SparsePro: an efficient 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 wenmin.zhang@mail.mcgill.ca or yueli@cs.mcgill.ca

Abstract

Identifying causal variants from genome-wide association studies (GWASs) is challenging due to widespread linkage disequilibrium (LD) and possible existence of multiple causal variants in the same genomic locus. Functional annotations of the genome may help to prioritize variants that are biologically relevant and thus improve fine-mapping of GWAS results. Classical fine-mapping methods conducting an exhaustive search of variant-level causal configurations have a high computational cost, especially when the underlying genetic architecture and LD patterns are complex. Here, we propose SparsePro to efficiently conduct fine-mapping integrating statistical evidence and relevant functional annotations. By grouping variants in high LD into effect groups and inferring group-level causal configurations, we

greatly reduced computation cost for fine-mapping. Within each effect group, we additionally used functional annotation information to prioritize the causal variant. With an efficient variational inference algorithm, we jointly infer both the causal status of effect groups as well as variant representations within effect group. We evaluate SparsePro through extensive simulations using resources from the UK Biobank. Compared to state-of-the-art methods, SparsePro achieved more accurate posterior inference with greatly reduced computation time. We demonstrate the utility of SparsePro by investigating the genetic architecture of five clinical biomarkers of vital organs. We show that, compared to other methods, the causal variants identified by SparsePro are highly enriched for tissue-specific annotations and capture a larger proportion of trait heritability. We also identify potential causal variants for multiple clinical biomarkers, and pinpoint target genes with potential pleiotropic effects. In summary, we have developed an accurate and efficient fine-mapping method for integrating statistical evidence and functional annotations. Our method can have wide utility in understanding the genetics of complex traits and 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 hundreds of thousands of associations between genetic variants and traits of interest, understanding the genetic architecture underlying these associations remains challenging [4–6], mainly because GWASs rely on univariate regression models, which are not able to distinguish causal variants from other variants in linkage disequilibrium (LD) [5, 7, 8].

Several fine-mapping methods have been proposed to identify causal variants from GWASs. For instance, BIMBAM [9], CAVIAR [10] and CAVIARBF [11] estimate the posterior inclusion probabilities (PIPs) in a genomic locus by exhaustively evaluating likelihoods of all possible causal configurations. FINEMAP [12] accelerates such calculations with a stochastic shotgun search focusing on the most likely subset of causal configurations. PAINTOR [13] additionally

uses functional annotation data to improve accuracy of fine-mapping and similarly, PolyFun [14] uses genome-wide heritability estimates from LD score regression to set the functional priors for fine-mapping methods. However, the total number of variant-level causal configurations can grow exponentially with the number of causal variants, which can lead to a very high computational cost in classical fine-mapping methods.

A promising approach to reduce the total number of causal configurations to be evaluated is to group correlated variants together and inferring group-level causal configurations. Recently, SuSiE [15] introduces the ”Sum of Single Effects“ model with multiple credible sets where each credible set consists of variants with similar marginal associations on traits. With an iterative Bayesian stepwise selection algorithm, SuSiE achieved great computation efficiency for fine-mapping [16].

In line with the idea of grouping correlated variants together into effect groups, we proposed *Sparse Projections to Causal Effects* (SparsePro) to further improve fine-mapping efficiency and accuracy. First, within each effect group, we additionally incorporate relevant functional annotation information to prioritize the causal variant. Second, we use an efficient variational inference algorithm to further simplify the intuitive algorithm proposed in SuSiE and improve computation efficiency. Third, we adopt heritability-based estimates in SparsePro to improve convergence of algorithm. We assessed the performance of SparsePro in simulation studies in comparison to other fine-mapping methods. We further examined the utility of SparsePro by investigating the genetic architecture of five clinical biomarkers of vital organs.

2 Results

2.1 SparsePro method overview

To fine-map causal variants, SparsePro takes two lines of evidence (**Figure** 1). First, with GWAS summary statistics and matched LD information, we jointly infer both the causal status of each effect group for traits as well as variant representation within each effect group. Second, we estimate the functional enrichment of causal variants and prioritize variants according to relevant

functional information. As outputs, our method yields variant-level PIP estimates, set-level posterior summary as well as enrichment estimates of functional annotations.

2.2 SparsePro demonstrates better performance in locus simulation

To evaluate the performance of SparsePro, we performed simulations with different number of causal variants and functional enrichment settings using the UK Biobank genotype data from different genomic loci (**Methods**). The results obtained from one locus (chr22: 31,000,000-32,000,000, Genome Assembly hg19) under one simulation setting: K = 5 (causal variants) and W = 2 (log-relative enrichment) are shown in **Figure** 2A and **Figure** 2B. An accurate method should identify true causal variants (red dots) with high PIP and identify non-causal variants (black dots) with low PIP and have a high area under precision recall curve (AUPRC).

We found that variant-level PIPs obtained from FINEMAP, SuSiE and SparsePro- (SparsePro implementation without functional information) were highly similar for most variants. Compared to FINEMAP, SparsePro- had fewer false positives; compared to SuSiE, SparsePro- was able to detect more causal variants with a higher PIP (**Figure** 2A). Aggregating all simulation replicates, SparsePro- achieved an AUPRC of 0.9059 in identifying true causal variants while the second best statistical fine-mapping method (SuSiE) achieved an AUPRC of 0.8685 (**Figure** 2B). Incorporating functional annotations can improve fine-mapping performance with SparsePro+ achieved an AUPRC of 0.9779 while PAINTOR+ achieved an AUPRC of 0.9349 (**Figure** 2B).

Moreover, SparsePro provided better set-level summary with a higher recall and precision and a comparable set size (**Supplementary Table** S1 and **Supplementary Figures** S1-S3). For example, in the showcased simulation setting, on average, 95.2% of the simulated causal variants were included in causal sets defined by SparsePro- (by combining all causal effect groups), higher than 70.4% by SuSiE and 90.8% by FINEMAP. Also, the mean precision for SparsePro causal sets was 0.684, higher than 0.483 of SuSiE and 0.522 of FINEMAP, despite its slightly larger size (11.8, compared to 10.5 of SuSiE and 11.3 of FINEMAP). With functional information incorporated, the causal sets defined by SparsePro+ on average included 98.8% of

the causal variants with the smallest mean size of 8.8 and the highest mean precision of 0.684 among all methods. The 95% credible sets derived from PAINTOR- were much larger with a mean size of 132.7 and consequently a low mean precision of 0.040. Similarly, 95% credible sets from PAINTOR+ had a mean size of 90.5 with a mean precision of 0.063. These trends were consistent for the different genomic loci (**Supplementary Table** S2).

In other simulation settings, SparsePro also consistently demonstrated the highest AUPRC among all methods (**Figure** 2C). When the number of causal variants was small, the AUPRC of FINEMAP was comparable to SparsePro-; however, as the number of causal variants increased, the performance of FINEMAP deteriorated (**Figure** 2C). Notably, when there was no actual functional enrichment (W = 0), none of the annotations demonstrated significance in the enrichment test (**Methods**) in SparsePro (**Supplementary Table** S3 and **Supplementary Figure** S4), hence SparsePro+ was not implemented. In contrast, PAINTOR did not select annotations by relevance, which could lead to worse performance of PAINTOR+ compared to PAINTOR- when there was no actual functional enrichment. For both PAINTOR+ and SparsePro+, stronger functional enrichment resulted in higher power for detecting and estimating functional enrichment (**Supplementary Tables** S3 and S4, and **Supplementary Figures** S4 and S5). Consequently, the magnitude of improvement of their performance became more evident compared to SparsePro- and PAINTOR-, respectively, (**Figure** 2C).

In these simulations, SparsePro also demonstrated superior computation efficiency (**Figure** 2D), with the computation cost only increasing marginally as the number of causal variants increased. In contrast, while FINEMAP was the fastest approach when there was only one causal variant, its computational cost increased substantially (**Figure** 2D).

2.3 SparsePro demonstrates better performance in genome-wide simulation

In addition to simulations on genomic loci, we also performed genome-wide simulations that were similar to analyzing published GWAS datasets (**Methods**). The results obtained for a W = 2 (log-relative enrichment) setting are shown in **Figure** 3A and **Figure** 3B. We found that

SparsePro- and SparsePro+ identified more true causal variants with a higher PIP than SuSiE and PolyFun-informed SuSiE, respectively (**Figure** 3A). With W = 2 (log-relative enrichment), both enrichment tests in SparsePro and PolyFun correctly identified the enriched annotations (**Supplementary Tables** S5 and S6, and **Supplementary Figures** S6 and S7). As a result, SparsePro+ achieved the highest AUPRC of 0.9779 while PolyFun-informed SuSiE had an AUPRC of 0.9623 (**Figure** 3B). Notably, when there was no functional enrichment, none of the annotations was deemed significant in the enrichment test performed by SparsePro. In contrast, PolyFun still detected significant functional enrichment in non-synonymous variants (**Supplementary Table** S6 and **Supplementary Figure** S7, resulting in an AUPRC of 0.9123, lower than 0.9286 achieved by SuSiE and 0.9382 achieved by SparsePro- (**Figure** 3B).

Consistent with the simulations on genomic loci, the causal sets defined by SparsePro also contained more causal variants and had a higher precision. For example, with W = 2 (log-relative enrichment), causal sets derived from SparsePro- on average covered 83.9% of the true causal variants with a mean size of 135.9, while the causal sets of SuSiE on average contained 64.4% with of the true causal variants with a mean size of 98.6 (**Figure** 3C and **Supplementary Table** S7). When functional annotations were incorporated, the causal sets derived by SparsePro+ had a smaller mean size of 117.9 compared to SparesePro-, with an improved mean coverage of 92.9% whereas PolyFun-informed SuSiE had a mean size of 104.8 with an mean coverage of 78.0% (**Figure** 3C and **Supplementary Table** S7).

SparsePro consistently demonstrated a high computational efficiency, as overall, SparsePro was approximately five times faster than SuSiE in genome-wide simulations (**Figure** 3C).

2.4 Functional enrichment for five clinical biomarkers of vital organs

We performed GWAS in the UK Biobank [1] for five clinical biomarkers: FEV1-FVC ratio (FFR; lung function), estimated glomerular filtration rate (eGFR; kidney function), pulse rate (heart function), blood gamma-glutamyl transferase (liver function) and blood glucose level (pancreatic islet function). The fine-mapping results for these five traits demonstrated that functional annotations (**Methods**) were informative for prioritizing causal variants. For example, causal

variants for all five traits were strongly enriched in non-synonymous variants (**Supplementary Figure** S8 and **Supplementary Table** S8). As expected, causal variants for eGFR identified by SparsePro- had an 11.7-fold (95% CI: 8.8-15.5) enrichment in non-synonymous variants. In SparsePro+, since relevant annotations were used for specifying priors, the resulting enrichment estimates (27.1 fold; 95% CI: 22.2-33.0 for eGFR) were even larger (**Supplementary Figure** S8 and **Supplementary Table** S8).

We used a different set of tissue-specific functional annotations to evaluate the biological relevance (**Methods**); and found that the magnitude of enrichment for each clinical biomarker was the strongest for the corresponding tissue as determined by SparsePro. For example, causal variants for eGFR were highly enriched in the annotation derived from histone marks of kidney chromatin, where there was a 4.13-fold (95% CI: 3.48-4.90) enrichment based on SparsePro- results, and a 5.42-fold (95% CI: 4.64-6.33) enrichment based on SparsePro+ results (**Figure** 4A and **Supplementary Table** S9). In comparison, causal variants identified by SuSiE and PolyFun-informed SuSiE had a 2.45-fold (95% CI: 2.14-2.80) and 4.25-fold (95% CI: 4.06-5.05) enrichment in the same histone annotation, respectively (**Figure** 4A and **Supplementary Table** S9).

2.5 Heritability explained by variants in causal set

Furthermore, variants in causal sets generated by SparsePro+ captured the highest proportion of trait variance compared to other methods (**Methods**and **Figure** 4B). For example, variants in the causal set of FFR derived from SparsePro- jointly had an adjusted R^2 of 0.093, higher than 0.088 achieved by SuSiE. Similarly, using functioal information, variants in the causal set derived from SparsePro+ achieved an adjusted R^2 of 0.099, while PolyFun-informed SuSiE achieved an adjusted R^2 of 0.092 (**Figure** 4B and **Supplementary Table** S10).

2.6 Evidence of genetic coordination of vital organs

The variants identified in causal sets mapped to genes in both core and regulatory pathways for these biomarkers (**Supplementary Tables** S11-S20). Six genes harbored causal sets for

four out of the five biomarkers derived from SparsePro+. Interestingly, we found variant rs1260326 (**Figure** 5B), a missense variant (Leu446Pro) in gene *GCKR*, was fine-mapped for glomerular filtration rate (PIP = 0.9997), blood glucose level (PIP = 0.9947), pulse rate (PIP = 0.8615) and gamma-GT level (PIP = 1.000). Notably, this specific variant is significantly associated with several glycemic traits [17] and quantitative traits for metabolic syndromes and comorbidities [18, 19], and has been implicated in the functions of the liver and other vital organs [20–22]. Other highly pleiotropic genes included four transcription factors *FOXN3*, *GLIS3*, *RREB1*, *SOX5*, and an uncharacterized gene *RP11-10A14.4* (**Supplementary Figures** S9-S13). These findings, along with other variants in the causal sets may present promising genetic targets for experimental validations in a larger effort towards understanding the mechanisms of genetic coordination among vital organs.

3 Discussion

Accurately identifying causal variants is fundamental to genetics and particularly important for interpreting GWAS results [5, 8]. In this work, we developed SparsePro, an efficient fine-mapping method to help prioritize causal variants for complex traits with GWAS summary statistics and functional information. Through simulation studies, we showed that SparsePro was more accurate and efficient compared to existing methods. By fine-mapping genetic associations with five clinical biomarkers for vital organ functions, we demonstrated that functional annotation were useful in prioritizing biologically relevant variants. We also showed that PIP estimates and causal sets defined by SparsePro were biologically relevant with meaningful enrichment in tissue-specific annotations and high proportion of trait variance explained.

Compared to classical fine-mapping methods that conduct an exhaustive search of variant-level causal configurations, both SuSiE and SparsePro greatly reduced the computational cost for fine-mapping by examining group-level causal configurations. Because each group-level causal configuration corresponds to multiple variant-level causal configurations with similar likelihoods, the number of configurations that need to be evaluated is greatly reduced, making the fine-mapping process much more efficient.

With SparsePro, we further improved the accuracy and computational efficiency of the intuitive Bayes factor based iterative Bayesian stepwise selection algorithm in SuSiE. In most simulations, compared with SuSiE, SparsePro detected more causal variants with higher posterior probabilities with less computation time. There are several characteristics in SparsePro contributing to these improvements. First, the paired mean field factorization based variational inference algorithm only requires statistics specified in **Algorithm** 1. Additional statistics used in SuSiE, for example, $\mathbf{y}^\top\mathbf{y}$, were not evaluated. Second, in SuSiE, Bayes factors were normalized by sum of Bayes factors across all variants while SparsePro used the $softmax$ function to normalize posterior probabilities which increased power for identifying causal variants. For example, in the comparison of PIP obtained by SparsePro- and SuSiE (**Figure** 2A and **Figure** 3A), for a number of causal variants (red dots), PIP obtained by SparsePro was close to 1 while PIP obtained by SuSiE was mostly in the range of 0.2 to 0.8. Third, SparsePro adopted accurate and robust HESS estimates [23] for specifying prior while SuSiE incorporated hyperparameter estimations in the iterative algorithm, which may lead to convergence issue.

Moreover, we also included functional information to further prioritize causal variants in SparsePro. As demonstrated in fine-mapping of functional biomarkers, including relevant functional information can improve fine-mapping on top of statistical evidence. However, incorporating irrelevant functional information can compromise fine-mapping performance. In order to address this issue, SparsePro incorporates a statistical test to assess the relevance of functional annotations before integrating them into the fine-mapping procedure. This ensures that functional annotations serve as complementary evidence when relevant, and do not compromise the accuracy of fine-mapping when they are not relevant. Furthermore, the effect of functional annotation is most evident when statistical evidence alone is not sufficient to discern causal variants. This is demonstrated in simulations as the precision-recall curve of SparsePro+ and SparsePro- separated at high recall region where we included more variants with relative smaller PIP. Compared to other fine-mapping methods that leverage functional annotations, such as PolyFun [14] and PAINTOR [13], this distinctive characteristic of SparsePro ensures its robustness to misspecification of functional data.

There are some cautions in fine-mapping analysis that also applies to SparsePro. First, Sparse-

Pro requires a close match between the GWAS study population and the reference panel used for LD calculation. Severe LD mismatch can lead to convergence issue of the algorithm. Second, the identification of causal variants in fine-mapping relies on the rigorousness of the study design in GWAS, and is only possible when unmeasured confounding factors such as population stratification have been carefully addressed. Third, SparsePro does not account for genetic effect heterogeneity, where causal effect sizes may vary across different subpopulations with different environmental exposures.

In summary, SparsePro is an accurate and efficient fine-mapping method integrating statistical evidence and 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.

4 Methods

4.1 SparsePro for genetic fine-mapping

In SparsePro, we use a generative model to integrate statistical evidence from GWAS summary statistics and functional evidence from variant annotations. (**Figure** 1). First, we use the functional annotation to derive prior inclusion probability $\tilde{\pi}_g$ for the gth variant ($g \in \{1, \ldots, G\}$):

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

where $\mathbf{A}_g$ is a $1 \times M$ row vector of M functional annotations for the gth variant; and $\mathbf{w}$ is a $M \times 1$ vector of logarithm of relative enrichment. Here, we use the $softmax$ function to ensure the prior probabilities are normalized. If no functional information is provided, the prior inclusion probability is considered equal for all variants, i.e. $\tilde{\pi}_g = \frac{1}{G}$. We assume that there are altogether K causal effect groups and the representation of the kth effect group is specified by a sparse

indicator $\mathbf{s}_k$:

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

where $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, ..., \tilde{\pi}_G)$. The effect size can be specified as:

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

For the continuous trait $\mathbf{y}_{N\times 1}$ over N individuals we have:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\sum_k \mathbf{s}_k\beta_k, \tau_y^{-1}\mathbf{I})$$

where τ_β^{-1} and τ_y^{-1} are used for specifying prior distributions for effect sizes and the residual errors for trait. We discuss heritability estimates based strategy for specifying prior in SparsePro in the **Supplementary Notes**.

In SparsePro, both variant-level PIP, set-level fine-mapping summary as well as functional enrichment estimates can be derived from the posterior distribution of sparse indicator s_k. We adopt an efficient variational inference algorithm [24, 25] adapted for GWAS summary statistics for posterior inference as showcased in **Algorithm** 1 (detailed in the **Supplementary Notes**).

4.2 Locus simulation studies

We conducted locus simulations to examine the performance of fine-mapping methods under different parameter settings. We randomly selected three 1-Mb regions, and extracted genotypes for 353,570 unrelated UK Biobank White British ancestry individuals [1]. For each locus, we generated 50 replicates for each combination of parameters: $K \in \{1, 2, 5, 10\}$ (number of causal variants) and $W \in \{0, 1, 2\}$ (log-relative enrichment) among variants that were annotated as “conserved sequences” [26], “DNase I hypersensitive sites” (DHS) [27], “non-synonymous” [28], or overlapping with histone marks H3K27ac [29] or H3K4me3 [27]. Causal variants in each simulation replicate were randomly assigned. Then, we used the GCTA GWAS simulation

pipeline [30] to simulate a continuous trait with a per-variant heritability of 1×10^{-4}. We tested the association between each variant and the simulated trait, and obtained GWAS summary statistics using the fastGWA software [31].

Next, we ran the different fine-mapping programs with the GWAS summary statistics and in-sample LD as inputs. For methods using functional annotations, we provided the aforementioned five annotations with enrichment of causal variants as well as five additional annotations without enrichment: "actively transcribed regions" [32], "transcription start sites" [32], "promoter regions" [33], "5'-untranslated regions" [28], and "3'-untranslated regions" [28]. The statistical fine-mapping results for SparsePro were denoted as "SparsePro-". Annotations with a G-test p-value $< 1 \times 10^{-5}$ were selected to conduct functionally informed fine-mapping, and the results were denoted as "SparsePro+". In addition, we conducted statistical fine-mapping using the stochastic shotgun search mode of FINEMAP (V1.4) and the function "susie_rss" from SuSiE (V0.12.16). The mcmc mode for PAINTOR (V3.0) was used to obtain the baseline model results and the annotated model results, separately denoted as "PAINTOR-" and "PAINTOR+".

4.3 Genome-wide simulation studies

Since the efficiency of our methods allows for genome-wide fine-mapping, we also conducted genome-wide simulations. As in locus simulations, we obtained genotypes of 353,570 unrelated UK Biobank White British individuals on chromosome 22 and sampled 100 causal variants with $W \in \{0, 1, 2\}$ (log-relative enrichment) amongst variants that were annotated as "conserved sequences" [26], "DNase I hypersensitive sites" (DHS) [27], "non-synonymous" [28], or that overlapped with histone marks H3K27ac [29] or H3K4me3 [27]. We used the GCTA GWAS simulation pipeline [30] to simulate a continuous trait with a per-variant heritability of 1×10^{-4}. We tested the association between each variant and the simulated trait, and obtained GWAS summary statistics using the fastGWA software [31]. This process was repeated 22 times to obtain genome-wide GWAS summary statistics. Additionally, we obtained LD information calculated using the UK Biobank participants from `https://alkesgroup.broadinstitute.org/UKBB_LD/` [14]. These LD matrices were generated for genome-wide variants 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, initially without any functional annotation information. We denoted the fine-mapping results as "SparsePro-". Next, we used the statistical fine-mapping results to estimate the relative enrichment for functional annotations. Annotations with a G-test p-value $< 1 \times 10^{-5}$ were selected to conduct functionally informed fine-mapping, and the results were denoted as "SparsePro+". PIPs for variants in the 1-Mb centre of each 3-Mb sliding window were extracted for comparison. We used the function "susie_rss" from SuSiE (V0.12.16). We aggregated summary statistics from 22 simulated chromosomes and used PolyFun with the "baselineLF2.2.UKB" model [14] 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 [14]. The estimated priors were provided to SuSiE via "prior_weights".

In both the locus-specific simulations and genome-wide simulations, we compared the performance of the different fine-mapping methods by (i) AUPRC calculated from variant-level PIP estimates; (ii) precision, recall and size of set-level summary; (iii) computation time on a 2.1 GHz CPU.

For set-level summary, 95% credible sets were used for classical fine-mapping methods: FINEMAP and PAINTOR. In SuSiE, variants in all identified 95% credible sets were combined to form a causal set. In SparsePro, variants from all 95% representation of identified causal effect groups were included in the causal set.

4.4 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, blood gamma-glutamyl transferase (gamma-GT)

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 [30, 31] to obtain summary statistics.

Using the summary statistics and the matched LD information [14], we performed genome-wide fine-mapping with SparsePro-, SparsePro+, SuSiE and PolyFun-informed SuSiE as described in the simulation analyses.

To evaluate the biological relevance of fine-mapping results obtained using different methods, we assessed their relative enrichment in 10 tissue-specific annotations derived from four histone marks H3K4me1, H3K4me3, H3K9ac, and H3K27ac by Finucane et al [34]. This set of annotation information was not used by any functional fine-mapping methods.

Additionally, we calculated the proportion of trait variance explained by variants in the causal sets. For each identified effect, we selected the variant with the largest PIP to represent this causal effect. 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) on causal set variants was used as a surrogate for heritability.

We used phenogram [35] to illustrate genes that harbored causal set variants for at least two traits to explore possible pleiotropic effects.

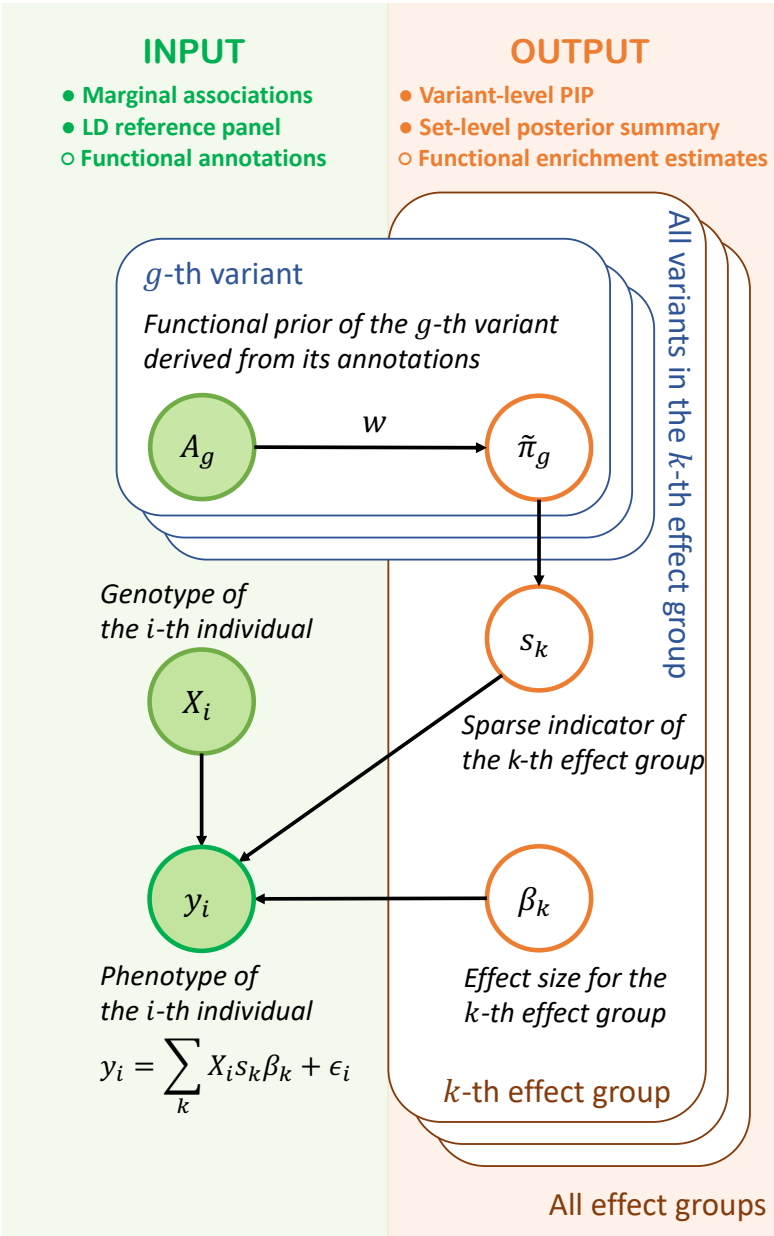

Figure 1: SparsePro for genetic fine-mapping. The data generative process in SparsePro is depicted in this graphical model. Green shaded nodes represent observed variables: functional annotation information A_g for the gth variant, genotype $\mathbf{X_i}$ and trait $\mathbf{y}_i$ for the ith individual. The orange unshaded nodes represent latent variables. Specifically, $\tilde{\pi}_g$ is the prior inclusion probability for the gth variant derived from functional annotation information. For the kth causal effect group, $\mathbf{s}_k$ is a sparse indicator specifying which variants are included in this effect group and β_k represents the corresponding effect size. As a result, posterior summary can be obtained from posterior distribution of $\mathbf{s}_k$. Here, we assume individual-level data are available and adaption to GWAS summary statistics is detailed in the **Supplementary Notes**, with an efficient variational inference algorithm for posterior inference.

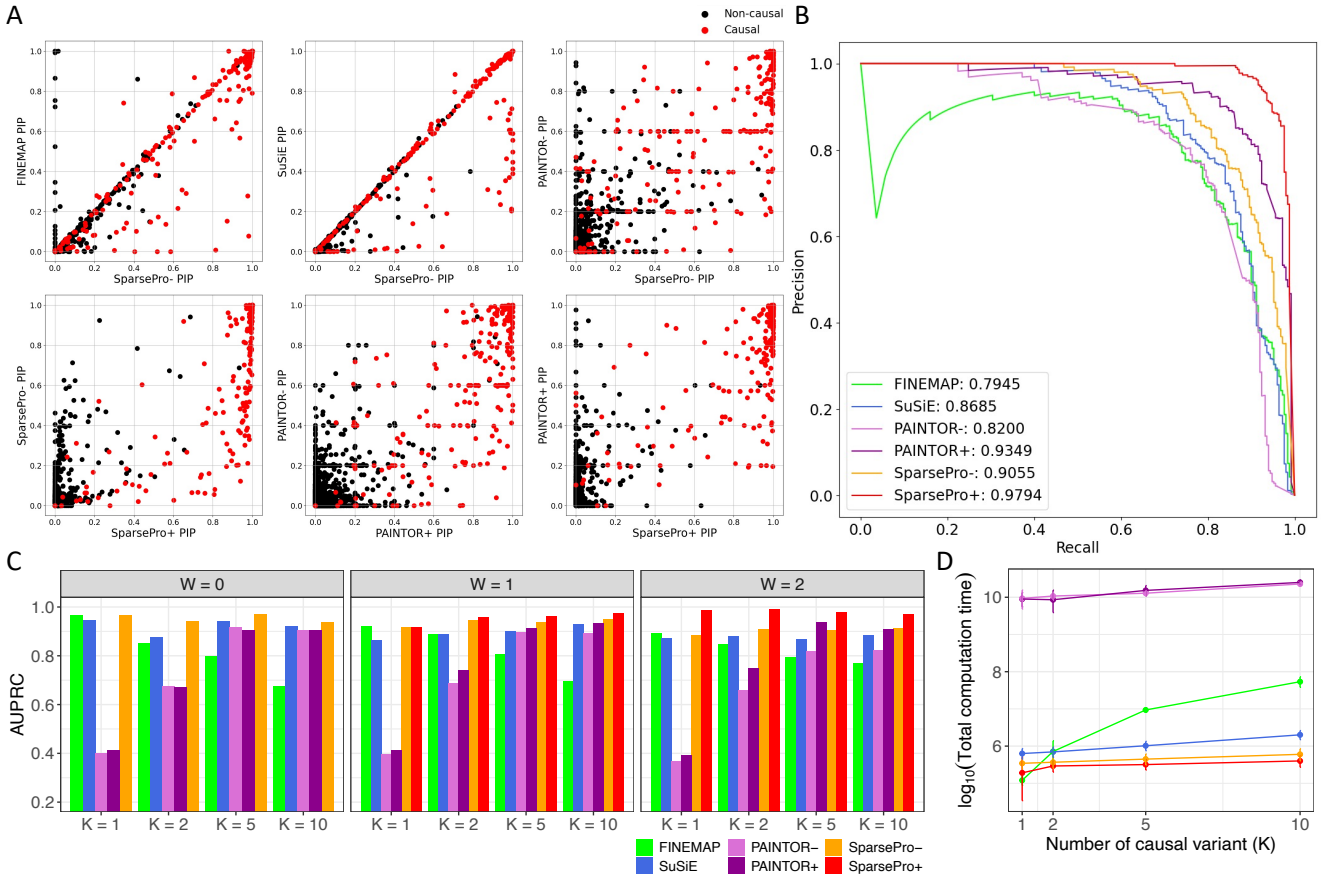

Figure 2: SparsePro demonstrated improved accuracy and computational efficiency in locus simulations. (A) Comparison of posterior inclusion probabilities obtained using different methods in the setting of K = 5 (number of causal variants) and W = 2 (log-relative enrichment). Each dot represents a variant. True causal variants are colored red and non-causal variants are colored black. (B) Precision-recall curves in the setting of K = 5 and W = 2. The area under the precision-recall curve (AUPRC) for each method is indicated. (C) Summary of AUPRC in different settings of K and W. (D) Summary of computation time in different settings.

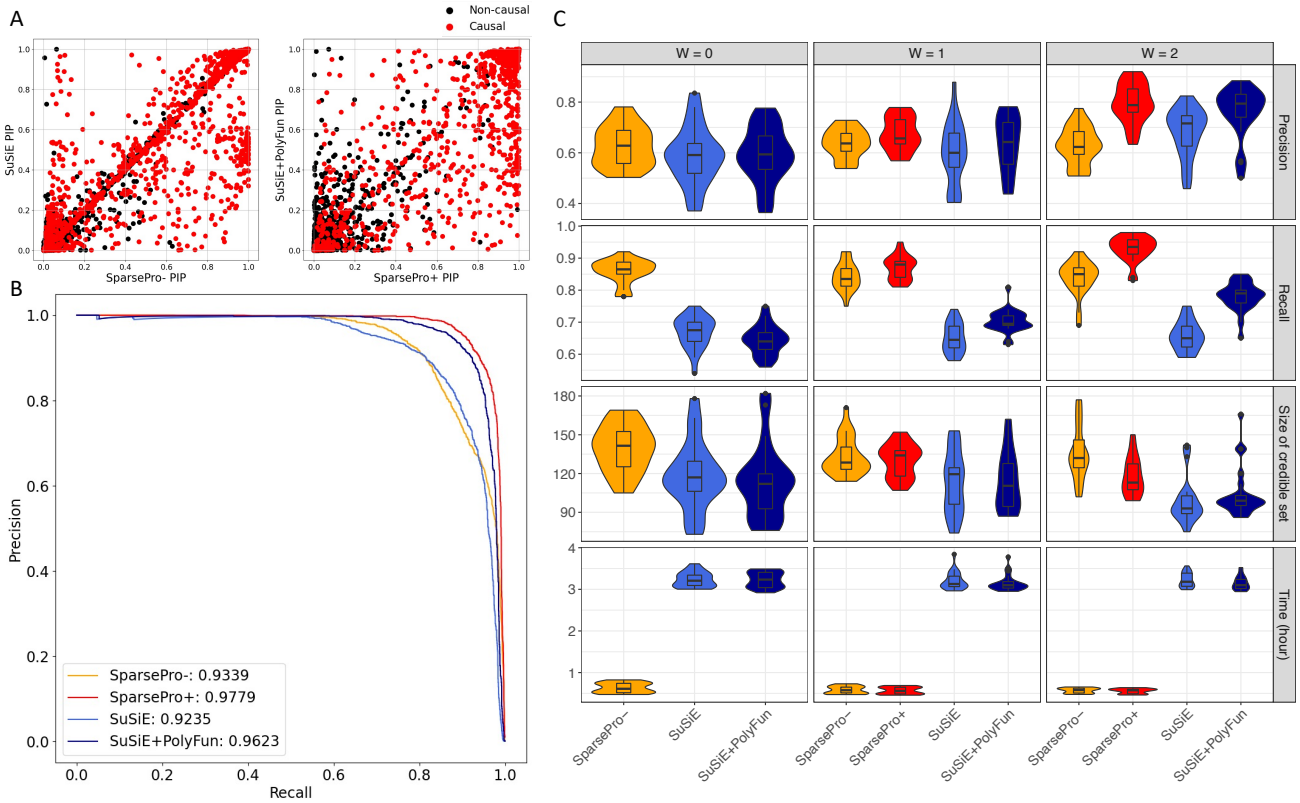

Figure 3: SparsePro demonstrated comparable or improved accuracy and computational efficiency in genome-wide simulations. (A) Comparison of posterior inclusion probabilities obtained using different methods in the setting of W = 2 (log-relative enrichment). (B) Precision-recall curves in the setting of W = 2. The area under the precision-recall curve (AUPRC) for each method is indicated. (C) Summary of computation time and precision, recall and size of credible sets under different parameter settings. Violin plots represent density of values. Boxes represent inter-quartile ranges with median values indicated by horizontal lines.

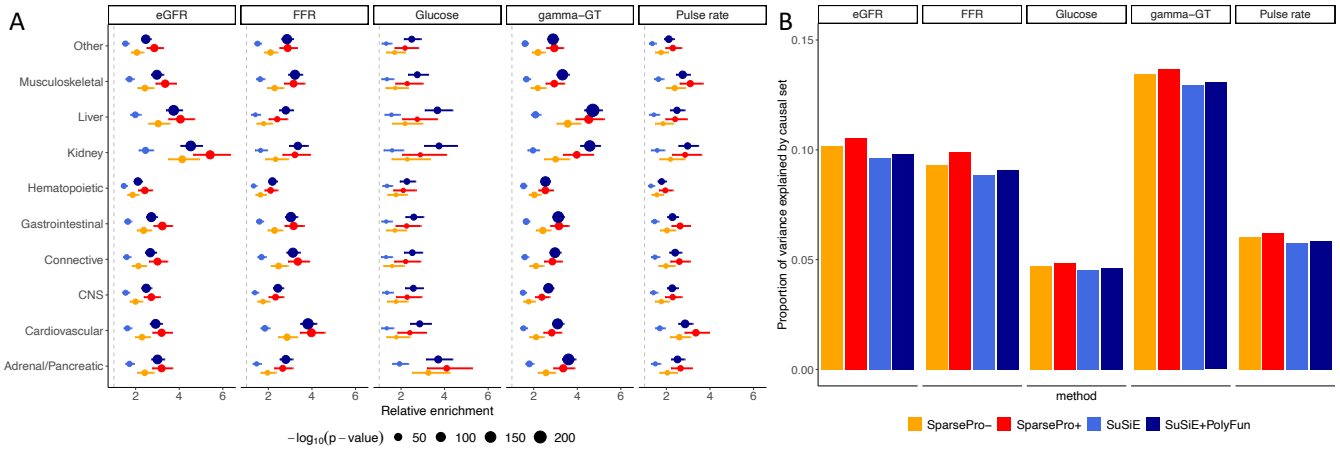

Figure 4: Biological relevance of fine-mapped genetic associations for five functional biomarkers of vital organs. (A) Relative enrichment of causal variants in tissue-specific annotations. Each row denote a tissue-specific annotation derived from histone marks (**Methods**) and each column denotes a functional biomarker. Error bars represent 95% confidence intervals for enrichment estimates. (B) Proportion of total trait variance captured by variants in causal sets.

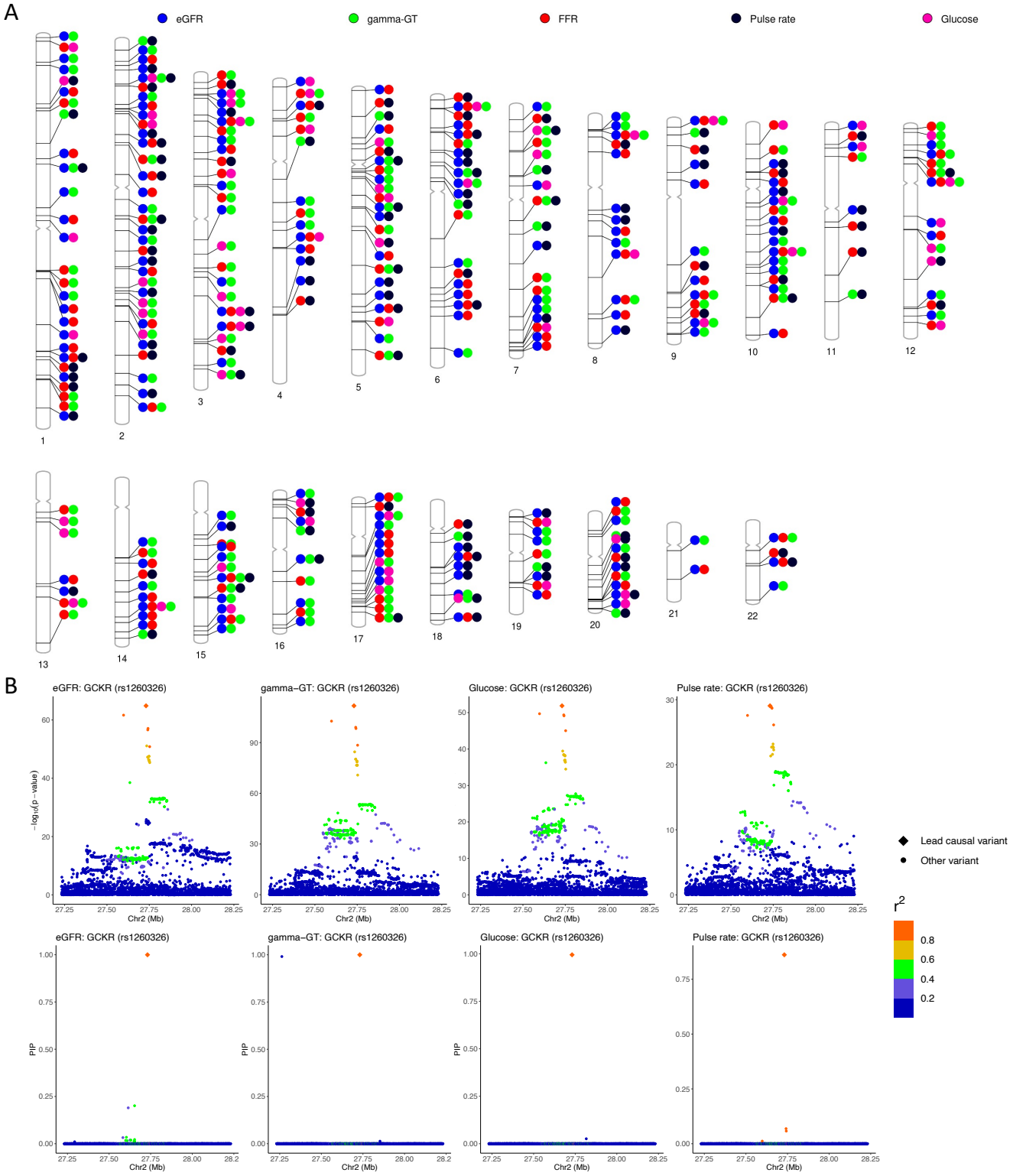

Figure 5: Genes harboring causal sets for five clinical biomarkers of vital organs. (A) Genome-wide distribution of genes harboring causal sets for at least two clinical biomarkers. (B) *GCKR* locus with fine-mapped variant rs1260326. This variants was deemed causal for eGFR, gamma-GT, glucose, and pulse rate. P-values from GWASs and inferred posterior inclusion probabilities are illustrated. Variants within a $\pm$500kb window are colored by their linkage disequilibrium r^2 with rs1260326.

5 Acknowledgements

Y.L. 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 and we thank NeuroHub for providing access to data resources. This study was enabled, in part, by support from Calcul Québec and Compute Canada. W.Z. has been supported by a doctoral training fellowship from the FRQNT and 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. We thank Dr. Robert Sladek for helpful discussion and suggestions.

6 Author contributions

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

7 Disclosures

The authors declare no conflict of interest.

8 Data and Software Availability

SparsePro is an open-access software and publicly available at `https://github.com/zhwm/SparsePro`. The simulation scripts are available at `https://github.com/zhwm/SparsePro_analysis`. 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 0.12.16) 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 annotation was downloaded from `https://storage.googleapis.com/broad-alkesgroup-public/LDSCORE/ct_and_ctg_bedfiles.tgz`

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

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

15. 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).

16. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the “Sum of Single Effects” model. *PLoS Genetics* **18,** e1010299 (2022).

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

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

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

20. 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).

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

22. 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).

23. 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).

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

25. 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).

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

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

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

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

30. 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).

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

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

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

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

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

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

37. Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* **5,** 3–55 (2001).

SparsePro Supplementary Information

1 A variational inference algorithm for fine-mapping

In SparsePro (**Methods**), we have:

$$\tilde{\boldsymbol{\pi}} = softmax(\mathbf{A}\mathbf{w})$$
$$\mathbf{s}_k \sim \mathbf{Multinomial}(1, \tilde{\boldsymbol{\pi}})$$
$$\beta_k \sim \mathcal{N}(0, \tau_\beta^{-1})$$
$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\sum_k \mathbf{s}_k\beta_k, \tau_y^{-1}\mathbf{I})$$

Denoting $\mathbf{S}_{G\times K} = [\mathbf{s}_1, ..., \mathbf{s}_K]$ and $\boldsymbol{\beta} = [\beta_1, ..., \beta_K]$, we have the joint probability:

$$p(\mathbf{y}, \mathbf{S}, \boldsymbol{\beta}|\mathbf{X}) = p(\mathbf{y}|\mathbf{X}, \mathbf{S}, \boldsymbol{\beta}) \prod_k p(\beta_k) \prod_k p(\mathbf{s}_k) \tag{1}$$

The goal of fine-mapping is to infer the posterior probability of the sparse projections $\mathbf{s}_k$ and the effect sizes β_k given $\mathbf{y}$ and $\mathbf{X}$ for $k \in \{1, ..., K\}$:

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

We use a paired mean field factorized variational family $q(\mathbf{S}, \boldsymbol{\beta})$ to approximate the posterior [25]:

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

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 a similar mode and shape as the desired pos-

terior distribution, and that such inference can achieve high accuracy with substantially improved computational efficiency [25]. 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) [24]:

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

To maximize the above ELBO, the following requirement should be satisfied for each k [24]:

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

where $E_{q(\mathbf{S}_{\backslash k}, \boldsymbol{\beta}_{\backslash k})}$ is the expectation with respect to the variational distribution excluding the kth component. With the joint probability provided in Equation (1) we have the following log likelihood:

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

Taking expectation with respect to the variational distribution excluding the kth component, we have the following logarithm of variational distribution:

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

where $\tilde{\boldsymbol{\beta}}_{\backslash k} = E_{q(\mathbf{S}_{\backslash k}, \boldsymbol{\beta}_{\backslash k})}[\sum_{k' \neq k}\mathbf{s}_k\beta_k]$. We recognize that

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

By matching sufficient statistics for this normal distribution, we obtain variational parameters:

$$\tau_{kg}^* = \tau_y \mathbf{X}_g^\top \mathbf{X}_g + \tau_\beta$$
$$\mu_{kg}^* = \frac{\tau_y}{\tau_{kg}^*} \mathbf{X}_g^\top (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_{\backslash k})$$

By integrating out β_k in Equation (2), we have:

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

Therefore, the posterior probability of the gth variant being causal in the kth effect group can be estimated as:

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

When functional annotations are available, we use an additional step to integrate relevant annotations. After obtaining variational estimates, we have:

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

In order to estimate functional enrichment weight for the mth annotation, we can take partial

derivatives of ELBO with respect to w_m:

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

where

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

Setting the derivative to 0, we have:

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

and solve for w_m:

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

This estimate is analogous to a relative risk estimate in a 2×2 contingency table. Suppose we have one binary 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 cor-

responds to the sum of posterior probability for variants without the annotation. The standard error of relative enrichment can be calculated based on the standard error of a relative risk:

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

We can also evaluate the statistical significance of enrichment with the log likelihood ratio test (G-test) [36] and select relevant functional information to further prioritize causal variants.

The information in individual-level genotype $\mathbf{X}$ and phenotype data $\mathbf{y}$ are used in the form of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}_g^\top\mathbf{y}$ throughout the algorithm. These statistics can be derived from publicly available GWAS summary statistics z-scores (the ratio of per-variant effect size estimate to its standard error) and matched LD information (estimates of pairwise variant-variant Pearson correlation coefficient). With standardized genotypes and phenotypes, we can derived required quantities using summary-level data:

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

$$\mathbf{X}^\top\mathbf{y} = \sqrt{N}\mathbf{z}$$

where N is the sample size, $\mathbf{LD}$ is the variant correlation matrix and $\mathbf{z}$ stands for z-scores in GWAS summary statistics. For binary traits, we can use z-scores derived from log-odds ratios and their standard errors to approximate the required quantities.

Apart from the required quantities derived from GWAS summary statistics, there are also hyperparameters required to be specified: K for the total number of causal effect groups, τ_β^{-1} for specifying prior distributions for effect sizes and τ_y^{-1} are for specifying residual errors for trait. We suggest setting K to be larger than the max number of causal signals in the locus, which can be determined through a pruning and thresholding procedure. The variance hyperparameters can be determined via heritability estimates. Specifically, we can obtain the local heritability ($\hat{h}^2$) in a locus as well as per-variant heritability ($\hat{h}_v^2$) with the HESS [23] estimator using GWAS summary statistics, and use them to set variance hyperparameters: $\tau_\beta^{-1} = \hat{h}_v^2$ and $\tau_y^{-1} = 1 - \hat{h}^2$. Shi et al [23] found that the HESS estimator is able to accurately and robustly estimate heritability in a variety of genetic architectures, thus provides suitable estimates for

specifying prior distribution in SparsePro.

After convergence of the variational inference algorithm we can easily obtain posterior summary for fine-mapping at both variant-level and set-level using variational estimates of $\mathbf{s}_k$. Specifically, for $k \in 1, ..., K$, by comparing the entropy difference in prior distribution of s_k and posterior distribution of s_k, we can identify causal effect groups and find variant representation for each causal effect group. Entropy is a measure of the uncertainty in probability distributions [37]. The prior distribution of $\mathbf{s}_k \sim \mathbf{Multinomial}(1, \tilde{\boldsymbol{\pi}})$ represents our initial belief of prior inclusion probabilities for each variant based on functional annotation information. For $k \in 1, ..., K$, if GWAS summary statistics data in support of the causal role of kth effect group, the posterior distribution $\mathbf{s}_k \sim \mathbf{Multinomial}(1, [\gamma^*_{k1}, \ldots, \gamma^*_{kG}])$ will have an extremely low entropy with most of the probability mass distributed on the causal variant and variants in high LD with the causal variant. Otherwise, without support from GWAS summary statistics data, the posterior distribution will be similar to the prior distribution and will have a high entropy, as it is in the prior distribution. In practice, we set the entropy difference cutoff to be $\log 20$ for identifying causal effect groups.

We obtain set-level summary by sorting variants by their posterior probability mass and including those with large probabilities, we can obtain a 95% representation for each causal effect group. This representation is created by including variants until the total probability mass exceeds 0.95. This process allows us to identify the most likely causal variants within each group and provide a summary of the set-level results.

For variant-level summary, we can obtain PIP for the gth variant as the maximum of all K posterior probabilities:

$$\gamma^*_g = \max(\gamma^*_{1g}, \ldots, \gamma^*_{Kg})$$

In summary, we have now derived **Algorithm** 1 for fine-mapping with SparsePro:

Algorithm 1: SparsePro for genetic fine-mapping

Data: $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}_g^\top\mathbf{y}$ calculated from GWAS summary statistics $\mathbf{z}$, N;
functional annotation $\mathbf{A}$;
LD information $\mathbf{LD}$;
the total number of causal effect groups in the locus K

Result: variant-level PIP;
set-level posterior summary;
log-relative enrichment w

```
estimate τ_β^{-1}, τ_y^{-1} with HESS from z, N;
while ELBO not converge do
    for k = 1 to K do
        update q(s_k);
        update q(β_k|s_k);
    end
    estimate w;
    update π̃;
end
obtain set-level posterior summary;
calculate variant-level PIP
```