal effect, we have:

$$\mathbf{s}_{\mathrm{k}} \sim \mathrm{Multinomial}(1, \tilde{\boldsymbol{\pi}})$$

$$\beta_{\mathrm{k}} \sim \mathrm{Normal}(0, \tau_{\beta_{\mathrm{k}}}^{-1})$$

$$\mathbf{y} = \mathrm{X}\sum_{\mathrm{k}} \mathbf{s}_{\mathrm{k}}\beta_{\mathrm{k}} + \boldsymbol{\epsilon}$$

with $\epsilon_{\mathrm{i}} \sim \mathrm{N}(0, \tau_{\mathrm{y}}^{-1})$. Therefore, we have the joint probability

$$\mathrm{p}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X}, \tilde{\boldsymbol{\pi}}, \tau_{\beta}, \tau_{\mathrm{y}}) = \mathrm{p}(\mathbf{y}|\mathrm{X}, \mathrm{S}, \boldsymbol{\beta}, \tau_{\mathrm{y}}) \prod_{\mathrm{k}} \mathrm{p}(\beta_{\mathrm{k}}|\tau_{\beta_{\mathrm{k}}}) \prod_{\mathrm{k}} \mathrm{p}(\mathbf{s}_{\mathrm{k}}|\tilde{\boldsymbol{\pi}}) \tag{4}$$

The goal of fine-mapping is to infer the posterior probability, and in particular, of the sparse projection S (from here we make the dependency on hyperparameters implicit for the ease of notation):

$$\mathrm{p}(\mathrm{S}, \boldsymbol{\beta}|\mathbf{y}, \mathrm{X}) = \frac{\mathrm{p}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X})}{\mathrm{p}(\mathbf{y}|\mathrm{X})}$$

We use a paired mean field factorized [18] variational family $\mathrm{q}(\mathrm{S}, \boldsymbol{\beta})$ to approximate the posterior:

$$\mathrm{q}(\mathrm{S}, \boldsymbol{\beta}) = \prod_{\mathrm{k}} \mathrm{q}(\mathbf{s}_{\mathrm{k}}, \beta_{\mathrm{k}}) = \prod_{\mathrm{k}} \mathrm{q}(\mathbf{s}_{\mathrm{k}}) \mathrm{q}(\beta_{\mathrm{k}}|\mathbf{s}_{\mathrm{k}})$$

Note that in this variational family, we do not specify the form of the distribution; rather, we only specify the dependency of β_{k} on $\mathbf{s}_{\mathrm{k}}$ and that all K causal effects are independent of each other. Also, the form of the variational family does not depend on any observed data.

To better approximate the posterior distribution with members of the variational family, we aim to minimize the KL divergence between the posterior distribution and the proposed variational distribution, which is equivalent to maximizing the ELBO [20]:

$$\mathrm{ELBO} = \mathrm{E}_{\mathrm{q}}[\mathrm{logp}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X})] - \mathrm{E}_{\mathrm{q}}[\mathrm{logq}(\mathrm{S}, \boldsymbol{\beta})]$$

To maximize the above ELBO, the following requirement should be satisfied for each k:

$$\mathrm{logq}(\mathbf{s}_{\mathrm{k}}, \beta_{\mathrm{k}}) = \mathrm{E}_{\mathrm{q}(\mathrm{k})}[\mathrm{logp}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X})]$$

where $\mathrm{E}_{\mathrm{q}(\mathrm{k})}$ denotes taking expectation with respect to the variational distribution excluding the k-th component. With the joint probability provided in Equation (4) we have

$$
\begin{aligned}
\mathrm{logp}(\mathbf{y}, \mathrm{S}, \boldsymbol{\beta}|\mathrm{X}) &= \mathrm{logp}(\mathbf{y}|\mathrm{X}, \mathrm{S}, \boldsymbol{\beta}) + \sum_{\mathrm{k}} \mathrm{logp}(\beta_{\mathrm{k}}|\tau_{\beta_{\mathrm{k}}}) + \sum_{\mathrm{k}} \mathrm{logp}(\mathbf{s}_{\mathrm{k}}|\tilde{\boldsymbol{\pi}}) \\
&= \frac{\mathrm{N}}{2}\log\frac{\tau_{\mathrm{y}}}{2\pi} - \frac{\tau_{\mathrm{y}}}{2}(\mathbf{y} - \mathrm{X}(\sum_{\mathrm{k}} \mathbf{s}_{\mathrm{k}}\beta_{\mathrm{k}}))^{\top}(\mathbf{y} - \mathrm{X}(\sum_{\mathrm{k}} \mathbf{s}_{\mathrm{k}}\beta_{\mathrm{k}})) \\
&+ \sum_{\mathrm{k}}(\frac{1}{2}\log\frac{\tau_{\beta_{\mathrm{k}}}}{2\pi} - \frac{\tau_{\beta_{\mathrm{k}}}}{2}\beta_{\mathrm{k}}^2) + \sum_{\mathrm{k}}\sum_{\mathrm{g}} \mathrm{s}_{\mathrm{kg}}\log\tilde{\pi}_{\mathrm{g}}
\end{aligned}
$$

If we denote the complete set of SNPs excluding the g-th SNP as $\mathrm{g}\prime$ and $\tilde{\boldsymbol{\beta}}_{(\mathrm{k})} = \mathrm{E}_{\mathrm{q}(\mathrm{k})}[\sum_{\mathrm{k}\prime\neq\mathrm{k}} \mathbf{s}_{\mathrm{k}}\beta_{\mathrm{k}}] = \sum_{\mathrm{k}'\neq\mathrm{k}} \boldsymbol{\gamma}^*_{\mathrm{k}'} \circ \boldsymbol{\mu}^*_{\mathrm{k}'}$, by taking expectation with respect to the variational distribution excluding the k-th component and plugging in $\mathrm{s}_{\mathrm{kg}} = 1$ and $\mathrm{s}_{\mathrm{kg}'} = 0$ for all SNPs excluding the g-th SNP, we can derive the following joint distribution of the k-th effect:

$$
\mathrm{logq}(\mathrm{s}_{\mathrm{kg}} = 1, \mathrm{s}_{\mathrm{kg}\prime} = 0, \beta_{\mathrm{k}}) = \mathrm{const} - \frac{\tau_{\beta_{\mathrm{k}}}}{2}\beta_{\mathrm{k}}^2 - \frac{\tau_{\mathrm{y}}}{2}\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}}\beta_{\mathrm{k}}^2 + \tau_{\mathrm{y}}\beta_{\mathrm{k}}\mathbf{X}_{\mathrm{g}}^{\top}(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}_{(\mathrm{k})}) + \log\tilde{\pi}_{\mathrm{g}} \tag{5}
$$

which takes the form of a normal distribution. Therefore, we recognize that

$$
\mathrm{q}(\beta_{\mathrm{k}}|\mathrm{s}_{\mathrm{kg}=1}, \mathrm{s}_{\mathrm{kg}\prime} = 0) \sim \mathrm{N}(\mu^*_{\mathrm{kg}}, \tau^*_{\mathrm{kg}})
$$

By matching sufficient statistics for this normal distribution, we can obtain the following variational parameters for updates:

$$
\tau^*_{\mathrm{kg}} = \tau_{\mathrm{y}}\mathbf{X}_{\mathrm{g}}^{\top}\mathbf{X}_{\mathrm{g}} + \tau_{\beta_{\mathrm{k}}}
$$

$$
\mu^*_{\mathrm{kg}} = \frac{\tau_{\mathrm{y}}}{\tau^*_{\mathrm{kg}}}\mathbf{X}_{\mathrm{g}}^{\top}(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}_{(\mathrm{k})})
$$

By integrating out β_{k} in Equation (5), we obtain that

$$
\mathrm{logq}(\mathrm{s}_{\mathrm{kg}} = 1, \mathrm{s}_{\mathrm{kg}\prime=0}) = \log\tilde{\pi}_{\mathrm{g}} - \frac{1}{2}\log\frac{\tau^*_{\mathrm{kg}}}{2\pi} + \frac{1}{2}\tau^*_{\mathrm{kg}}\mu^{*2}_{\mathrm{kg}} + \mathrm{const}
$$

hence the posterior probability of the g-th SNP being causal in the k-th effect can be estimated as:

$$\gamma^*_{\mathrm{kg}} := \mathrm{q}(\mathrm{s}_{\mathrm{kg}} = 1, \mathrm{s}_{\mathrm{kg}'} = 0) = \mathrm{softmax}(\log\tilde{\pi}_{\mathrm{g}} - \frac{1}{2}\log\frac{\tau^*_{\mathrm{kg}}}{2\pi} + \frac{1}{2}\tau^*_{\mathrm{kg}}\mu^{*2}_{\mathrm{kg}})$$

This completes the variational expectation step in our inference algorithm. When functional annotations are available, we use the following maximization step to integrate relevant annotation.

After the expectation step, we have that

$$\begin{aligned}
\mathrm{ELBO} &= \mathrm{const} + \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} \log\tilde{\pi}_{\mathrm{g}} \\
&= \mathrm{const} + \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} \log\frac{\exp(\mathbf{A}_{\mathrm{g}}\mathbf{w})}{\sum_{\mathrm{g}}\exp(\mathbf{A}_{\mathrm{g}}\mathbf{w})} \\
&= \mathrm{const} + \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} [\mathbf{A}_{\mathrm{g}}\mathbf{w} - \log(\sum_{\mathrm{g}}\exp(\mathbf{A}_{\mathrm{g}}\mathbf{w}))]
\end{aligned}$$

To maximize ELBO with respect to the relative enrichment of the m-th candidate annotation, we can take derivatives of ELBO with respect to w_{m} and set it to 0 to solve for w_{m}:

$$
\begin{aligned}
\frac{\partial \mathrm{ELBO}}{\partial \mathrm{w_m}} &= \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} [\mathrm{A_{gm}} - \frac{\sum_{\mathrm{g}} \mathrm{A_{gm}} \exp(\mathbf{A}_{\mathrm{g}} \mathbf{w}))}{\sum_{\mathrm{g}} \exp(\mathbf{A}_{\mathrm{g}} \mathbf{w}))}] \\
&= \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} [\mathrm{A_{gm}} - \frac{\sum_{\mathrm{g}} \mathrm{A_{gm}} \exp(\mathrm{A_{gm}} \mathrm{w_m}) \exp(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}}))}{\sum_{\mathrm{g}} \exp(\mathrm{A_{gm}} \mathrm{w_m}) \exp(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}}))}] \\
&= \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{k,g}} [\mathrm{A_{gm}} - \frac{\sum_{\mathrm{g}} \mathrm{A_{gm}} \exp(\mathrm{A_{gm}} \mathrm{w_m}) \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})}{\sum_{\mathrm{g}} \exp(\mathrm{A_{gm}} \mathrm{w_m}) \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})}] \\
&= \sum_{\mathrm{k,g}} [\mathrm{A_{gm}} = 1] \gamma^*_{\mathrm{kg}} \\
&- \sum_{\mathrm{k,g}} \gamma^*_{\mathrm{kg}} \frac{\mathrm{e^{w_m}} \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 1] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})}{\mathrm{e^{w_m}} \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 1] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}}) + \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 0] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})} \\
&= \mathrm{r_1} - (\mathrm{r_1} + \mathrm{r_0}) \frac{\mathrm{k_1} \mathrm{e^{w_m}}}{\mathrm{k_1} \mathrm{e^{w_m}} + \mathrm{k_0}} \\
&= 0
\end{aligned}
$$

with

$$
\mathrm{k_1} = \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 1] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})
$$

$$
\mathrm{k_0} = \sum_{\mathrm{g}} [\mathrm{A_{gm}} = 0] \mathrm{softmax}(\sum_{\mathrm{m'} \neq \mathrm{m}} \mathrm{A_{gm'}} \mathrm{w_{m'}})
$$

$$
\mathrm{r_1} = \sum_{\mathrm{k,g}} [\mathrm{A_{gm}} = 1] \gamma^*_{\mathrm{kg}}
$$

$$
\mathrm{r_0} = \sum_{\mathrm{k,g}} [\mathrm{A_{gm}} = 0] \gamma^*_{\mathrm{kg}}
$$

We can obtain

$$
\frac{\mathrm{k_1} \mathrm{e^{w_m}}}{\mathrm{k_1} \mathrm{e^{w_m}} + \mathrm{k_0}} = \frac{\mathrm{r_1}}{\mathrm{r_1} + \mathrm{r_0}}
$$

i.e.

$$\mathrm{w_m} = \log(\frac{\mathrm{r_1}/\mathrm{r_0}}{\mathrm{k_1}/\mathrm{k_0}})$$

Notably, this estimate is analogous to a relative risk estimate in a 2×2 contigency table. Suppose we consider one annotation, then $\mathrm{k_1}$ corresponds to the number of variants with this specific annotation while $\mathrm{k_0}$ corresponds to the number of variants without the annotation. Meanwhile, $\mathrm{r_0}$ corresponds to the sum of posterior probability for variants with the annotation while $\mathrm{r_1}$ corresponds to the sum of posterior probability for variants without the annotation.

Similarly, the standard error of this estimate can be calculated based on the standard error of a relative risk:

$$\mathrm{se}(\hat{\mathrm{w}}_\mathrm{m}) = \sqrt{\frac{1}{\mathrm{r_1}} + \frac{1}{\mathrm{r_0}} - \frac{1}{\mathrm{k_1}} - \frac{1}{\mathrm{k_0}}}$$

We can also evaluate the statistical significance of enrichment with the log likelihood ratio test (G-test) [21].