A novel Bayesian fine-mapping model using a continuous global-local shrinkage prior with applications in prostate cancer analysis

Xiang Li[1], Pak Chung Sham[2,3], Yan Dora Zhang[1*]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China.
[2]Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
[3]Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

*Corresponding author(s). E-mail(s): doraz@hku.hk;
Contributing authors: freddyl@connect.hku.hk; pcsham@hku.hk;

Abstract

The aim of fine-mapping is to identify genetic variants causally contributing to complex traits or diseases. Existing fine-mapping methods employ discrete Bayesian mixture priors and depend on a pre-specified maximum number of causal variants which may lead to sub-optimal solutions. In this work, we propose a novel fine-mapping method called h2-D2, utilizing a continuous global-local shrinkage prior. We also present an approach to define credible sets of causal variants in continuous prior settings. Simulation studies demonstrate that h2-D2 outperforms the state-of-art fine-mapping methods such as SuSiE and FINEMAP in accurately identifying causal variants and estimating their effect sizes. We further applied h2-D2 to prostate cancer analysis and discovered some previously unknown causal variants. In addition, we inferred 385 target genes associated with the detected causal variants and several pathways that were significantly over-represented by these genes, shedding light on their potential roles in prostate cancer development and progression.

Keywords: fine-mapping, global-local shrinkage prior, causal variant, prostate cancer

Introduction

Genome-wide association studies (GWAS) have discovered numerous genetic variants associated with a wide range of complex traits and diseases [1]. However, pinpointing the specific variants that have causal effects on the traits is challenging due to the presence of high linkage disequilibrium (LD) among single nucleotide polymorphisms (SNPs) and their small effect sizes [2–4]. The goal of statistical fine-mapping is to identify the causal variants that have nonzero effects on the trait, which is essentially a statistical problem known as "variable selection". Since it is difficult to distinguish a causal variant from other variants highly correlated with it without extra information, penalized regression methods sometimes fail to select the true causal variants. On the other hand, Bayesian methods are more appropriate for fine-mapping by providing posterior "credible sets" (CSs) [4]. A level $1-\alpha$ CS is defined as a set of variants that contains at least one causal variant with posterior probability no less than $1-\alpha$ [5, 6]. A CS may contain multiple highly correlated candidate causal variants for further functional validation.

To date, many Bayesian fine-mapping methods have been developed, including CAVIAR [2], CAVIARBF [7], PAINTOR [3], JAM [8], DAP [9], FINEMAP [10, 11], and SuSiE [12]. All these methods are based on discrete mixture priors, specifying a prior probability for each variant being causal. Suppose there are M SNPs in the region of interest, the number of possible models is 2^M. To reduce computational cost, these methods need to set limit on the maximum number of causal variants. However, misspecifying the number may lead to decrease in performance[12]. In addition, existing methods rely on exhaustive search, shotgun stochastic search, or stepwise selection to explore the space of causal configurations, which can be time-consuming or lead to poor suboptimal solutions [5, 12].

In Bayesian analysis, there is another class of shrinkage priors termed "continuous global-local shrinkage priors". Existing continuous priors have been shown to be efficient variable selection tools [13–20] and have been successfully applied in genetic studies, including polygenic risk prediction [21]. However, continuous shrinkage priors are hardly used in fine-mapping. One shortcoming of continuous priors is that they require additional procedures in order to perform variable selection, as the posterior mean of regression coefficients is not sparse almost surely. Existing approaches include hard thresholding methods [15, 22], penalized credible regions [23, 24], and posterior variable selection summary [25]. Nonetheless, these approaches can only produce a single sparse model instead of several candidate models, and cannot generate credible sets similar to those obtained using discrete mixture priors.

In this paper, we introduce a novel fine-mapping method based on a continuous global-local shrinkage prior, called the heritability-induced Dirichlet decomposition (h2-D2) prior, which is a variant of R2-D2 prior [20]. R2-D2 prior possesses both unbounded density around the origin and very heavy tails, thus enabling it to model the extremely sparse structure of the fine-mapping coefficients. Our proposed h2-D2 prior inherits the same desirable properties as R2-D2 and is adapted specifically to GWAS data. Without loss of generality, we will refer to our method, which represents for the entire fine-mapping process, as h2-D2 throughout the manuscript.

Moreover, in order to address the limitations of continuous priors, inspired by the principles of frequentist hypothesis testing, we propose a statistic, termed "credible level", which can be easily computed from posterior samples, to quantify how likely one or a set of SNPs have nonzero effects. We further define credible sets in the framework of continuous priors, offering a selection of candidate variants in the post-selection process.

Our simulation studies show that h2-D2 has better performance in identifying causal variants and accurately estimating effect sizes than the state-of-art fine-mapping methods such as SuSiE and FINEMAP. The CSs produced by h2-D2 exhibit superior power and achieve the target level of coverage when accurate LD matrices are provided. Finally, we apply h2-D2 to prostate cancer GWAS, identifying some novel causal signals that were not previously reported. The identified credible causal variants show significant enrichment in active gene regulatory regions and binding sites of specific transcription factors. In addition, we infer a total of 385 likely target genes associated with these credible causal variants. These genes are significantly over-represented in several pathways, providing valuable insights into the potential biological mechanisms underlying prostate cancer development and progression. We conclude with a discussion of future topics and further describe our software tool h2-D2 to implement the method for public use.

Results

Simulations

We conducted simulation studies using UK Biobank imputed genotype data from $N = 275,768$ unrelated British individuals [26] (Methods). We chose 100 nearly independent LD blocks on chromosome 2 (Supplementary Table 1). For each block, we simulated quantitative traits based on four scenarios. For the first three scenarios, we used the genotypes of all $N = 275,768$ individuals and considered different combinations of local heritability and numbers of causal variants: (1) $h^2 = 0.1\%$, $n_{\text{causal}} = 5$; (2) $h^2 = 0.05\%$, $n_{\text{causal}} = 5$; and (3) $h^2 = 0.1\%$, $n_{\text{causal}} = 10$. In the last scenario, we simulated eQTL studies, where the sample sizes were small ($N = 1,000$), but the effect sizes of causal SNPs were large ($h^2 = 10\%$), and $n_{\text{causal}} = 5$. To examine the influence of LD matrix accuracy on the fine-mapping performance, we computed four LD matrics from different reference panels with varying sample sizes for each block (Methods).

We compared h2-D2 with two state-of-art fine-mapping methods, FINEMAP [10, 11] and SuSiE-RSS [12]. On the SNP level, we evaluated the performance of variable selection using the area under the precision-recall curve (AUPRC), which was computed based on the credible level of each SNP for h2-D2 or the marginal posterior inclusion probability (PIP) of each SNP for SuSiE and FINEMAP. In addition, we assessed the accuracy of effect size estimation using the sum of squared error (SSE) of $\boldsymbol{\beta}$ based on its posterior mean. When using in-sample LD matrices, h2-D2 consistently outperformed SuSiE and FINEMAP in terms of both AUPRC and SSE across all scenarios (Fig. 1a,b). As expected, all methods exhibited degraded performance as the accuracy of the LD matrices decreased. In most cases, h2-D2 still demonstrated

superior performance. Additionally, h2-D2's credible levels were better calibrated than PIPs of SuSiE and FINEMAP, particularly when inaccurate LD matrices were used (Supplementary Fig. 1). The performance of SuSiE was close to that of h2-D2. However, FINEMAP had significantly larger SSE and performed much worse in Scenario 3 where the true number of causal variants was 10.

To gain further insights into the differences among the three methods, we compared the AUPRC for each simulated dataset between h2-D2 and the other two methods (Supplementary Fig. 2). While the AUPRC values were generally close for all three methods across most datasets, h2-D2 exhibited significantly better performance in certain datasets. By visualizing the fine-mapping results of these datasets, we noticed that in many cases if there was a non-causal SNP having moderate LD with one or more causal SNPs and having a stronger marginal association than causal SNPs, SuSiE and FINEMAP tended to select that non-causal SNP instead of the causal ones. Supplementary Fig. 3 provides two examples illustrating this issue. This phenomenon may be attributed to the stepwise selection nature of SuSiE and the shotgun stochastics search algorithm employed by FINEMAP. Once a marginally significant variant is included in the model, it is difficult for discrete-mixture prior-based methods to remove it, i.e., the algorithms are more prone to be trapped into suboptimal solutions. It appears that the refinement step of SuSiE cannot always alleviate this problem. On the other hand, continuous shrinkage prior-based methods allow for the continuous updating of coefficients, enabling smoother transitions among different local modes, and making the Markov Chain Monte Carlo (MCMC) algorithm to explore the space of causal configurations more extensively.

We also compared differences among the three methods in effect size estimation. We grouped the variants into causal and non-causal categories and analysed the prediction error for each group (Methods, Supplementary Figs. 4-5). Although SuSiE and h2-D2 produced similar estimation of causal variant effect sizes, h2-D2 had smaller prediction errors for the non-causal variant effect sizes, suggesting that h2-D2 had lower FDR than SuSiE. While FINEMAP demonstrated the lowest SSE for non-causal variant effect sizes, it grossly underestimated causal variant effect sizes, presumably from excessive shrinkage, resulting in larger SSE compared with SuSiE and h2-D2.

Next, we compared the level 95% CSs produced by the three methods. As shown in Fig. 1c-g, when using $\widehat{\boldsymbol{R}}$ or $\widehat{\boldsymbol{R}}_{\mathrm{UKBB},3000}$, the numbers of 95% CSs generated by the three methods were comparable, and CSs from h2-D2 exhibited higher coverage and greater power in most cases. When using $\widehat{\boldsymbol{R}}_{\mathrm{UKBB},500}$ or $\widehat{\boldsymbol{R}}_{\mathrm{1KG}}$, SuSiE and FINEMAP detected more CSs with higher power but lower coverage, while h2-D2 detected fewer CSs with lower power but higher coverage. These results suggested that the CSs from h2-D2 have a lower false discovery rate (FDR) even when low accuracy LD matrices are used. Although the CSs based on continuous priors may not guarantee the frequentist coverage, we found that the coverage was generally higher or close to the target level of 0.95, except when using $\widehat{\boldsymbol{R}}_{\mathrm{1KG}}$. It is not surprising that 95% CSs from h2-D2 had larger sizes and lower purity, since SuSiE and FINEMAP focus on regions with high posterior probability density and select "best candidates" among a set of SNPs in high LD. In contrast, h2-D2 samples from the full posterior distribution, providing a more comprehensive representation of the uncertainty in the fine-mapping results.

Finally, we compared runtime of the three methods (Fig. 1h). The computational complexity of h2-D2 is proportional to M^2 (where M is the number of variants) and the number of MCMC iterations (n_{MCMC}), while the computational complexity of SuSiE is proportional to M^2 and the maximum number of single effects L. When $n_{\mathrm{MCMC}} = 10000$ and $L = 5$, the runtime of h2-D2 were approximately three times as long as the runtime of SuSiE. The computational complexity of FINEMAP is primarily determined by the maximum number of causal variants and the number of iterations, so the runtime of FINEMAP didn't significantly vary with the number of variants.

Fine-mapping causal variants of prostate cancer

We applied h2-D2 to identify candidate causal variants of prostate cancer (PrCa) using summary GWAS data from a large meta-analysis of European ancestry [27] (Methods). We excluded the major histocompatibility complex MHC region (chr6 25-33 M) from our analysis. The remaining autosomal regions were partitioned into $3,717$ nonoverlapping regions with approximately independent LD (Methods). There are 126 risk variants out of $3,717$ regions (i.e., contain at least one SNP with $P < 5 \times 10^{-8}$). We performed fine-mapping analysis on all $3,717$ regions, which encompassed a total of $6,446,747$ SNPs ($1,342,667$ tag SNPs) with MAF ≥ 0.01 and INFO score ≥ 0.8.

Overall, we identified 164 CSs at 95% level (Supplementary Table 2), containing $4,706$ credible causal variants (366 tags). Among these CSs, 93 overlaped with the 106 CSs in autosomal risk loci reported by ref. [28] and 86 overlaped with the CSs identified by ref. [29]. Out of the 3,717 regions analysed, 92 regions contained a single CS, while 23 regions contained multiple CSs. 6 CSs were detected within non-risk regions. The region with the highest number of CSs was chr8 127708268-128658961, where 15 CSs were detected. This finding is consistent with previous research that chr8q24 region harbors multiple loci associated with PrCa susceptibility [30]. The sizes of the CSs ranged from 1 to 282 variants, with a median size of 12 variants. There were 22 CSs containing only a single variant, including some well-established causal variants of PrCa, such as rs77559646, which disrupts *ANO7* mRNA splicing and protein expression [31], and rs61752561, which affects glycosylation and function of prostate-specific antigen [32] (Table 1).

In our analysis, we identified some novel independent association signals that have not been previously reported. One such example is chr11 68810837-69542062, where four 95% CSs were detected (Fig. 2a, Supplementary Fig. 6a). CS:11-88-1 is represented by rs12275055 ($P = 3.7\times10^{-98}$), which is known to have pleiotropic associations with multiple cancer types [33]. This SNP acts as an eQTL in multiple tissues for *TPCN2*, which plays a role in autophagy progression and extracellular vesicle secretion in cancer cells [34]. The location of CS:11-88-2 overlaps with CS:11-88-1. Hi-C data from the normal prostate cell line RWPE1 indicated that several SNPs within CS:11-88-2 are located in an enhancer region that looping to the promoter of the cell cycle related gene *CCND1* [35]. Furthermore, an interaction between the *TPCN2* promoter and the *CCND1* promoter was detected by H3K27ac HiChIP in the LNCaP prostate cancer cell line [29]. These findings suggest a possible mechanism involving a three-way interaction between an enhancer harboring the causal SNPs in CS:11-88-1 and CS:11-88-2, the *TPCN2* promoter, and the *CCND1* promoter. We also identified

Table 1 Single-SNP credible sets of prostate cancer causal variants

Fine-mapping region[1]	Variant[2]	rsID[3]	AAF[4]	P value[5]	CL[6]	Target gene(s)[7]	Association type(s)[8]
chr2:62482371-64700760	2_63301164_G_A	rs6545977	0.50	7.35×10^{-46}	1		
chr2:241912029-243041411	2_242135265_G_A	rs77559646	0.02	9.93×10^{-21}	1	*ANO7*	NSM[9],INT[10]
chr3:169194244-170170389	3_170083629_C_G	rs61436251	0.21	1.76×10^{-63}	1	*SKIL*	INT
chr5:1279701-1551138	5_1287194_G_A	rs2853677	0.58	0.02	1	*TERT*	INT
chr5:1279701-1551138	5_1288547_T_C	rs2853676	0.73	8.86×10^{-12}	1	*TERT*	INT
chr5:1279701-1551138	5_1292118_G_A	rs71595003	0.03	1.78×10^{-16}	0.96	*TERT*	INT
chr5:1279701-1551138	5_1298017_G_A	rs148297846	0.07	4.98×10^{-15}	0.97		
chr5:1551930-2131681	5_1895829_C_T	rs12653946	0.42	9.58×10^{-22}	1	*IRX4*	INT,Cis eQTL (GTEx)
chr6:159951830-161847113	6_160581374_A_G	rs651164	0.69	2.15×10^{-36}	1	*SOD2*,*ACAT2*, *TCP1*,*MRPL18*	Enhancer (H3K27ac HiChIP, LNCaP)
chr6:159951830-161847113	6_160581502_T_C	rs4646283	0.14	1.31×10^{-5}	0.99	*SOD2*,*ACAT2*, *TCP1*,*MRPL18*	Enhancer (H3K27ac HiChIP, LNCaP)
chr8:127708268-128658961	8_128540776_C_G	rs12549761	0.12	5.20×10^{-77}	1		
chr8:128659713-129297518	8_128665480_C_T	rs4385433	0.37	6.91×10^{-8}	0.96		
chr10:50839567-53146331	10_51549496_T_C	rs10993994	0.62	2.29×10^{-147}	1	*TIMM23B*,*MSMB*, *NCOA4*	INT,R5[11], Cis eQTL (GTEx)
chr11:124697216-125111546	11_125054793_C_T	rs138466039	0.01	2.01×10^{-11}	1	*PKNOX2*	INT,Cis eQTL (GTEx)
chr12:12101106-12922339	12_12871099_T_G	rs2066827	0.24	2.31×10^{-9}	1	*CDKN1B*	NSM,R3[12]
chr13:73847474-74347673	13_74084684_G_A	rs61957204	0.07	3.29×10^{-11}	1		
chr14:23251130-23598976	14_23305649_T_C	rs1004030	0.42	1.55×10^{-8}	0.98	*MMP14*	R5
chr17:7251713-8007416	17_7571752_T_G	rs78378222	0.01	1.73×10^{-9}	1	*TP53*	U3[13]
chr17:45308047-47211560	17_46816630_C_A	rs189183876	0.01	0.27	1		
chr17:45308047-47211560	17_46832497_C_T	rs146240770	0.02	0.05	1		
chr19:51254187-51450534	19_51361382_G_A	rs61752561	0.04	2.33×10^{-8}	1	*KLK3*	NSM,INT
chr22:42872086-43649657	22_43500212_G_T	rs5759167	0.50	5.55×10^{-71}	1		

[1]Chromosome (chr) number and boundary of fine-mapping region (GRCh37/hg19).

[2]Variant ID in the format {chr}_{pos}_{ref_seq}_{alt_seq}.

[3]dbSNP (build 151, GRCh37/hg19) rsID.

[4]Alternative allele frequency of controls in meta-analysis.

[5]Meta-analysis p-value.

[6]h2-D2 redible level.

[7]Putative target gene(s) of the variant.

[8]Association type(s) between the variant and its putative target gene(s).

[9]Non-synonymous missense.

[10]In Intron.

[11]In 5' gene region.

[12]In 3' gene region.

[13]In 3' UTR.

two other CSs, CS:11-88-3 and CS:11-88-4, near the gene *CCND1*. Within CS:11-88-3, 4 out of 17 variants are located in the 5' flanking region of *CCND1*. In CS:11-88-4, the most likely causal variant is the lead SNP rs3212870 ($P = 1.5 \times 10^{-3}$), which is located intronic in *CCND1*. The associations between CS:11-88-4 and PrCa have not been previously reported, because of the weak marginal associations, which can be explained by the moderate LD between CS:11-88-1, CS:11-88-2, and CS:11-88-4 (Supplementary Fig. 6a). Another interesting example is chr4 73256856-74885359 (Fig. 2b, Supplementary Fig. 6b), where we identified a novel CS, CS:4-88-3, insignificantly associated with PrCa (minimum $P = 6.5 \times 10^{-4}$). The lead SNP in CS:4-88-3, rs72649118, is a non-synonymous missense SNP of *RASSF6*, a member of the RASSF family of tumor suppressors [36].

Functional enrichment of prostate cancer credible causal variants

We used the hypergeometric tests to investigate the enrichment of credible causal variants (CCVs) in specific genomic features, including prostate-specific DNaseI hypersensitivity sites, ChIP-seq peaks of transcription factors and histone marks (Methods, Supplementary Table 3). We observed significant enrichment of CCVs in active regulatory regions (defined by H3K27ac and H3K4me1 marks), active gene promoters (defined by H3K4me3 and H3K9ac marks), actively transcribed gene bodies (defined by H3K36me3 and H3K79me2 marks), and DNaseI hypersensitivity sites (Fig. 3a). CCVs were also significantly enriched in the binding sites of various transcription factors (Fig. 3b, Supplementary Table 3), including *AR* (androgen receptor), *NR3C1* (glucocorticoid receptor), *ASH2L*, and *FOXA1*.

To formally evaluate the relationship between the biological functions associated with SNPs and their contributions to the risk of PrCa, we fitted a linear model for the logarithm of per-SNP heritability (i.e., the posterior mean of squared effect size) of all $1,342,667$ tag SNPs using the following functional annotations of SNPs as predictors: (i) 11 gene-based annotations extracted from the dbSNP database (build 151) [37]; (ii) cis- and trans-eQTLs within PrCa tissues from the TCGA database [38]; (iii) cis-eQTLs within normal prostate tissues from the GTEx v8 database [39]; (iv) DNaseI hypersensitivity sites, ChIP-seq peaks of 48 transcription factors and 9 histone modifications from normal prostate or prostate cancer cell lines, obtained from the Cistrome Data Browser [40]; (v) enhancer elements identified by Hi-C data and H3K27ac ChIP-seq peaks in normal prostate (RWPE1) and prostate cancer (C42B and 22Rv1) cell lines [35]; (vi) enhancer elements predicted by H3K27ac HiChIP in the prostate cancer cell line LNCaP [29]. In addition, $\log(f(1-f))$ and log(LD score) were included as covariates, where f is the minor allele frequency of SNP. This analysis revealed that cis-eQTL (TCGA) ($P_{\text{adj}} = 5.4 \times 10^{-137}$), cis-eQTL (GTEx) ($P_{\text{adj}} = 8.2 \times 10^{-52}$), and enhancer (H3K27ac HiChIP, LNCaP) ($P_{\text{adj}} = 1.8 \times 10^{-50}$) were the most significant three annotations associated with per-SNP heritability (Fig. 3c, Supplementary Table 4). Trans-eQTL (TCGA) ($P_{\text{adj}} = 1.5 \times 10^{-8}$) exhibited the largest effect size (0.36). Notably, HDAC1 (histone deacetylase 1) binding site was the only significant functional annotation with a negative effect on per-SNP heritability. These findings

suggested that genetic variants influencing gene expression levels and enhancer activity play a crucial role in the development and progression of PrCa.

Putative target genes of prostate cancer credible causal variants

To identify potential target genes of CCVs, we integrated various sources of information, including gene-based annotations from the dbSNP database (build 151), eQTL data, and enhancer-promoter interaction data from Hi-C and HiChIP experiments (Methods). As a result, we identified 385 protein-coding genes as potential target genes of CCVs across all 95% CSs (Fig. 4a, Supplementary Table 2).

We further conducted pathway enrichment analysis to gain insights into the biological functions and processes associated with these putative target genes. Our analysis revealed significant over-representation of these genes in 52 non-redundant pathways at an FDR of 0.05 (Fig. 4b, Supplementary Table 5). Notable enriched pathways included prostate gland development, DNA damage response (only ATM dependent), positive regulation of transcription by RNA polymerase II, and regulation of mitotic cell cycle. The enrichment of putative target genes in cellular response to BMP (bone morphogenetic protein) stimulus, collagen catabolic process, and definitive hemopoiesis pathways may be attributed to the involvement of these processes in PrCa bone metastasis [41–43]. Furthermore, putative target genes were also over-represented in toxin transport pathway. Although previous studies have reported associations between PrCa and several genes in toxin transport pathway, such as *SLC22A1–A3* [44, 45], the relationship between PrCa and this pathway is not well elucidated and needs further investigation.

Discussion

In this article, we present h2-D2, a fine-mapping method that utilizes a continuous global-local shrinkage prior. As an extension of R2-D2, h2-D2 is designed for GWAS data where the phenotype values are standardized. Unlike existing fine-mapping methods that rely on discrete mixture priors, h2-D2 does not impose a constraint on the maximum number of causal variants and allows for the exploration of a wider range of causal configurations. In addition, h2-D2 does not reply on assumptions regarding the distribution of causal variant effect sizes, compatible with infinitesimal effect assumption for non-causal variants, which has been adopted by some recent works in fine-mapping [46, 47]. These features ensure the applicability and flexibility of h2-D2 in various scenarios.

We develop an efficient MCMC algorithm for h2-D2 to sample from the posterior distribution. We utilize several strategies to accelerate the mixing of MCMC chains, allowing for a more extensive exploration of the model space. Simulation studies show that h2-D2 is less likely to get trapped into local optima and performs better in variable selection than discrete-mixture-prior-based methods including SuSiE and FINEMAP. This may be due to the property of continuous priors that the coefficients are updated continuously and the transitions among local modes can be smoothly. Our results also highlight the importance of using accurate LD matrices derived from adequately large reference panels, which concurs with previous discoveries [12, 48].

Another important contribution of our work is that we propose an inference approach to define credible sets in the framework of continuous priors, which addresses the limitation of continuous priors that do not yield selection results directly. Simulation studies show that the CSs produced by h2-D2 can achieve the target level of coverage and are well-powered when using in-sample LD matrices, and exhibit an improved control of FDR when using mismatched LD matrices. These results suggest the robustness and effectiveness of our proposed approach. Theoretical properties of the credible level defined for multiple SNPs deserve further investigation. Additionally, we acknowledge that the greedy search algorithm used to identify credible sets may not always yield the optimal sets and may miss some sets (Supplementary Note 3). Further refinement and improvement of the algorithm are needed to enhance its performance.

In the real data application on prostate cancer GWAS, we identified several novel signals that have not been previously reported. Variants in 95 % CSs are significantly over-represented in prostate-specific epigenetic marks associated with activation of gene transcription. Through integrating gene-based annotation of SNPs, eQTL, Hi-C, and HiChIP data in prostate cell lines, we identified 385 potential target genes of variants in 95 % CSs. These genes are enriched in prostate development and cancer related pathways.

As a future direction, fine-mapping resolutions may be improved by integrating functional annotations into the h2-D2 prior. Stratified LD score regression-based methods like PolyFun [49] are well suited to be incorporated with h2-D2, since h2-D2 prior is imposed on the per-SNP heritability directly. Furthermore, h2-D2 can also be extended to multi-trait fine-mapping. Given the widespread existence of pleiotropy, fine-mapping multiple traits simultaneously has the potential to enhance the power of identifying shared causal variants among traits [50–52]. Jointly analyzing multiple traits may provide valuable insights into the genetic architecture underlying complex diseases and traits, and improve our understanding of the shared genetic basis between different phenotypes.

Methods

h2-D2 prior

For a GWAS of quantitative trait with N individuals, consider a region containing M variants. The relationship between phenotypes and genotypes can be modeled by a multiple linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a vector of standardized phenotype values for N individuals, $\boldsymbol{X}$ is an $N \times M$ column-standardized genotype matrix for N individuals and M variants, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\top$ is an M-vector of effect sizes to be estimated, and $\boldsymbol{\varepsilon}$ is an N-vector of error terms. We assume that $\boldsymbol{\varepsilon} \sim \mathrm{N}_N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_N)$, where $\boldsymbol{I}_N$ is an $N \times N$ identity matrix and $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ denotes the k-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$.

We introduce a prior for $\boldsymbol{\beta}$ satisfying $E(\boldsymbol{\beta}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is an $M \times M$ diagonal matrix with diagonal elements $\sigma_1^2, \ldots, \sigma_M^2$. The narrow-sense heritability h^2 of the quantitative trait explained by the M SNPs can be expressed as

$$\begin{aligned} h^2 =& \frac{\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta})}{\mathrm{Var}(\boldsymbol{y})} = \mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta}) \\ =& \mathrm{E}\left(\boldsymbol{\beta}^\top \boldsymbol{R}\boldsymbol{\beta}\right) = \mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}) = \sum_{j=1}^{M} \sigma_j^2 \leq 1, \end{aligned} \tag{2}$$

where $\boldsymbol{R}$ is the linkage disequilibrium (LD) matrix of the M variants. Then σ_j^2 can be interpreted as the per-variant heritability of variant j.

To achieve an ideal prior that shrinks most elements of $\boldsymbol{\beta}$ toward 0 while retaining some large coefficients, we impose a Dirichlet prior on the variance terms:

$$\left(\sigma_1^2, \ldots, \sigma_M^2, 1 - h^2\right) \sim \mathrm{Dir}\left(a_1, \ldots, a_M, b\right), \tag{3}$$

where $a_1, \ldots, a_M \in (0, 1)$ and $b > 0$ are hyperparameters. Additionally, a double-exponential prior is assigned to each element of $\boldsymbol{\beta}$:

$$\beta_j \big| \sigma_j^2 \sim \mathrm{DE}\left(\sqrt{\sigma_j^2/2}\right), \quad j = 1, \ldots, M, \tag{4}$$

where $\mathrm{DE}(\delta)$ denotes a double-exponential distribution with mean 0 and variance $2\delta^2$.

Same as many other fine-mapping methods, h2-D2 requires GWAS summary data only [53]. Assume the single-SNP summary statistics $D = \{\widehat{\beta}_j, \widehat{e}_j\}_{j=1}^M$ are provided, where $\widehat{\beta}_j$ is the marginal effect size estimate of SNP j, and $\widehat{e}_j$ is its standard error. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_M)^\top$, $\widehat{s}_j = \left(\widehat{e}_j^2 + N^{-1}\widehat{\beta}_j^2\right)^{1/2}$ for $j = 1, \ldots, M$, and $\widehat{\boldsymbol{S}} = \mathrm{diag}(\widehat{s}_1, \ldots, \widehat{s}_M)$. The LD matrix is estimated from some reference panel as $\widehat{\boldsymbol{R}}$. The RSS likelihood of $\widehat{\boldsymbol{\beta}}$ [53] is given by

$$\widehat{\boldsymbol{\beta}} | \boldsymbol{\beta}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{R}} \sim \mathrm{N}_M\left(\widehat{\boldsymbol{S}}\widehat{\boldsymbol{R}}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{\beta}, \widehat{\boldsymbol{S}}\widehat{\boldsymbol{R}}\widehat{\boldsymbol{S}}\right). \tag{5}$$

The h2-D2 prior can also be applied to binary traits by considering the observed-scale heritability (Supplementary Note 1). An MCMC algorithm that is compatible with both quantitative traits and binary traits is developed to obtain samples from the posterior distribution (Supplementary Note 2).

Credible level and credible set

For the j-th SNP, consider the null hypothesis $H_{0j} : \beta_j = 0$. In the frequentist framework, H_{0j} can be rejected at the level of α if 0 is not contained in a confidence interval at the level of $1-\alpha$. We migrate this approach to the Bayesian framework by replacing the confidence interval with the Bayesian credible interval. We propose the following

statistic to evaluate how likely SNP j is causal:

$$\mathrm{CL}_j \triangleq \left|\widehat{\Pr}\left(\beta_j > 0|D\right) - \widehat{\Pr}\left(\beta_j < 0|D\right)\right| \in [0,1], \tag{6}$$

where the posterior probability $\widehat{\Pr}(\cdot|D)$ is estimated from the MCMC samples. We term this statistic as the "credible level" of SNP j, since it can be interpreted as the maximum probability such that the corresponding equal-tailed credible interval of β_j doesn't cover 0.

Next, we extend this concept to multiple SNPs and define credible sets (CSs) accordingly. Consider a set of SNPs $\mathcal{C} = \{j_1, \ldots, j_k\}$. In the frequentist framework, claiming that $\mathcal{C}$ is a level $1-\alpha$ CS is equivalent to rejecting the null hypothesis $H_{0\mathcal{C}} : \boldsymbol{\beta}_{\mathcal{C}} \triangleq (\beta_{j_1}, \ldots, \beta_{j_k})^\top = \mathbf{0}$ at the significance level of α, which can be declared if the null hypothesis $\boldsymbol{v}^\top \boldsymbol{\beta}_{\mathcal{C}} = 0$ can be rejected at the significance level of α for at least one $\boldsymbol{v} \in \mathbb{R}^k$. Therefore, the credible level of $\mathcal{C}$ is defined as

$$\begin{aligned} \widetilde{\mathrm{CL}}_{\mathcal{C}} &\triangleq \max_{\boldsymbol{v}\in\mathbb{R}^k} \mathrm{CL}_{\mathcal{C}}(\boldsymbol{v}) \\ &= \max_{\boldsymbol{v}\in\mathbb{R}^k} \left|\widehat{\Pr}\left(\boldsymbol{v}^\top \boldsymbol{\beta}_{\mathcal{C}} > 0|D\right) - \widehat{\Pr}\left(\boldsymbol{v}^\top \boldsymbol{\beta}_{\mathcal{C}} < 0|D\right)\right|. \end{aligned} \tag{7}$$

However, it is computationally infeasible to find the $\boldsymbol{v}$ that maximize the credible level when the number of variants exceeds 2. Instead, we choose a single $\boldsymbol{v}$ to provide a lower bound of the credible level. The coefficients of $\boldsymbol{v}$ are selected to satisfy the condition that for positively correlated variants in $\mathcal{C}$, the corresponding coefficients have the same signs, while for negatively correlated variants, the corresponding coefficients have different signs. We choose $\boldsymbol{v}$ as an eigenvector of $\widehat{\boldsymbol{R}}_{\mathcal{C}}$ corresponding to its largest eigenvalue, where $\widehat{\boldsymbol{R}}_{\mathcal{C}}$ denotes the estimated LD matrix of SNPs in $\mathcal{C}$. If $\mathrm{CL}_{\mathcal{C}}(\boldsymbol{v}) \geq 1-\alpha$, we have $\widetilde{\mathrm{CL}}_{\mathcal{C}} \geq 1-\alpha$, and we conclude that $\mathcal{C}$ is a level $1-\alpha$ credible set. A greedy algorithm is designed to search all CSs achieving a pre-specified level (Supplemental Note 3).

Choice of hyper-parameters

In the Dirichlet prior (3), a smaller a_j leads to a higher concentration around 0 for β_j, while a larger b indicates stronger global shrinkage. When incorporating external information, such as functional annotations, if the j-th SNP is more likely to be causal, a larger a_j can be set. By default, we suggest setting $a_1 = \ldots = a_M = a \in [0.001, 0.01]$ for general fine-mapping tasks. A smaller a would make the MCMC chain converge slowly, while a larger a can be considered if there are evidences that the region may harbor a large number of causal variants (e.g. more than 10).

As for the choice of b, if an in-sample or highly accurate LD matrix $\widehat{\boldsymbol{R}}$ is available, we recommend estimating the local heritability using some well-known estimation procedures, such as the HESS estimator [54], which is defined as:

$$\widehat{h^2} = \frac{N\widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{\beta}} - M}{N - M}. \tag{8}$$

Then, b can be chosen as follows:

$$b = \frac{\left(1 - \widehat{h^2}\right) \sum_{j=1}^{M} a_j}{\widehat{h^2}}. \tag{9}$$

However, if the accuracy of $\widehat{\boldsymbol{R}}$ is poor, the HESS estimator may exhibit large bias. In this scenario, even if the true heritability is known, setting b according to Equation (9) can lead to large effect size estimates for some non-causal variants in h2-D2. This is consistent with a recent finding that significant miscalibration due to external LD matrices can produce suspicious results in meta-analysis fine-mapping studies [55]. To address this, we suggest performing quality-control to filter out outlier variants before fine-mapping, and setting $b \in \left[10^4 \sum_{j=1}^{M} a_j, 2 \times 10^5 \sum_{j=1}^{M} a_j\right]$ for GWAS fine-mapping tasks or setting $b \in \left[10 \sum_{j=1}^{M} a_j, 200 \sum_{j=1}^{M} a_j\right]$ for eQTL fine-mapping tasks.

UK Biobank data preprocessing

We selected British individuals from the UK Biobank database based on specific criteria. The selection process involved the following steps: (i) Only individuals with available genotype data were included. (ii) We specifically chose individuals who self-identified as "White British" to ensure homogeneity in the population. (iii) Genetic sex was confirmed to be consistent with self-reported sex. (iv) Outlier individuals were identified and excluded based on heterozygosity or missingness. (v) Individuals with close familial relationships were removed to avoid any potential bias in the analysis. After applying these filtering criteria, a total of $275,768$ individuals were retained for further analysis.

Subsequently, we focused on variants that met the following criteria: (i) Variants with at most one alternative allele were considered to ensure simplicity in the analysis. (ii) Variants with a minor allele frequency of at least 1% were selected to ensure a reasonable frequency of the variant in the population. (iii) Variants with an information score of at least 0.8 were included to ensure high-quality genotype data.. The rsID of selected variants were labeled based on dbSNP database (build 151).

Partition LD blocks

We noticed that the LD blocks partitioned by LDetect [56] based on 1000 Genome reference panel are not optimal for UKBB reference panel. We developed a method to divide the whole genome into nearly independent LD blocks, so as to improve computational efficiency and achieve accurate fine-mapping results.

For a given LD matrix $\widehat{\boldsymbol{R}}$ of M SNPs, we defined the optimal splitting as the solution to the following optimization problem:

$$\underset{k \in \{1,\ldots,M\}}{\arg\min} \frac{\sum_{1 \le j_1 \le k, k < j_2 \le M} r_{j_1 j_2}^2}{k(M-k)},$$

i.e., minimizing the average squared correlation r^2 between two blocks. Our algorithm iteratively identifies optimal splitting points between consecutive LD blocks obtained from LDetect. If the loss in optimal splitting, defined as the difference in the objective function value before and after the split, is smaller than 0.001 and the size of the split block is not smaller than 50, the split point is accepted. This process is performed recursively for each split block until no further split points satisfying the conditions can be found.

As a result, we divided the entire autosomal region (excluding the major histocompatibility complex [MHC] regions) into a total of $3,717$ nearly independent LD blocks. We provide the script and the full list of LD blocks on our GitHub repository at https://github.com/xiangli428/PrCaFineMapping. This approach allows for improved efficiency and accuracy in fine-mapping analyses using the UK Biobank reference panel.

Simulations

We conducted simulation studies using UK Biobank imputed genotype data from $N = 275,768$ unrelated British individuals [26]. For the simulations, we selected 100 nearly independent LD blocks on chromosome 2 (Supplementary Table 1), and included variants with MAF ≥ 0.01 and INFO score ≥ 0.8. We pruned SNPs such that the absolute correlation $|r|$ between any two SNPs was less than 0.99. Each block contained a varying number of SNPs, ranging from 288 to $1,122$, and had a length between 0.25 and 2.

We designed four simulation scenarios with varying sample sizes, local heritabilities, and numbers causal variants. Genotype values of each SNP were standardized. In each scenario, for each block, the causal variants were chosen randomly and the effect sizes of causal variants were sampled from a normal distribution with mean 0. The phenotype values were then computed according to the multiple regression model (1), where the error term $\boldsymbol{\varepsilon}$ were sampled from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_\varepsilon^2 \boldsymbol{I}_N$. σ_ε^2 was chosen such that $\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta}) / \left(\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta}) + \sigma_\varepsilon^2\right)$ equaled h^2 in each scenario. After standardizing the phenotype values and scaling the effect sizes of causal variants consistently, we computed summary statistics for each variant.

To assess the influence of LD matrix accuracy on the fine-mapping performance, we computed four LD matrics for each block. The first one was an in-sample LD matrix computed from all $275,768$ UKBB individuals ($\widehat{\boldsymbol{R}}$). The second one and the third one were down-sample LD matrices, computed from randomly sampled $3,000$ or 500 UKBB individuals, denoted by $\widehat{\boldsymbol{R}}_{\mathrm{UKBB,3000}}$ and $\widehat{\boldsymbol{R}}_{\mathrm{UKBB,500}}$, respectively. The fourth one was an out-of-sample LD matrix computed from 522 unrelated European ancestry individuals using the genotype data from the $1,000$ Genomes Project on GRCh38 [57, 58], denoted by $\widehat{\boldsymbol{R}}_{\mathrm{1KG}}$. When using mismatched LD matrices, we applied SLALOM to all pairs of SNPs with $|r| \geq 0.8$ and remove outlier non-causal variants with DENTIST-S statistics ≥ 40 [55].

Compared methods

We performed a comprehensive comparison between h2-D2 and two state-of-art fine-mapping methods requiring only summary statistics, FINEMAP [10, 11] and SuSiE-RSS [12]. FINEMAP utilizes a general discrete distribution as prior for the number of causal SNPs,

$$\Pr(\text{number of causal SNPs is } k) = p_k, k = 1, \ldots, K,$$

where $K \ll M$ is the maximum number of causal variants, and uses a shotgun stochastic search algorithm to identify models with high posterior probabilities. SuSiE [5] is a novel variable selection method that decomposes the effect size as the sum of single-effect vectors and imposes a multinomial prior distribution on each single-effect vector. SuSiE adopts an iterative Bayesian stepwise selection algorithm to optimize a variational approximation to the posterior distribution, as well as a refinement procedure to address the convergence problem of the algorithm.

As for the choices of hyper-parameters, for h2-D2, we set $a_1 = \ldots = a_M = a = 0.005$. When using LD matrices $\widehat{\boldsymbol{R}}$ or $\widehat{\boldsymbol{R}}_{\text{UKBB},3000}$, we set b according to Equations (8) and (9). When using $\widehat{\boldsymbol{R}}_{\text{UKBB},500}$ or $\widehat{\boldsymbol{R}}_{\text{1KG}}$, we set $b = 2 \times 10^5 Ma$ for scenario 1-3, and $100Ma$ for scenario 4, respectively. For SuSiE, we set options "refine=TRUE" and "estimate_residual_variance=TRUE". For both SuSiE and FINEMAP, we set the number of single-effect vectors or the maximum number of causal variants equals the true number of causal variants in each scenario (5 or 10).

Comparison of causal variant effect sizes and their posterior mean in simulation studies

In each simulation setting, we merged the results from 100 datasets together. Since causal variants with small effect sizes are difficult to be identified by fine-mapping methods, we used the following piecewise linear model to assess the relationship between the true effect sizes (β) of causal SNPs and their posterior means ($\overline{\beta}$):

$$\overline{\beta} = \begin{cases} k\,(\beta - \beta_0) & \text{for } \beta > \beta_0, \\ 0 & \text{for } |\beta| \leq \beta_0, \\ k\,(\beta + \beta_0) & \text{for } \beta < -\beta_0, \end{cases} \tag{10}$$

where k and $\beta_0 > 0$ are the coefficients to be estimated. We used least square method to estimate these coefficients. The fitted curves are shown in Supplementary Fig. 4.

Prostate cancer GWAS data preprocessing

The prostate cancer GWAS summary statistics were obtained from a meta-analysis involving $79,148$ cases and $61,106$ controls of European ancestry. We conducted the following QC process for the PrCa GWAS summary data.

- Duplicated SNPs and SNPs with imputation $r^2 < 0.3$ or MAF$= 0$ were filtered out.

- Effect sizes and standard errors on the allelic scale were transformed to standardized dosage scale by multiplying them with $\sqrt{2\text{MAF}(1-\text{MAF})}$. Standard errors of effect sizes should be closed to $\sqrt{N/(N_1 N_0)} \approx 5.4 \times 10^{-3}$, where N is the total sample size, N_1 is the number of cases, and N_0 is the number of controls. We filtered out SNPs with standard errors $< 5 \times 10^{-3}$ or $> 10^{-2}$.
- The remaining SNPs were compared with UKBB reference panel (Methods). Only SNPs that were present in UKBB reference panel with a logit(MAF) difference less than 0.5 were retained for the fine-mapping analysis.
- Since mismatched LD matrices were used, to avoid unreliable results, for each pair of SNPs with an absolute correlation $|r| \geq 0.8$, we checked if the pattern of LD and GWAS summary statistics is suspicious using DENTIST-S statistic [55]. If DENTIST-S statistic was greater than or equal to 30, the less significant SNP would be removed.

After the QC process, we had a total of $6,446,747$ common SNPs for the fine-mapping analysis. For each LD block, before fine-mapping, the variants were pruned such that all pairwise absolute correlation were less than 0.95. This resulted in $1,342,667$ tag SNPs that were retained for the fine-mapping analysis. We used h2-D2 with specific hyper-parameters ($a_1 = \ldots a_M = 0.005$ and $b = 2 \times 10^5$) to fine-map each region and compute 95% CSs. Each 95% CS includes a set of tag SNPs with a joint credible level ≥ 0.95, as well as the pruned SNPs that are in high LD with them.

Annotations of variants

The gene-based annotations of variants and their associated genes were extracted from the dbSNP database (build 151) with GRCh37.p13 as the reference assembly. These annotations include: NSF (non-synonymous frameshift), NSM (non-synonymous missense), NSN (non-synonymous nonsense), SYN (synonymous), U3 (in 3' UTR), U5 (in 5' UTR), ASS (in acceptor splice site), DSS (in donor splice-site), INT (in intron), R3 (in 3' gene region), and R5 (in 5' gene region).

Prostate cancer-specific cis- and trans-eQTL data were obtained from PancanQTL database (http://gong_lab.hzau.edu.cn/PancanQTL/). Cis-eQTLs from normal prostate tissues mapped in European-American subjects were obtained from GTEx V8 database (https://gtexportal.org/home/).

DNaseI peaks, ChIP-seq peaks of histone marks and transcription factor binding sites in prostate-derived cell lines were obtained from Cistrome database (http://cistrome.org/). Details of downloaded data are shown in Supplementary Table 3. The peak coordinates were converted from hg38 to hg19 reference assembly using LiftOver. Variants located within these peaks were selected using BEDTools.

Enhancer-promoter loops identified from Hi-C data in RWPE1, C42B, and 22Rv1 cell lines were obtained from Supplementary Table 5A-C of ref. [35]. Annotated H3K27ac HiChIP loops in LNCaP cell line were obtained from Table S7 of ref. [29]. Variants located within the identified enhancers were selected using BEDTools.

Pathway enrichment analysis

Potential target genes of credible causal variants (CCVs) were derived by merging (1) associated genes of CCVs annotated in dbSNP database (build 151); (2) associated genes of eQTLs in CCVs in PancanQTL and GTEx V8 databases; (3) genes whose promoters interact with enhancers covering CCVs in Hi-C or H3K27ac HiChIP data. Protein-coding genes were retained based on GENCODE v42 annotations mapped to GRCh37 assembly.

Enrichment analyses for pathways from GO Biological Process [59] and WikiPathways [60] were carried out using GeneCodis [61] (https://genecodis.genyo.es/). To remove redundant pathways, we computed Dice coefficients for all pairs of pathways. If the Dice coefficient between two pathways is larger than 0.3, only the more significant one was retained.

Data availablity

UK Biobank data are accessed under Application Number 58942 (https://www.ukbiobank.ac.uk/). 1000 Genomes on GRCh38 data are publicly available at https://www.internationalgenome.org/. Prostate cancer summary data are available from the PRACTICAL Consortium (http://practical.icr.ac.uk/blog/?page_id=8164). The complete contents of dbSNP database (build 151) with GRCh37.p13 as reference assembly are available at https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/. Prostate cancer-specific cis- and trans-eQTLs within PrCa tissues from the TCGA database are available at PancanQTL database http://gong_lab.hzau.edu.cn/PancanQTL/. Cis-eQTLs from normal prostate tissues mapped in European-American subjects are available at GTEx V8 database https://gtexportal.org/home/. DNaseI peaks, ChIP-seq peaks of histone marks and transcription factor binding sites in prostate-derived cell lines are available at Cistrome database http://cistrome.org/. Enhancer-promoter loops identified from Hi-C data in RWPE1, C42B, and 22Rv1 cell lines are available at https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-12079-8/MediaObjects/41467_2019_12079_MOESM7_ESM.xlsx. Annotated H3K27ac HiChIP loops in LNCaP cell line are available at https://ars.els-cdn.com/content/image/1-s2.0-S0002929721004195-mmc3.csv.

Code availability

The software h2D2 is available at https://github.com/xiangli428/h2D2. Scripts and data related to PrCa fine-mapping analysis are available at https://github.com/xiangli428/PrCaFineMapping.

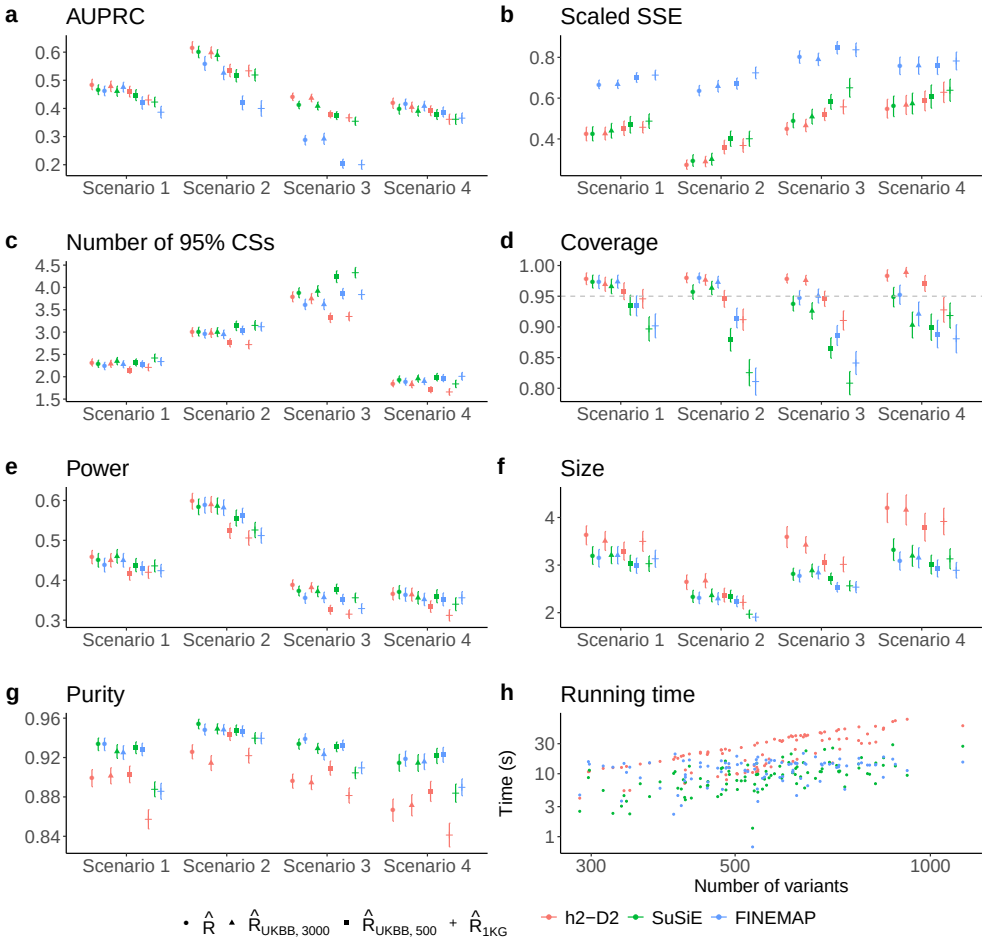

Fig. 1 Performance comparison of h2-D2, SuSiE and FINEMAP on simulated data. In **a-g**, all values are the average ones across 100 datasets, with standard errors indicated by the error bars. **a** Area under the precision-recall curve (AUPRC) based on the credible level of each SNP for h2-D2 or the marginal posterior inclusion probability (PIP) of each SNP for SuSiE and FINEMAP. **b** Sum of squared error (SSE) of $\boldsymbol{\beta}$ based on its posterior mean, scaled by h^2 in each scenario. **c** Number of detected 95% credible sets (CSs). **d** Coverage of 95% CS (the proportion of CSs that capture at least one causal variant). **e** Power of 95% CS (the proportion of causal variants captured by at least one CS). **f** Size of 95% CS (the number of variants in each CS). **g** Purity of 95% CS (the minimum absolute correlation among all pairs of SNPs in each CS). **h** Running time of the three methods against the number of variants in scenario 1. Each point represents a simulated dataset. For h2-D2, MCMC ran 10,000 iterations. For SuSiE, $L = 5$. For FINEMAP, $K = 5$.

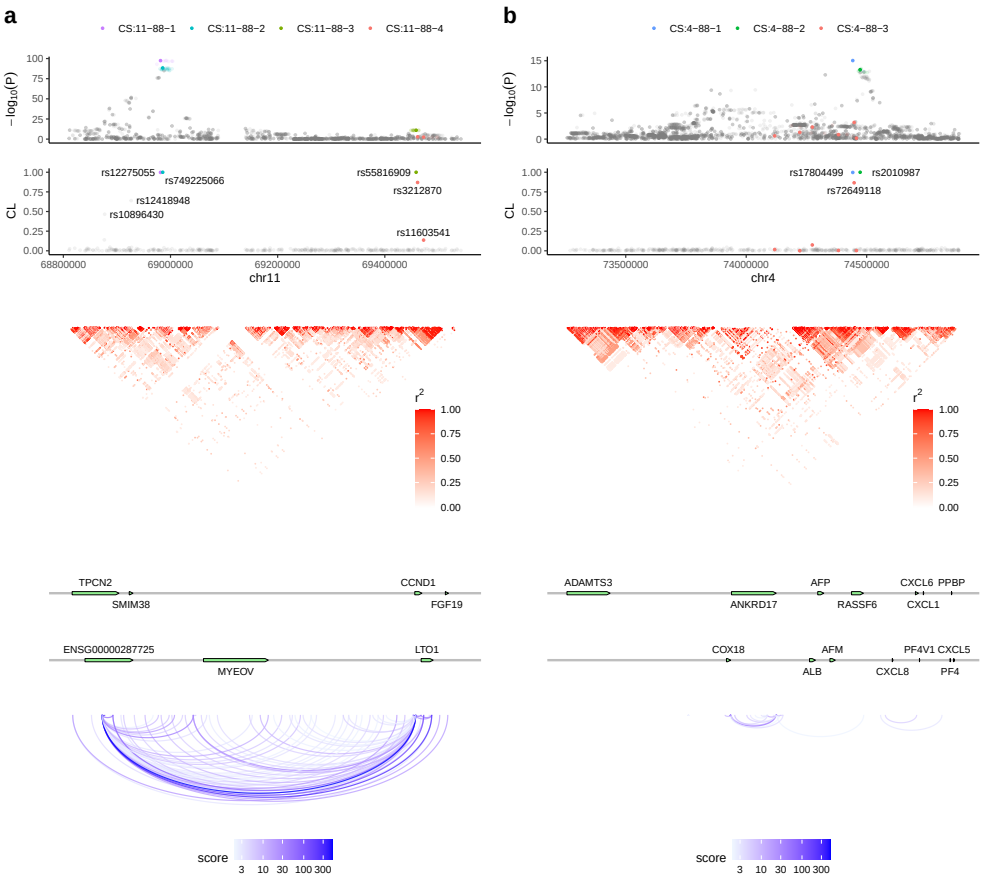

Fig. 2 Fine-mapping results of two genomic regions in prostate cancer data analysis. a chr11 68810837-69542062; **b** chr4 73256856-74885359. The top panel depicts the marginal associations of variants ($-\log_{10}(P)$) from the GWAS meta-analysis data. The second panel illustrates the credible levels of tag SNPs computed by h2-D2. In the first two panels, each color represents a 95% credible set (CS). The CS is named in the format CS:{chromosome ID}-{region ID}-{index}. The third panel demonstrates the patterns of linkage disequilibrium of the genomic region. The fourth panel displays the positions of genes in the corresponding regions. The bottom panel shows the H3K27ac HiChIP loops detected in the LNCaP prostate cancer cell line [29].

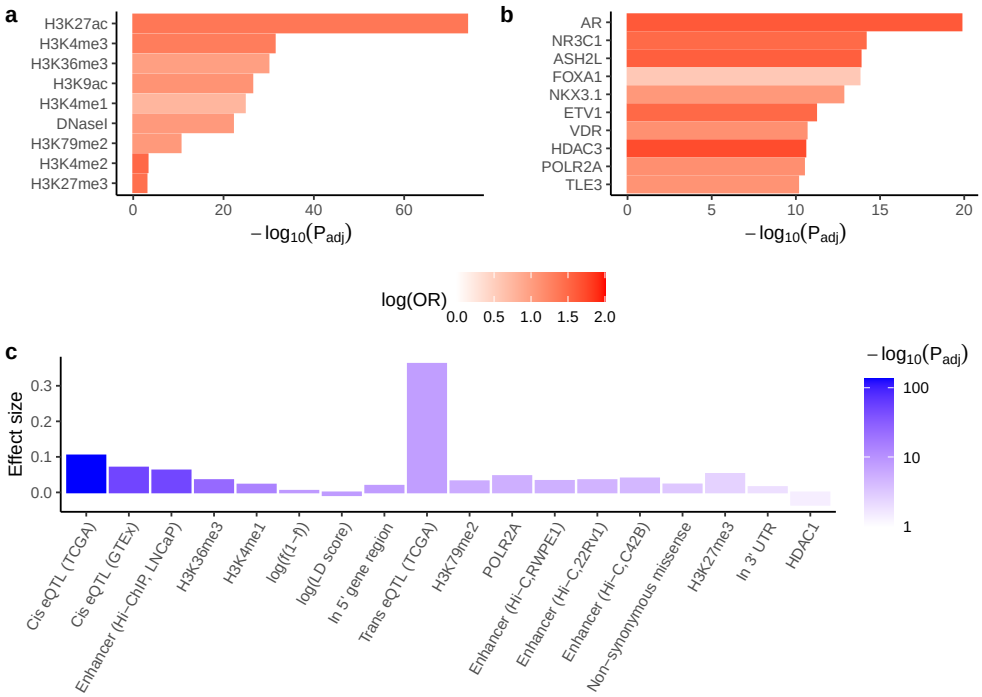

Fig. 3 **a**,**b** Enrichment of credible causal variants in prostate-specific (**a**) histone marks and DNaseI hypersensitivity sites (**b**) top 10 transcription factor binding sites. Hypergeometric test P values are adjusted using the Benjamini-Hochberg (BH) method. **c** A linear regression model is fitted for the logarithm of per-SNP heritabilities of tag SNPs using the functional annotations of tag SNPs as predictors. Effect sizes and adjusted P values of significant functional annotations are shown. P values are adjusted using the BH method. Significance is defined as $P_{\mathrm{adj}} < 0.05$.

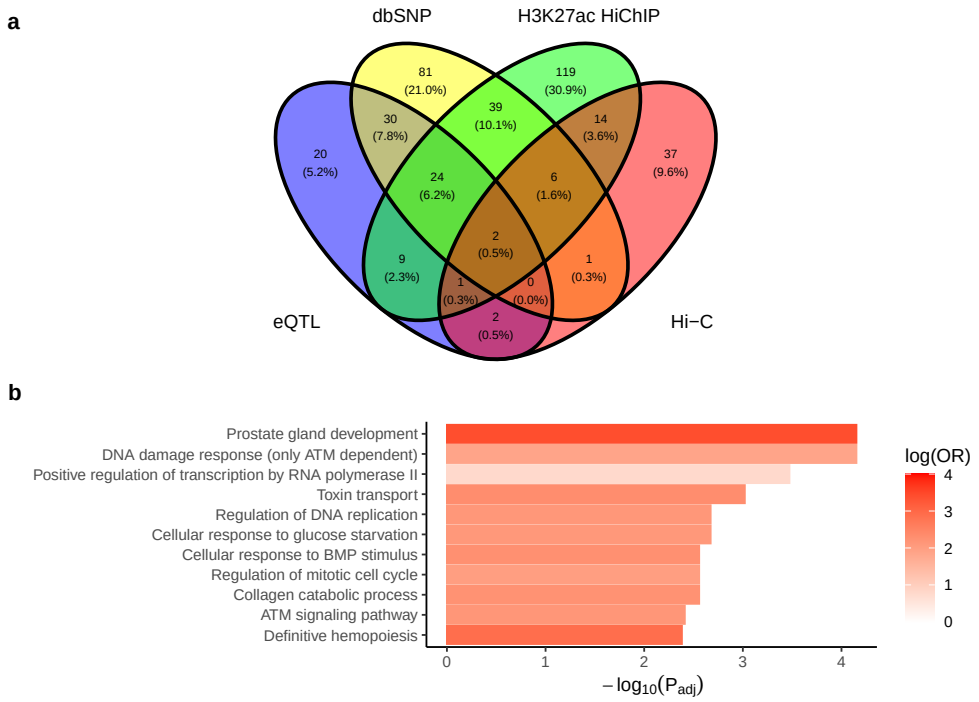

Fig. 4 **a** Venn diagram showing the numbers of putative target genes inferred from different sources of information. **b** Enrichment of putative target genes in pathways from Gene Ontology Biological Processes and WikiPathways. Hypergeometric test P values are adjusted using the BH method. Pathways with $P_{\mathrm{adj}} < 0.005$ are shown.

References

[1] Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.*: The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research **47**(D1), 1005–1012 (2019)

[2] Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E.: Identifying causal variants at loci with multiple signals of association. Genetics **198**(2), 497–508 (2014)

[3] Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., Pasaniuc, B.: Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS genetics **10**(10), 1004722 (2014)

[4] Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)

[5] Wang, G., Sarkar, A., Carbonetto, P., Stephens, M., *et al.*: A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society Series B **82**(5), 1273–1300 (2020)

[6] Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., *et al.*: Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature genetics **44**(12), 1294–1301 (2012)

[7] Chen, W., Larrabee, B.R., Ovsyannikova, I.G., Kennedy, R.B., Haralambieva, I.H., Poland, G.A., Schaid, D.J.: Fine mapping causal variants with an approximate bayesian method using marginal test statistics. Genetics **200**(3), 719–736 (2015)

[8] Newcombe, P.J., Conti, D.V., Richardson, S.: Jam: a scalable bayesian framework for joint analysis of marginal snp effects. Genetic epidemiology **40**(3), 188–201 (2016)

[9] Wen, X., Lee, Y., Luca, F., Pique-Regi, R.: Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. The American Journal of Human Genetics **98**(6), 1114–1129 (2016)

[10] Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S., Pirinen, M.: Finemap: efficient variable selection using summary data from genome-wide association studies. Bioinformatics **32**(10), 1493–1501 (2016)

[11] Benner, C., Havulinna, A.S., Salomaa, V., Ripatti, S., Pirinen, M.: Refining fine-mapping: effect sizes and regional heritability. BioRxiv, 318618 (2018)

[12] Zou, Y., Carbonetto, P., Wang, G., Stephens, M.: Fine-mapping from summary data with the "sum of single effects" model. PLoS genetics **18**(7), 1010299 (2022)

[13] Park, T., Casella, G.: The bayesian lasso. Journal of the American Statistical Association **103**(482), 681–686 (2008)

[14] Griffin, J.E., Brown, P.J.: Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis **5**(1), 171–188 (2010)

[15] Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika **97**(2), 465–480 (2010)

[16] Armagan, A., Dunson, D.B., Lee, J.: Generalized double pareto shrinkage. Statistica Sinica **23**(1), 119 (2013)

[17] Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet–laplace priors for optimal shrinkage. Journal of the American Statistical Association **110**(512), 1479–1490 (2015)

[18] Bhadra, A., Datta, J., Polson, N.G., Willard, B.: The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis **12**(4), 1105–1131 (2017)

[19] Bai, R., Ghosh, M.: Large-scale multiple hypothesis testing with the normal-beta prime prior. Statistics **53**(6), 1210–1233 (2019)

[20] Zhang, Y.D., Naughton, B.P., Bondell, H.D., Reich, B.J.: Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. Journal of the American Statistical Association **117**(538), 862–874 (2020)

[21] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., Smoller, J.W.: Polygenic prediction via bayesian regression and continuous shrinkage priors. Nature communications **10**(1), 1–10 (2019)

[22] ISHWARAN, H., RAO, J.S.: Spike and slab variable selection: Frequentist and bayesian strategies. The Annals of statistics **33**(2), 730–773 (2005)

[23] Bondell, H.D., Reich, B.J.: Consistent high-dimensional bayesian variable selection via penalized credible regions. Journal of the American Statistical Association **107**(500), 1610–1624 (2012)

[24] Zhang, Y., Bondell, H.D.: Variable selection via penalized credible regions with dirichlet-laplace global-local shrinkage priors. Bayesian Analysis (2018)

[25] Hahn, P.R., Carvalho, C.M.: Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. Journal of the American Statistical Association **110**(509), 435–448 (2015)

[26] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer,

A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.*: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)

[27] Schumacher, F.R., Al Olama, A.A., Berndt, S.I., Benlloch, S., Ahmed, M., Saunders, E.J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., *et al.*: Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nature genetics **50**(7), 928–936 (2018)

[28] Dadaev, T., Saunders, E.J., Newcombe, P.J., Anokian, E., Leongamornlert, D.A., Brook, M.N., Cieza-Borrella, C., Mijuskovic, M., Wakerell, S., Olama, A.A.A., *et al.*: Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. Nature communications **9**(1), 1–19 (2018)

[29] Giambartolomei, C., Seo, J.-H., Schwarz, T., Freund, M.K., Johnson, R.D., Spisak, S., Baca, S.C., Gusev, A., Mancuso, N., Pasaniuc, B., *et al.*: H3k27ac hichip in prostate cell lines identifies risk genes for prostate cancer susceptibility. The American Journal of Human Genetics **108**(12), 2284–2300 (2021)

[30] Al Olama, A.A., Kote-Jarai, Z., Giles, G.G., Guy, M., Morrison, J., Severi, G., Leongamornlert, D.A., Tymrakiewicz, M., Jhavar, S., Saunders, E., *et al.*: Multiple loci on 8q24 associated with prostate cancer susceptibility. Nature genetics **41**(10), 1058–1060 (2009)

[31] Wahlström, G., Heron, S., Knuuttila, M., Kaikkonen, E., Tulonen, N., Metsälä, O., Löf, C., Ettala, O., Boström, P.J., Taimen, P., *et al.*: The variant rs77559646 associated with aggressive prostate cancer disrupts ano7 mrna splicing and protein expression. Human Molecular Genetics **31**(12), 2063–2077 (2022)

[32] Srinivasan, S., Stephens, C., Wilson, E., Panchadsaram, J., DeVoss, K., Koistinen, H., Stenman, U.-H., Brook, M.N., Buckle, A.M., Klein, R.J., *et al.*: Prostate cancer risk-associated single-nucleotide polymorphism affects prostate-specific antigen glycosylation and its function. Clinical chemistry **65**(1), 1–9 (2019)

[33] Rashkin, S.R., Graff, R.E., Kachuri, L., Thai, K.K., Alexeeff, S.E., Blatchins, M.A., Cavazos, T.B., Corley, D.A., Emami, N.C., Hoffman, J.D., *et al.*: Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nature communications **11**(1), 4423 (2020)

[34] Sun, W., Yue, J.: Tpc2 mediates autophagy progression and extracellular vesicle secretion in cancer cells. Experimental cell research **370**(2), 478–489 (2018)

[35] Rhie, S.K., Perez, A.A., Lay, F.D., Schreiner, S., Shi, J., Polin, J., Farnham, P.J.: A high-resolution 3d epigenomic map reveals insights into the creation of the prostate cancer transcriptome. Nature communications **10**(1), 4154 (2019)

[36] Allen, N., Donninger, H., Vos, M., Eckfeld, K., Hesson, L., Gordon, L., Birrer, M., Latif, F., Clark, G.: Rassf6 is a novel member of the rassf family of tumor

suppressors. Oncogene **26**(42), 6203–6211 (2007)

[37] Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbsnp: the ncbi database of genetic variation. Nucleic acids research **29**(1), 308–311 (2001)

[38] Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., *et al.*: Pancanqtl: systematic identification of cis-eqtls and trans-eqtls in 33 cancer types. Nucleic acids research **46**(D1), 971–976 (2018)

[39] Consortium, G.: The gtex consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

[40] Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., *et al.*: Cistrome: an integrative platform for transcriptional regulation studies. Genome biology **12**(8), 1–10 (2011)

[41] Paiva, A.E., Lousado, L., Almeida, V.M., Andreotti, J.P., Santos, G.S., Azevedo, P.O., Sena, I.F., Prazeres, P.H., Borges, I.T., Azevedo, V., *et al.*: Endothelial cells as precursors for osteoblasts in the metastatic prostate cancer bone. Neoplasia **19**(11), 928–931 (2017)

[42] Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., Liu, L.: The role of collagen in cancer: from bench to bedside. Journal of translational medicine **17**, 1–22 (2019)

[43] Decker, A., Jung, Y., Cackowski, F., Taichman, R.: The role of hematopoietic stem cell niche in prostate cancer bone metastasis. Journal of Bone Oncology **5**(3), 117–120 (2016)

[44] Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., *et al.*: Integrative molecular concept modeling of prostate cancer progression. Nature genetics **39**(1), 41–51 (2007)

[45] Grisanzio, C., Werner, L., Takeda, D., Awoyemi, B.C., Pomerantz, M.M., Yamada, H., Sooriakumaran, P., Robinson, B.D., Leung, R., Schinzel, A.C., *et al.*: Genetic and functional analyses implicate the nudt11, hnf1b, and slc22a3 genes in prostate cancer pathogenesis. Proceedings of the National Academy of Sciences **109**(28), 11252–11257 (2012)

[46] Cui, R., Elzur, R.A., Kanai, M., Ulirsch, J.C., Weissbrod, O., Daly, M., Neale, B., Fan, Z., Finucane, H.K.: Improving fine-mapping by modeling infinitesimal effects. BioRxiv, 2022–10 (2022)

[47] Cai, M., Wang, Z., Xiao, J., Hu, X., Chen, G., Yang, C.: Xmap: Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias.

bioRxiv, 2023–03 (2023)

[48] Benner, C., Havulinna, A.S., Järvelin, M.-R., Salomaa, V., Ripatti, S., Pirinen, M.: Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. The American Journal of Human Genetics **101**(4), 539–551 (2017)

[49] Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., Van De Geijn, B., Reshef, Y., Márquez-Luna, C., *et al.*: Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nature genetics **52**(12), 1355–1363 (2020)

[50] Hernández, N., Soenksen, J., Newcombe, P., Sandhu, M., Barroso, I., Wallace, C., Asimit, J.: The flashfm approach for fine-mapping multiple quantitative traits. Nature Communications **12**(1), 6147 (2021)

[51] Arvanitis, M., Tayeb, K., Strober, B.J., Battle, A.: Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. The American Journal of Human Genetics **109**(2), 223–239 (2022)

[52] Zou, Y., Carbonetto, P., Xie, D., Wang, G., Stephens, M.: Fast and flexible joint fine-mapping of multiple traits via the sum of single effects model. bioRxiv, 2023–04 (2023)

[53] Zhu, X., Stephens, M.: Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The annals of applied statistics **11**(3), 1561 (2017)

[54] 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**(1), 139–153 (2016)

[55] Kanai, M., Elzur, R., Zhou, W., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., *et al.*: Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell genomics **2**(12), 100210 (2022)

[56] Berisa, T., Pickrell, J.K.: Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics **32**(2), 283 (2016)

[57] Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., Consortium, .G.P.: Alignment of 1000 genomes project reads to reference assembly grch38. Gigascience **6**(7), 038 (2017)

[58] Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P., Consortium, .G.P., et al.: Variant calling on the grch38 assembly with the data from phase three of the 1000 genomes project. Wellcome Open Research **4** (2019)

[59] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[60] Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A. Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., *et al.*: Wikipathways: connecting communities. Nucleic acids research **49**(D1), 613–621 (2021)

[61] García-Moreno, A., López-Domínguez, R., Ramirez-Mena, A., Pascual-Montano, A., Aparicio-Puerta, E., Hackenberg, M., Carmona-Saez, P.: Genecodis 4: Expanding the modular enrichment analysis to regulatory elements. bioRxiv, 2021–04 (2021)

Author contributions

Y.D.Z. and X.L. conceived and designed the model. X.L. designed the algorithm, implemented the software, and conducted analysis of simulation and real data. X.L. wrote the original draft of the manuscript. Y.D.Z. and P.C.S. contributed to the critical revision of the manuscript.