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 Sciences, McGill University, Montreal, Canada

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

[3]McGill Genome 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 data, we enable a linear search of causal variants instead of a combinatorial search of causal configurations used in most

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 show that, compared to other methods, the causal variants identified by SparsePro are highly enriched for expression quantitative trait loci and explain a larger proportion of trait heritability. We also 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. SparsePro software is available on GitHub at `https://github.com/zhwm/SparsePro`.

1 Introduction

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], mainly 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 combinatorially as the number of causal variants increases, thus tremendously increasing the computational cost if multiple causal variants exist. SuSiE [13] introduces an iterative 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] adopts stratified LD score regression [17] to effectively partition total trait heritability into annotation-dependent heritability estimates, and uses these estimates of annotation-tagged heritability to specify functional priors for fine-mapping methods.

In this work, we present a unified probabilistic framework called *Sparse Projections to Causal Effects* (SparsePro) for statistical fine-mapping with the capacity to incorporate functional annotations. Accompanied with an efficient variational expectation-maximization inference algorithm [18], SparsePro achieves superior accuracy in identifying causal variants as well as computational efficiency compared to the state-of-the-art approaches in both simulation studies and real data analyses. We further demonstrate the utility of SparsePro in genome-wide fine-mapping of functional biomarkers for five vital organs in human.

2 Materials and Methods

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 as-

sess their effects jointly. Then, we infer the contribution of each SNP towards each group of causal effect separately to obtain posterior inclusion probabilities (PIPs). Second, optionally, if we have knowledge about any functional annotations which may be enriched for the causal SNPs, we can estimate the relative enrichment of these annotations, and re-prioritize SNPs according to the enrichments of these annotations. As outputs, our model yields functionally informed PIP for each SNP and the enrichment estimates of candidate functional annotations.

2.2 Our contributions in the context of the existing methods

Our work is related to two existing methods, SuSiE [13] and PolyFun [16]. Inspired by the "sum of single effects" model in SuSiE, we introduce 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 exhaustively evaluating the combinatorial number of causal configurations. For statistical inference, SuSiE adopts an iterative Bayesian stepwise selection algorithm that operates on the Bayes Factors (BFs) [13]. Here, we use a paired mean field variational inference algorithm [18] to jointly update the variational parameters for the causal effects and causal indicators of each SNP. Moreover, we have adapted our algorithm to directly work with GWAS summary statistics and provided appropriate estimates for the hyperparamters including trait variance and heritability estimates. To enable functionally informed fine-mapping, PolyFun uses genome-wide heritability estimates from LD score regression to set the functional priors for fine-mapping methods [16]. In contrast, we aggregate the genome-wide statistical fine-mapping evidence by maximizing the evidence lower bound of SparsePro to prioritize relevant annotations and robustly derive genome-wide functional priors.

2.3 SparsePro model specification

We assume the following data generative process (**Figure** 1) for a continuous polygenic trait. First, the prior probability $\tilde{\pi}_g$ for the g-th SNP ($g \in \{1, \ldots, G\}$) being causal is defined as:

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

where $\mathbf{A}_g$ is the $1 \times M$ annotation row vector of M candidate annotations for the g-th SNP; and $\mathbf{w}$ is the $M \times 1$ 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. $\tilde{\pi}_g = \frac{1}{G}$.

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

$$\mathbf{s}_k \sim Multinomial(1, \tilde{\boldsymbol{\pi}})$$

where $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, ..., \tilde{\pi}_G)$ and $\mathbf{s}_k$ is a binary indicator vector of length G indicating which SNP is the causal SNP under the k-th ($k \in \{1, ..., K\}$) causal effect.

Then, the causal effect sizes are sampled from a normal distribution, i.e.

$$\beta_k \sim \mathcal{N}(0, \tau_{\beta_k}^{-1})$$

Finally, the continuous trait $\mathbf{y}_{N\times1}$ over N individuals is generated as follows:

$$\mathbf{y} = \mathbf{X}\sum_k \mathbf{s}_k\beta_k + \boldsymbol{\epsilon}$$

or in matrix form:

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

where $\mathbf{X}_{N\times G}$ is the full genotype matrix, $\mathbf{S}_{G\times K}$ is the sparse projection matrix, $\boldsymbol{\beta}_{K\times1}$ is the causal effect vector, and $\boldsymbol{\epsilon}_{N\times1} \sim \mathcal{N}(0, \tau_y^{-1}\mathbf{I}_N)$ denotes the variance not attributable to the modelled ge-

netic effects.

2.4 A variational inference algorithm for Bayesian fine-mapping

With this model specification (**Figure** 1), finding the causal variants is equivalent to inferring the sparse projections $\mathbf{s}_k$ and the effect sizes β_k given $\mathbf{y}$ and $\mathbf{X}$ for $k \in \{1, ..., K\}$:

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

As the number of possible causal configurations grows combinatorial with G, the exact posterior solution is intractable because of the marginal likelihood in the denominator. Unlike most existing 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]:

$$q(\mathbf{S}, \boldsymbol{\beta}) = \prod_k q(\mathbf{s}_k, \beta_k) = \prod_k q(\mathbf{s}_k) q(\beta_k|\mathbf{s}_k)$$

This variational distribution preserves the dependency between $\mathbf{s}_k$ and β_k. 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].

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]:

$$ELBO = E_q[\log p(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X}, \tilde{\boldsymbol{\pi}}, \tau_\beta, \tau_y)] - E_q[\log q(\mathbf{S}, \boldsymbol{\beta})]$$

Based on the mean field assumptions [18], this optimization can be conducted iteratively for the k^{th} causal effect and the g^{th} SNP with the following closed-form updates until convergence (derivation details are available in **Supplementary Notes**).

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

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

with

$$\tau_{kg}^* = \mathbf{X}_g^\top \mathbf{X}_g \tau_y + \tau_{\beta_k} \tag{2}$$

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

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

$$\gamma_{kg}^* = softmax(\log \tilde{\pi}_g - \frac{1}{2} \log \frac{\tau_{kg}^*}{2\pi} + \frac{\tau_{kg}^* \mu_{kg}^{*2}}{2}) \tag{3}$$

For fine-mapping, we take the maximum of these K probabilities as the PIP for SNP g: $\gamma_g^* = \max(\gamma_{1g}^*, \ldots, \gamma_{Kg}^*)$.

2.5 Adaptation to GWAS summary statistics

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

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

1. If the genotypes are standardized, we have

$$\mathbf{X}_g^\top \mathbf{X}_g = N$$

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

$$\mathbf{X}_g^\top \mathbf{y} = N\hat{\beta}_g$$

where N is the sample size.

2. If the genotypes are not standardized, we have

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

$$se_g = \sqrt{var(\mathbf{y})(\mathbf{X}_g^\top \mathbf{X}_g)^{-1}}$$

Therefore,

$$\mathbf{X}_g^\top \mathbf{X}_g = \frac{var(\mathbf{y})}{(se_g^2)}$$

$$\mathbf{X}^\top \mathbf{X} = LD * (\mathbf{s}\mathbf{e}^\top \mathbf{s}\mathbf{e})$$

$$\mathbf{X}_g^\top \mathbf{y} = \mathbf{X}_g^\top \mathbf{X}_g * \hat{\beta}_g$$

Notably, if $\mathbf{y}$ has been standardized to have unit variance prior to a GWAS, we naturally supply $var(\mathbf{y}) = 1$. Otherwise, it can be estimated as $var(\mathbf{y}) = 2Np(1-p)se^2$ 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.

2.6 Variational expectation-maximization for integrating functional annotations

To estimate the relative enrichment of functional annotations and further prioritize variants, we adopt a 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 $\tilde{\pi}_g$ (prior probability of being causal) for each SNP.

Suppose we have M candidate annotations and A_{gm} ($m \in \{1, ..., M\}$) 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 w_m to 0 and solving for w_m, we have the following estimate for the logarithm of rele-

vant enrichment (detailed in **Supplementary Notes**),

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

where

$$k_1 = \sum_g [A_{gm} = 1] softmax(\sum_{m' \neq m} A_{gm'} w_{m'})$$
$$k_0 = \sum_g [A_{gm} = 0] softmax(\sum_{m' \neq m} A_{gm'} w_{m'})$$
$$r_1 = \sum_{k,g} [A_{gm} = 1] \gamma^*_{kg}$$
$$r_0 = \sum_{k,g} [A_{gm} = 0] \gamma^*_{kg}$$

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

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

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,

$$\tilde{\pi}_g = softmax(\sum_m A_{gm} \hat{w}_m)$$

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

2.7 Hyperparameter settings

We have three hyperparameters: number of causal effect K, inverse of the unexplained variance τ_y and inverse variance of causal effect sizes τ_{β_k} in our model. As we show in **Supplementary Notes**, 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

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

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

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

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

2.8 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 ancestry 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 that 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 imitate a GWAS. We ob-

tained 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 five annotations without enrichment: "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.

2.9 Method comparisons using simulated data

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 each COJO-identified lead SNP. 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 optimal performance of PolyFun [16]. The estimated priors were provided to SuSiE via "prior_weights" and to FINEMAP via the

`--prior-snps` option, 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.

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

To investigate the genetic 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 z-scores that had 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 the simulation analyses (**Section** 2.9), except that the number of causal effects was set to 9 for each LD region to account for potentially 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 identified 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 calculated trait heritability conferred by fine-mapped SNPs with 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 z-scores (i.e. inverse normal transformed trait residuals after regressing out covariate effects) 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 the 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].

3 Results

3.1 SparsePro demonstrates superior performance in simulation

We performed simulations based on real genotype data from UK Biobank (**Materials and Methods**). We observed that 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 precision at the same recall rates (**Figure** 2A). For example, at the recall rate of 25%, SparsePro achieved greater than 95% precision, which is highly desirable in fine-mapping because only a small number of the prioritized SNPs will be experimented validated *in vivo* or *in vitro* in practice (**Figure** 2A).

Moreover, SparsePro can incorporate functional priors (**Supplementary Table** S1) with improved fine-mapping 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. As expected, 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. It has been shown that for a well-calibrated fine-mapping method, the mean PIP of all SNPs with a PIP above a certain threshold should be equal to the precision if these SNPs were to be considered causal variants [16]. Here, we found that the mean PIP of all SNPs considered to be causal variants by SparsePro was almost identical to the desired precision at any threshold (**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 as causal variants, SparsePro- and SparsePro+ would both have a median precision across simulations of 95% and 100% respectively (**Figure** 2C). The selected SNPs by FINEMAP (median precision = 50% only)

and SuSiE (median precision = 71% only) included an excessive proportion of false positives, even with functional priors (median precision = 77% for FINEMAP and 79% for SuSiE; **Figure** 2C). The high precision by SparsePro was consistent for all frequently used PIP thresholds (**Figure** 2C) although FINEMAP and SuSiE sometimes have a slightly higher recall.

Furthermore, SparsePro conferred not only higher fine-mapping precision, but also higher computational efficiency. In our simulation, it took only an hour to fine-map chromosome 22, which was 6.5 times faster than FINEMAP and 3 times faster than SuSiE (**Figure** 2D and **Supplementary Table** S3).

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

We performed GWAS in the UK Biobank [1] for five functional biomarkers: FEV1-FVC ratio (lung function), estimated glomerular filtration rate (kidney function), pulse rate (heart function), total protein (liver function) and blood glucose level (pancreatic islet function). Genome-wide fine-mapping of five functional biomarkers based on the UK Biobank population using SparsePro identified multiple potentially causal variants (**Supplementary Table** S4). To assess biological relevance of the fine-mapping results, we estimated the relative enrichment of causal signals in tissue-specific eQTL for each trait (**Materials and Methods**). 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, the fine-mapped SNPs by SparsePro+ exhibited the highest biological relevance, being 4.06-fold (95% CI: 3.31-4.97) more likely to be liver-specific eQTL.

Moreover, at most PIP thresholds, the SNPs fine-mapped by SparsePro- explained a higher proportion of phenotypic variance based on all UK Biobank subjects (**Methods**) compared to

the same number of the most likely causal SNPs identified by SuSiE (**Figure** 3B and **Supplementary Table** S5). With the functional annotations (**Supplementary Table** S5), 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 compared to the PolyFun-informed SuSiE; although for glucose and pulse rate, PolyFun-informed SuSiE was able to identify SNPs with a slightly higher predictive performance at certain PIP thresholds (**Figure** 3C and **Supplementary Table** S6).

3.3 Pleiotropic effects of SNPs rs1260326 and rs5742915 on the functions of multiple vital organs

Overall, we observed considerable polygenicity for the five biomarkers of the vital organs (**Figure** 4A). Interestingly, at the PIP threshold of 0.80, we found two potentially causal variants for three of the five biomarkers. Specifically, SNP rs1260326 (**Figure** 4B), a missense variant (Leu446Pro) in gene *GCKR*, 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). Notably, this specific variant has been found to be significantly associated with a wide variety of glycemic traits [36] and other quantitative traits for metabolic syndromes and comorbidities [37, 38], and has been implicated in the functions of liver and other vital organs [39–41].

Another SNP, rs5742915 (**Figure** 4C), a missense variant (Phe645Leu) in gene *PML* was fine-mapped for FEV1-FVC ratio (PIP = 0.858), pulse rate (PIP = 1.000) and total protein level (PIP = 0.987). This variant has also been associated with other quantitative biomarkers of polygenic traits featuring development and metabolism, including birth weight [42], height [43], appendicular lean mass [44], and age at menarche [45]. These findings, along with other SNPs exhibiting pleiotropic effects at somewhat lower PIP thresholds (**Supplementary Table** S4) presented promising genetic targets for experimental validations in a larger effort towards understanding the mechanisms of genetic coordination among vital organs.

4 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 coordination among vital organs.

Compared to the existing methods, SparsePro has three important features. First of all, we use 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 the existing fine-mapping methods, such as FINEMAP [12], and is more than twice as fast as SuSiE [13], which is a similar variable selection framework but implements an iterative Bayesian stepwise selection procedure. The substantially improved computational efficiency enables 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-identified lead SNPs, such a genome-wide fine-mapping requires neither a pre-specified p-value threshold (e.g. $p < 5 \times 10^{-8}$) 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 with little additional computational overhead.

Second, we utilize a paired mean field variational family, where the causal effect and the causal indicator are coupled in the variational distribution. This ensures that our approximation matches closely with the true posterior distribution of the causal variant indicators [18]. As a result, SparsePro yielded better-calibrated PIPs compared to existing fine-mapping methods.

Third, 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 showed improved control of false positives, demonstrated higher precision in identifying causal variants in simulation and obtained stronger enrichment for tissue-specific eQTL in real data application.

Last, we propose and implement a probabilistic model that coherently integrates statistical evidence and functional prior information. The key difference between SparsePro+ and other methods that leverage 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. Based on our results, it seems that this approach distills better prior information from the functional annotations compared to the aforementioned methods.

We note that SparsePro can be 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 proper convergence. While we advocate the public availability of the in-sample LD information along with the GWAS summary statistics, a more robust model is needed to account for mismatched LD information. Second, SparsePro currently supports only binary annotations while compatibility with continuous annotations is also desirable. Last, the current variational expectation-maximization scheme might not accurately estimate the joint enrichment of highly correlated annotations. Performing variable selection beforehand or effectively aggregating enrichment estimates may enable the inclusion of multiple correlated informative annotations, such as cell type-specific annotations to further improve the utility of SparsePro.

In summary, SparsePro is an efficient genome-wide fine-mapping method with the ability of integrate functional annotations. We envision its wide utility in understanding the genetic architecture of complex traits, identifying target genes, and increasing the yield of functional follow-up studies of GWASs.

5 Figure Legends

Figure 1. SparsePro overview. The data generating process of SparsePro is depicted in a plate model with shaded nodes represent observed variables and unshaded nodes represent latent variables. The trait y is generated from K causal effects, where the k-th causal effect size $\beta_k \sim \mathcal{N}(0, \tau_{\beta_k})$. We use a sparse projection $\mathbf{s}_k \sim Multinomial(1, \hat{\boldsymbol{\pi}})$ of genotype to indicate causal variant for the k-th effect. Given the causal effect sizes and sparse indicators of causal variants, the target trait y_i for individual i follows a normal distribution $y_i \sim \mathcal{N}(\mathbf{X}_i \sum_k \mathbf{s}_k \beta_k, \tau_y^{-1})$. To help prioritize variants with functional annotations, we assume the prior probability of being causal $\hat{\pi}_g$ for the g-th variant as $\hat{\pi}_g = softmax(\mathbf{A}_g \mathbf{w})$ where $\mathbf{A}_g$ is a $M \times 1$ functional annotation vector and $\mathbf{w}$ is the $M \times 1$ vector of annotation enrichment coefficients. We adopt an efficient variational inference algorithm to jointly estimate both causal effect sizes and sparse indicators and an expectation-maximization scheme for estimating annotation enrichment coefficients $\mathbf{w}$ as detailed in Section 2.

Figure 2. SparsePro demonstrated improved accuracy and computational efficiency in genome-wide simulation results. (A) Precision-Recall curves. The inset shows the area under the precision recall curve (AUPRC) for each method. (B) Calibration of posterior inclusion probabilities (PIPs). The y-axis is the mean PIPs for all SNPs considered as causal variants, corresponding to the expected precision at different PIP cutoffs. The x-axis represents the actual precision at different PIP cutoffs. The black dashed line indicates an optimal calibration, where the expected precision perfectly matches the observed precision. (C) Precision and recall rates obtained at five frequently used PIP 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. The color legends are displayed at the bottom of the figure.

Figure 3. Biological relevance of fine-mapped SNPs for five biomarkers, each for a distinct vital organ. (A) Relative enrichment of causal signals in tissue-specific eQTL. 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 PIP thresholds. As a surrogate of SNP heritability, the proportion of trait variance explained was obtained from multivariate linear regression adjusted R^2. In this multivariate regression, we regress the inverse normal transformed trait residuals against all fine-mapped SNPs after adjusting for covariate effects. We selected the same number of top-ranked SNPs for each method separately at each PIP threshold (**Materials and Methods**).

Figure 4. Fine-mapping genetic associations for five functional biomarkers of vital organs. (A) Illustration of genome-wide distribution of fine-mapped SNPs on 22 chromosomes. SNPs with a posterior inclusion probability > 0.80 were indicated as colored solid circles. Two loci with potential pleiotropic effects on four and three vital organ biomarkers respectively were highlighted by red dashed rectangles. Locus zoom plots were 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.

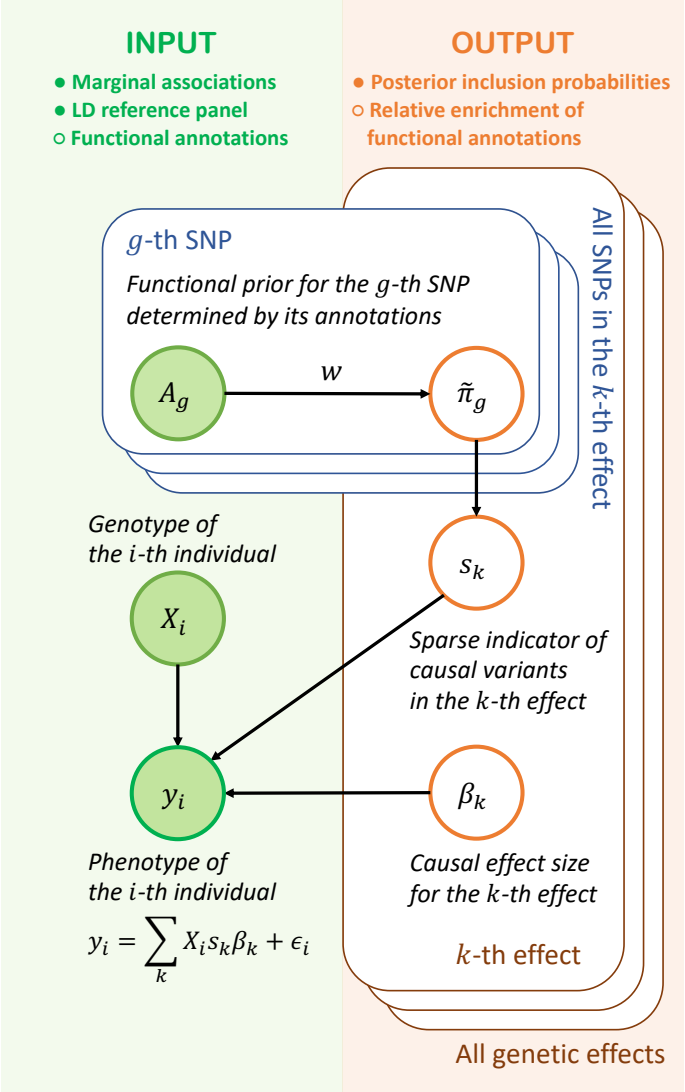

Figure 1: SparsePro overview. The data generating process of SparsePro is depicted in a plate model with shaded nodes represent observed variables and unshaded nodes represent latent variables. The trait y is generated from K causal effects, where the k-th causal effect size $\beta_k \sim \mathcal{N}(0, \tau_{\beta_k})$. We use a sparse projection $\mathbf{s}_k \sim Multinomial(1, \hat{\boldsymbol{\pi}})$ of genotype to indicate causal variant for the k-th effect. Given the causal effect sizes and sparse indicators of causal variants, the target trait y_i for individual i follows a normal distribution $y_i \sim \mathcal{N}(\mathbf{X}_i \sum_k \mathbf{s}_k \beta_k, \tau_y^{-1})$. To help prioritize variants with functional annotations, we assume the prior probability of being causal $\hat{\pi}_g$ for the g-th variant as $\hat{\pi}_g = softmax(\mathbf{A}_g \mathbf{w})$ where $\mathbf{A}_g$ is a $M \times 1$ functional annotation vector and $\mathbf{w}$ is the $M \times 1$ vector of annotation enrichment coefficients. We adopt an efficient variational inference algorithm to jointly estimate both causal effect sizes and sparse indicators and an expectation-maximization scheme for estimating annotation enrichment coefficients $\mathbf{w}$ as detailed in Section 2.

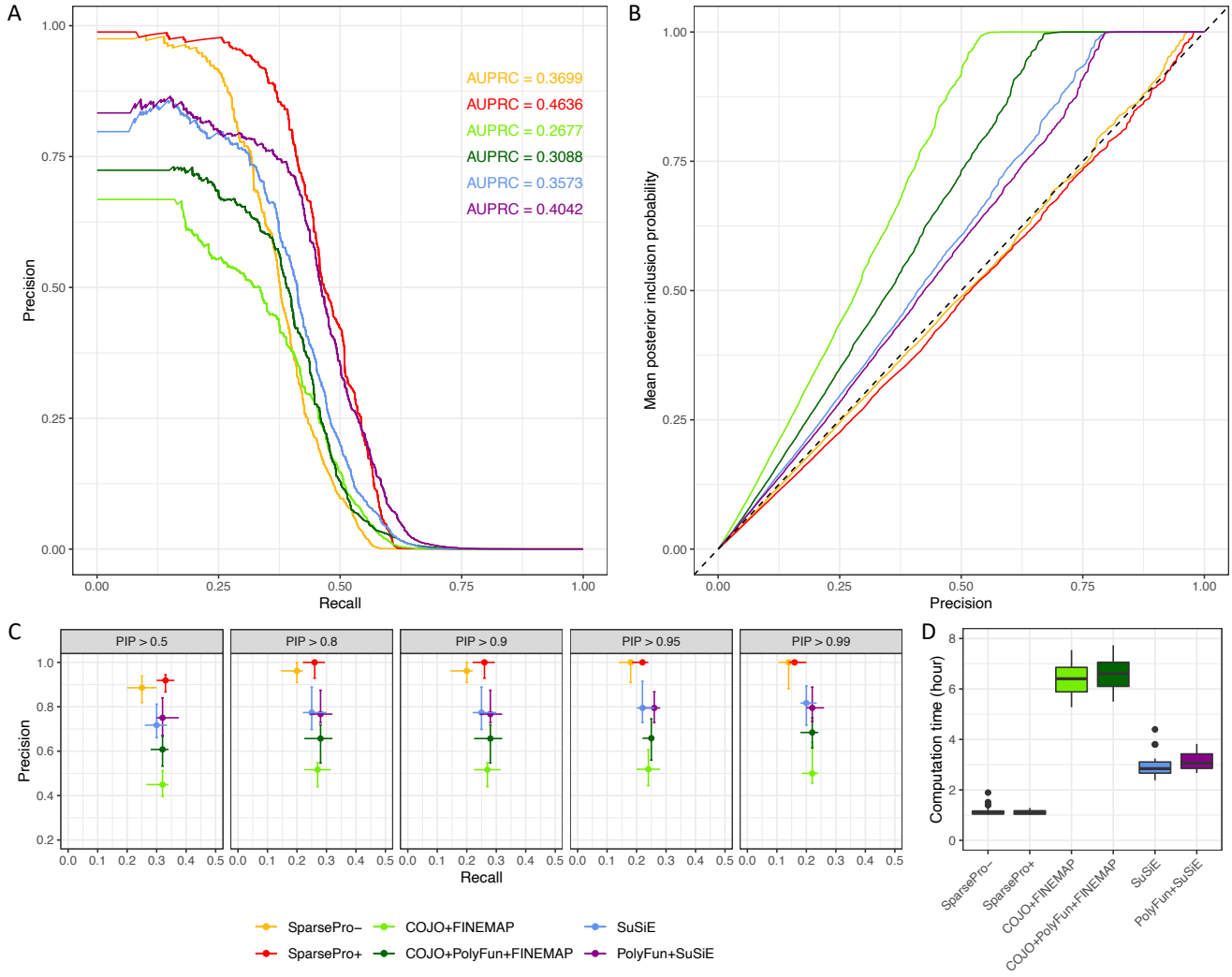

Figure 2: SparsePro demonstrated improved accuracy and computational efficiency in genome-wide simulation results. (A) Precision-Recall curves. The inset shows the area under the precision recall curve (AUPRC) for each method. (B) Calibration of posterior inclusion probabilities (PIPs). The y-axis is the mean PIPs for all SNPs considered as causal variants, corresponding to the expected precision at different PIP cutoffs. The x-axis represents the actual precision at different PIP cutoffs. The black dashed line indicates an optimal calibration, where the expected precision perfectly matches the observed precision. (C) Precision and recall rates obtained at five frequently used PIP 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. The color legends are displayed at the bottom of the figure.

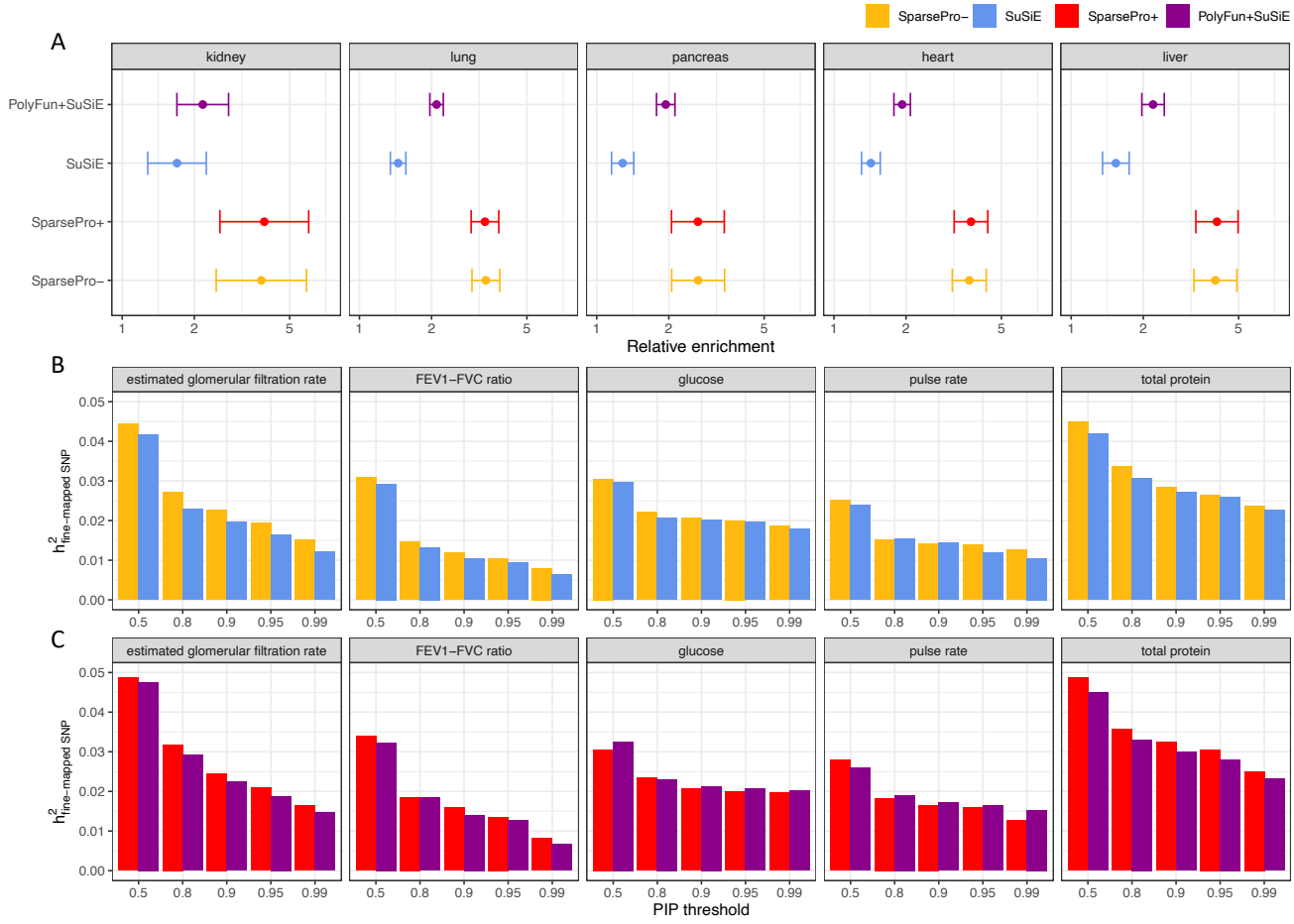

Figure 3: Biological relevance of fine-mapped SNPs for five biomarkers, each for a distinct vital organ. (A) Relative enrichment of causal signals in tissue-specific eQTL. Target traits and the corresponding organs are indicated. Estimates of relative enrichment with 95% confidence intervals are plotted 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 PIP thresholds. As a surrogate of SNP heritability, the proportion of trait variance explained was obtained from multivariate linear regression adjusted R^2. In this multivariate regression, we regress the inverse normal transformed trait residuals against all fine-mapped SNPs after adjusting for covariate effects. We selected the same number of top-ranked SNPs for each method separately at each PIP threshold (**Materials and Methods**).

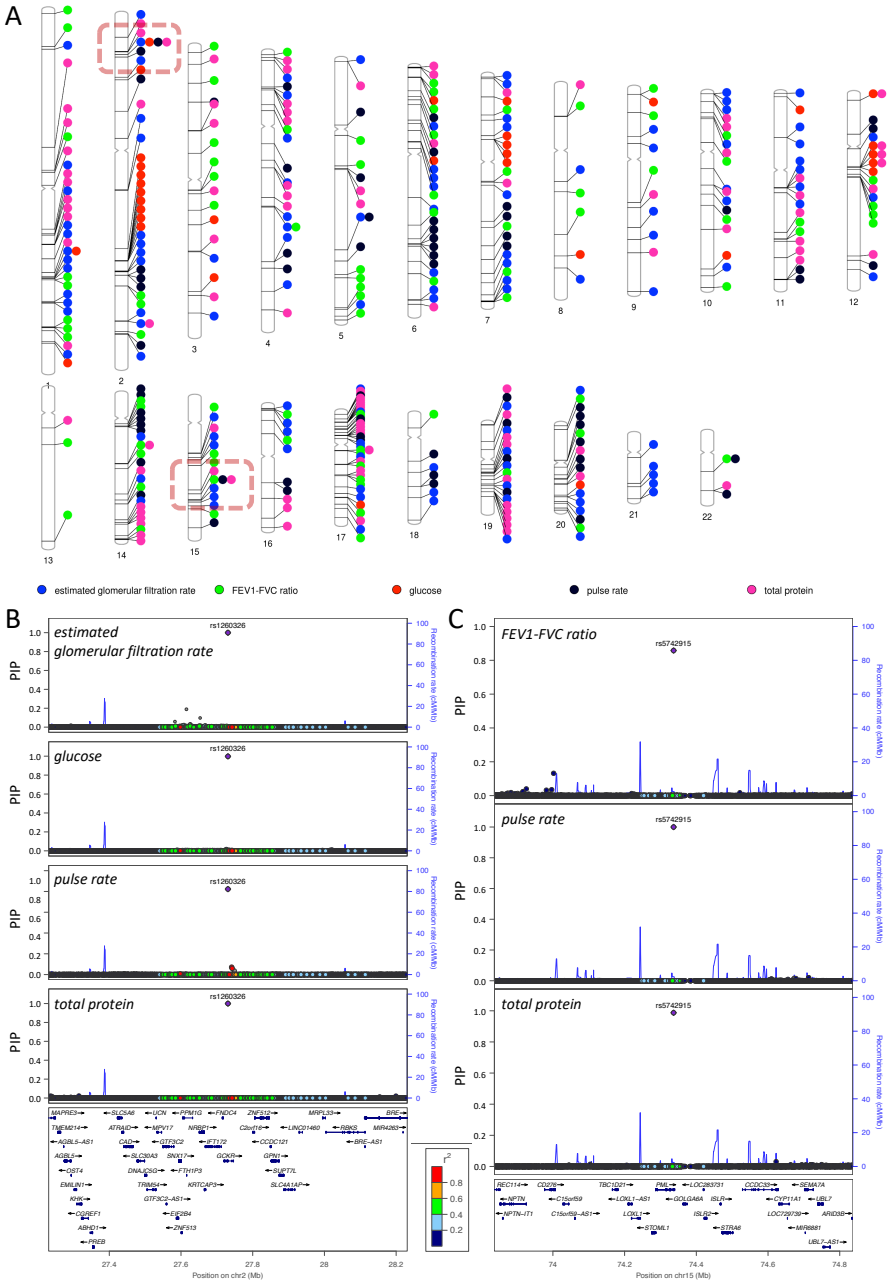

Figure 4: Fine-mapping genetic associations for five functional biomarkers of vital organs. (A) Illustration of genome-wide distribution of fine-mapped SNPs on 22 chromosomes. SNPs with a posterior inclusion probability > 0.80 were indicated as colored solid circles. Two loci with potential pleiotropic effects on four and three vital organ biomarkers respectively were highlighted by red dashed rectangles. Locus zoom plots were 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 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).H.S.N. holds a Canada Research Chair funded by the Canadian Institutes of Health Research.

7 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.S.N and Y.L supervised this study and revised the manuscript critically.

8 Disclosures

The authors declare no conflict of interest.

9 Data and Software Availability

SparsePro is an open-access software and publicly available at `https://github.com/zhwm/SparsePro`. All simulation and plotting scripts to reproduce this study are 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).

36. Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nature genetics* **53,** 840–860 (2021).

37. Huang, L. O. *et al.* Genome-wide discovery of genetic loci that uncouple excess adiposity from its comorbidities. *Nature Metabolism* **3,** 228–243 (2021).

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

39. Chen, V. L. *et al.* Genome-wide association study of serum liver enzymes implicates diverse metabolic and liver pathology. *Nature communications* **12,** 1–13 (2021).

40. Pazoki, R. *et al.* Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nature communications* **12,** 1–12 (2021).

41. Bell, S. *et al.* A genome-wide meta-analysis yields 46 new loci associating with biomarkers of iron homeostasis. *Communications biology* **4,** 1–14 (2021).

42. Warrington, N. M. *et al.* Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nature genetics* **51,** 804–814 (2019).

43. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics* **46,** 1173–1186 (2014).

44. Pei, Y.-F. *et al.* The genetic architecture of appendicular lean mass characterized by association analysis in the UK Biobank study. *Communications biology* **3,** 1–13 (2020).

45. Kichaev, G. *et al.* Leveraging polygenic functional enrichment to improve GWAS power. *The American Journal of Human Genetics* **104,** 65–75 (2019).

SparsePro Supplementary Information

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

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

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

[3]McGill Genome 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 effects K is an important hyperparameter in statistical fine-mapping. In methods that exhaustively search through causal configurations, the computation time increases combinatorially with K since the number of candidate causal configurations also grows combinatorially. In contrast, in SparsePro, the computation time increases linearly with K. In practice, most of the computation time is spent on loading LD information, thus the computation time varies only slightly with $K \in \{5, 7, 9, 11\}$. 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 $K = 7$, 9, or 11, the resulting PIPs were extremely highly correlated with those based on $K = 5$, 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:

$$\hat{h}_g = \frac{N\hat{\boldsymbol{\beta}}^T\mathbf{R}^{-1}\hat{\boldsymbol{\beta}} - P}{N - P}$$

where $\mathbf{R}$ 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:

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

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

1.3 Full derivation of variational EM algorithm:

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

$$\mathbf{s}_k \sim Multinomial(1, \tilde{\boldsymbol{\pi}})$$

$$\beta_k \sim \mathcal{N}(0, \tau_{\beta_k}^{-1})$$

$$\mathbf{y} = X\sum_k \mathbf{s}_k\beta_k + \boldsymbol{\epsilon}$$

with $\epsilon_i \sim N(0, \tau_y^{-1})$. Therefore, we have the joint probability:

$$p(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X}, \tilde{\boldsymbol{\pi}}, \tau_\beta, \tau_y) = p(\mathbf{y}|\mathbf{X}, \mathbf{S}, \boldsymbol{\beta}, \tau_y) \prod_k p(\beta_k|\tau_{\beta_k}) \prod_k p(\mathbf{s}_k|\tilde{\boldsymbol{\pi}}) \tag{4}$$

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

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

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

$$q(\mathbf{S}, \boldsymbol{\beta}) = \prod_k q(\mathbf{s}_k, \beta_k) = \prod_k q(\mathbf{s}_k) q(\beta_k|\mathbf{s}_k)$$

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 $\mathbf{s}_k$ 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]:

$$ELBO = E_{q(\mathbf{S},\boldsymbol{\beta})}[\log p(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X})] - E_{q(\mathbf{S},\boldsymbol{\beta})}[\log q(\mathbf{S}, \boldsymbol{\beta})]$$

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

$$\log q(\mathbf{s}_k, \beta_k) = E_{q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})}[(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X})]$$

where $E_{q(\mathbf{S}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})}$ is the expectation with respect to the variational distribution excluding the k-th

component. With the joint probability provided in Equation (4) we have

$$\begin{aligned}\log p(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X}) &= \log p(\mathbf{y}|\mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) + \sum_k \log p(\beta_k|\tau_{\beta_k}) + \sum_k (\mathbf{s}_k|\tilde{\boldsymbol{\pi}}) \\ &= \frac{N}{2}\log\frac{\tau_y}{2\pi} - \frac{\tau_y}{2}(\mathbf{y} - \mathbf{X}(\sum_k \mathbf{s}_k\beta_k))^\top(\mathbf{y} - \mathbf{X}(\sum_k \mathbf{s}_k\beta_k)) \\ &+ \sum_k(\frac{1}{2}\log\frac{\tau_{\beta_k}}{2\pi} - \frac{\tau_{\beta_k}}{2}\beta_k^2) + \sum_k\sum_g s_{kg}\log\tilde{\pi}_g\end{aligned}$$

Taking expectation with respect to the variational distribution excluding the k-th component and plugging in $s_{kg} = 1$ and $\mathbf{s}_{k\setminus g} = \mathbf{0}$ for all SNPs excluding the g-th SNP, we can obtain the joint distribution of the k-th effect as:

$$\log q(s_{kg} = 1, \mathbf{s}_{k\setminus g} = \mathbf{0}, \beta_k) = const - \frac{\tau_{\beta_k}}{2}\beta_k^2 - \frac{\tau_y}{2}\mathbf{X}_g^\top\mathbf{X}_g\beta_k^2 + \tau_y\beta_k\mathbf{X}_g^\top(\mathbf{y} - X\tilde{\boldsymbol{\beta}}_{\setminus k}) + log\tilde{\pi}_g \quad (5)$$

where

$$\tilde{\boldsymbol{\beta}}_{\setminus k} = E_{q(\mathbf{s}_{\setminus k}, \boldsymbol{\beta}_{\setminus k})}[\sum_{k'\neq k}\mathbf{s}_k\beta_k] = \sum_{k'\neq k}\gamma^*_{k'} \circ \mu^*_{k'}$$

We recognize that

$$q(\beta_k|s_{kg=1}, \mathbf{s}_{k\setminus g} = \mathbf{0}) \sim \mathcal{N}(\mu^*_{kg}, \tau^*_{kg})$$

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

$$\begin{aligned}\tau^*_{kg} &= \tau_y\mathbf{X}_g^\top\mathbf{X}_g + \tau_{\beta_k} \\ \mu^*_{kg} &= \frac{\tau_y}{\tau^*_{kg}}\mathbf{X}_g^\top(\mathbf{y} - X\tilde{\beta}_{\setminus k})\end{aligned}$$

By integrating out β_k in Equation (5), we obtain that

$$\log q(s_{kg} = 1, \mathbf{s}_{k\backslash g} = \mathbf{0}) = \log \tilde{\pi}_g - \frac{1}{2} \log \frac{\tau^*_{kg}}{2\pi} + \frac{1}{2} \tau^*_{kg} \mu^{*2}_{kg} + const$$

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

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

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

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

To maximize ELBO with respect to the relative enrichment of the m-th candidate annotation,

we take partial derivatives of ELBO with respect to w_m and set it to 0 to solve for w_m:

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

where

$$
\begin{aligned}
k_1 &= \sum_g [A_{gm}=1]softmax(\sum_{m'\neq m} A_{gm'}w_{m'}) \\
k_0 &= \sum_g [A_{gm}=0]softmax(\sum_{m'\neq m} A_{gm'}w_{m'}) \\
r_1 &= \sum_{k,g}[A_{gm}=1]\gamma^*_{kg} \\
r_0 &= \sum_{k,g}[A_{gm}=0]\gamma^*_{kg}
\end{aligned}
$$

We then obtain:

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

and solve for:

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

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

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

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

2 Supplementary Table Legends

Supplementary Table S1 Relative enrichment of functional priors in simulation studies.

Supplementary Table S2 Comparison of fine-mapping results based on different hyperparameter settings of the number of causal effects K.

Supplementary Table S3 Details of computation time by each method.

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

589 **Supplementary Table S5** Relative enrichment of functional annotations for five functional
590 biomarkers.

591 **Supplementary Table S6** Comparison of fine-mapped SNP heritability for five functional
592 biomarkers.