

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

Xiang Li¹, Pak Chung Sham^{2,3}, Yan Dora Zhang^{1*}

¹Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China.

²Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

³Centre 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

1 Introduction

2 Genome-wide association studies (GWAS) have discovered numerous genetic variants
3 associated with a wide range of complex traits and diseases [1]. However, pinpointing
4 the specific variants that have causal effects on the traits is challenging due to the
5 presence of high linkage disequilibrium (LD) among single nucleotide polymorphisms
6 (SNPs) and their small effect sizes [2–4]. The goal of statistical fine-mapping is to
7 identify the causal variants that have nonzero effects on the trait, which is essentially
8 a statistical problem known as “variable selection”. Since it is difficult to distinguish a
9 causal variant from other variants highly correlated with it without extra information,
10 penalized regression methods sometimes fail to select the true causal variants [5]. On
11 the other hand, Bayesian methods are more appropriate for fine-mapping by providing
12 posterior “credible sets” (CSs) [4]. A level $1 - \alpha$ CS is defined as a set of variants that
13 contains at least one causal variant with posterior probability no less than $1 - \alpha$ [6, 7].
14 A CS may contain multiple highly correlated candidate causal variants for further
15 functional validation.

16 To date, many Bayesian fine-mapping methods have been developed, including
17 CAVIAR [2], CAVIARBF [5], PAINTOR [3], JAM [8], DAP [9], FINEMAP [10, 11],
18 and SuSiE [12]. All these methods are based on discrete mixture priors, specifying a
19 prior probability for each variant being causal. Suppose there are M SNPs in the region
20 of interest, the number of possible models is 2^M . To reduce computational cost, these
21 methods need to set limit on the maximum number of causal variants. However, mis-
22 specifying the number may lead to decrease in performance [12]. In addition, existing
23 methods rely on exhaustive search, shotgun stochastic search, or stepwise selection to
24 explore the space of causal configurations, which can be time-consuming or lead to
25 poor suboptimal solutions [6, 12].

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

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

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

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

64 Material and methods

65 Overview of h2-D2

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (\text{Equation 1})$$

69 where \mathbf{y} is a vector of standardized phenotype values for N individuals, \mathbf{X} is an
70 $N \times M$ column-standardized genotype matrix for N individuals and M variants, $\boldsymbol{\beta} =$
71 $(\beta_1, \dots, \beta_M)^\top$ is an M -vector of effect sizes to be estimated, and $\boldsymbol{\varepsilon}$ is an N -vector of
72 error terms. We assume that $\boldsymbol{\varepsilon} \sim N_N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_N)$, where \mathbf{I}_N is an $N \times N$ identity matrix
73 and $N_k(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ denotes the k -variate normal distribution with mean $\boldsymbol{\mu}$ and covariance
74 matrix $\boldsymbol{\Lambda}$.

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

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

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

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

$$(\sigma_1^2, \dots, \sigma_M^2, 1 - h^2) \sim \text{Dir}(a_1, \dots, a_M, b), \quad (\text{Equation 3})$$

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

$$\beta_j | \sigma_j^2 \sim \text{DE}\left(\sqrt{\sigma_j^2/2}\right), \quad j = 1, \dots, M, \quad (\text{Equation 4})$$

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

85 Same as many other fine-mapping methods, h2-D2 requires GWAS summary
86 data only [26]. Assume the single-SNP summary statistics $D = \{\hat{\beta}_j, \hat{e}_j\}_{j=1}^M$
87 are provided, where $\hat{\beta}_j$ is the marginal effect size estimate of SNP j , and \hat{e}_j is its stan-
88 dard error. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_M)^\top$, $\hat{s}_j = \left(\hat{e}_j^2 + N^{-1}\hat{\beta}_j^2\right)^{1/2}$ for $j = 1, \dots, M$, and
89 $\hat{\mathbf{S}} = \text{diag}(\hat{s}_1, \dots, \hat{s}_M)$. The LD matrix is estimated from some reference panel as $\hat{\mathbf{R}}$.
90 The RSS likelihood of $\hat{\boldsymbol{\beta}}$ [26] is given by

$$\hat{\boldsymbol{\beta}} | \boldsymbol{\beta}, \hat{\mathbf{S}}, \hat{\mathbf{R}} \sim N_M\left(\hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}^{-1}\boldsymbol{\beta}, \hat{\mathbf{S}}\hat{\mathbf{R}}\hat{\mathbf{S}}\right). \quad (\text{Equation 5})$$

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

95 Credible level and credible set

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

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

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

105 Next, we extend this concept to multiple SNPs and define credible sets (CSs)
106 accordingly. Consider a set of SNPs $\mathcal{C} = \{j_1, \dots, j_k\}$. In the frequentist framework,

107 claiming that \mathcal{C} is a level $1 - \alpha$ CS is equivalent to rejecting the null hypothesis
 108 $H_{0\mathcal{C}} : \beta_{\mathcal{C}} \triangleq (\beta_{j_1}, \dots, \beta_{j_k})^\top = \mathbf{0}$ at the significance level of α , which can be declared if
 109 the null hypothesis $\mathbf{v}^\top \beta_{\mathcal{C}} = 0$ can be rejected at the significance level of α for at least
 110 one $\mathbf{v} \in \mathbb{R}^k$. Therefore, the credible level of \mathcal{C} is defined as

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

111 However, it is computationally infeasible to find the \mathbf{v} that maximize the credible
 112 level when the number of variants exceeds two. Instead, we choose a single \mathbf{v} to pro-
 113 vide a lower bound of the credible level. The coefficients of \mathbf{v} are selected to satisfy
 114 the condition that for positively correlated variants in \mathcal{C} , the corresponding coeffi-
 115 cients have the same signs, while for negatively correlated variants, the corresponding
 116 coefficients have different signs. We choose \mathbf{v} as an eigenvector of $\widehat{\mathbf{R}}_{\mathcal{C}}$ corresponding
 117 to its largest eigenvalue, where $\widehat{\mathbf{R}}_{\mathcal{C}}$ denotes the estimated LD matrix of SNPs in \mathcal{C} . If
 118 $\text{CL}_{\mathcal{C}}(\mathbf{v}) \geq 1 - \alpha$, we have $\widetilde{\text{CL}}_{\mathcal{C}} \geq 1 - \alpha$, and we conclude that \mathcal{C} is a level $1 - \alpha$ credible
 119 set. A greedy algorithm is designed to search all CSs achieving a pre-specified level
 120 (supplemental method 3).

121 Choice of hyper-parameters

122 In the Dirichlet prior (Equation 3), a smaller a_j leads to a higher concentration around
 123 0 for β_j , while a larger b indicates stronger global shrinkage. When incorporating
 124 external information, such as functional annotations, if the j -th SNP is more likely
 125 to be causal, a larger a_j can be set. By default, we suggest setting $a_1 = \dots = a_M =$
 126 $a \in [0.001, 0.01]$ for general fine-mapping tasks. A smaller a would make the MCMC
 127 chain converge slowly, while a larger a can be considered if there are evidences that
 128 the region may harbor a large number of causal variants (e.g. more than 10).

129 As for the choice of b , if an in-sample or highly accurate LD matrix $\widehat{\mathbf{R}}$ is available,
 130 we recommend estimating the local heritability using some well-known estimation
 131 procedures, such as the HESS estimator [27], which is defined as:

$$\widehat{h}^2 = \frac{N \widehat{\beta}^\top \widehat{\mathbf{R}}^{-1} \widehat{\beta} - M}{N - M}. \quad (\text{Equation 8})$$

132 Then, b can be chosen as follows:

$$b = \frac{(1 - \widehat{h}^2) \sum_{j=1}^M a_j}{\widehat{h}^2}. \quad (\text{Equation 9})$$

133 However, if the accuracy of $\widehat{\mathbf{R}}$ is poor, the HESS estimator may exhibit large bias. In
 134 this scenario, even if the true heritability is known, setting b according to Equation
 135 9 can lead to large effect size estimates for some non-causal variants in h2-D2. This
 136 is consistent with a recent finding that significant miscalibration due to external LD

137 matrices can produce suspicious results in meta-analysis fine-mapping studies [28]. To
138 address this, we suggest performing quality-control to filter out outlier variants before
139 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
140 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.

141 UK Biobank data preprocessing

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

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

157 Partition LD blocks

158 We noticed that the LD blocks partitioned by LDetect [29] based on 1000 Genome
159 reference panel are not optimal for UKBB reference panel. We developed a method
160 to divide the whole genome into nearly independent LD blocks, so as to improve
161 computational efficiency and achieve accurate fine-mapping results.

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

$$\arg \min_{k \in \{1, \dots, M\}} \frac{\sum_{1 \leq j_1 \leq k, k < j_2 \leq M} r_{j_1 j_2}^2}{k(M-k)},$$

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

169 As a result, we divided the entire autosomal region (excluding the major histo-
170 compatibility complex [MHC] regions) into a total of 3,717 nearly independent LD

171 blocks. We provide the script and the full list of LD blocks on our GitHub repos-
172 itory at <https://github.com/xiangli428/PrCaFineMapping>. This approach allows for
173 improved efficiency and accuracy in fine-mapping analyses using the UK Biobank
174 reference panel.

175 Simulations

176 We conducted simulation studies using UK Biobank imputed genotype data from
177 $N = 275,768$ unrelated British individuals [30]. For the simulations, we selected 100
178 nearly independent LD blocks on chromosome 2 (Table S1), and included variants
179 with $\text{MAF} \geq 0.01$ and INFO score ≥ 0.8 . We pruned SNPs such that the absolute
180 correlation $|r|$ between any two SNPs was less than 0.99. Each block contained a
181 varying number of SNPs, ranging from 288 to 1,122, and had a length between 0.25
182 and 2 Mb.

183 We designed four simulation scenarios with varying sample sizes, local heritabili-
184 ties, and numbers causal variants. For the first three scenarios, we used the genotypes
185 of all $N = 275,768$ individuals and considered different combinations of local heri-
186 tability and numbers of causal variants: (1) $h^2 = 0.1\%$, $n_{\text{causal}} = 5$; (2) $h^2 = 0.05\%$,
187 $n_{\text{causal}} = 5$; and (3) $h^2 = 0.1\%$, $n_{\text{causal}} = 10$. In the last scenario, we simulated eQTL
188 studies, where the sample sizes were small ($N = 1,000$), but the effect sizes of causal
189 SNPs were large ($h^2 = 10\%$), and $n_{\text{causal}} = 5$. Genotype values of each SNP were
190 standardized. In each scenario, for each block, the causal variants were chosen ran-
191 domly and the effect sizes of causal variants were sampled from a normal distribution
192 with mean 0. The phenotype values were then computed according to the multiple
193 regression model (Equation 1), where the error term $\boldsymbol{\varepsilon}$ were sampled from a multi-
194 variate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_N$. $\sigma_{\boldsymbol{\varepsilon}}^2$ was chosen
195 such that $\text{Var}(\mathbf{X}\boldsymbol{\beta}) / (\text{Var}(\mathbf{X}\boldsymbol{\beta}) + \sigma_{\boldsymbol{\varepsilon}}^2)$ equaled h^2 in each scenario. After standardiz-
196 ing the phenotype values and scaling the effect sizes of causal variants consistently,
197 we computed summary statistics for each variant.

198 To assess the influence of LD matrix accuracy on the fine-mapping performance,
199 we computed four LD matrices for each block. The first one was an in-sample LD
200 matrix computed from all 275,768 UKBB individuals ($\widehat{\mathbf{R}}$). The second one and the
201 third one were down-sample LD matrices, computed from randomly sampled 3,000
202 or 500 UKBB individuals, denoted by $\widehat{\mathbf{R}}_{\text{UKBB},3000}$ and $\widehat{\mathbf{R}}_{\text{UKBB},500}$, respectively. The
203 fourth one was an out-of-sample LD matrix computed from 522 unrelated European
204 ancestry individuals using the genotype data from the 1,000 Genomes Project on
205 GRCh38 [31, 32], denoted by $\widehat{\mathbf{R}}_{\text{1KG}}$. When using mismatched LD matrices, we applied
206 SLALOM to all pairs of SNPs with $|r| \geq 0.8$ and remove outlier non-causal variants
207 with DENTIST-S statistics ≥ 40 [28].

208 Compared methods

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

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

213 where $K \ll M$ is the maximum number of causal variants, and uses a shotgun stochastic
214 search algorithm to identify models with high posterior probabilities. SuSiE [6] is
215 a novel variable selection method that decomposes the effect size as the sum of single-
216 effect vectors and imposes a multinomial prior distribution on each single-effect vector.
217 SuSiE adopts an iterative Bayesian stepwise selection algorithm to optimize a varia-
218 tional approximation to the posterior distribution, as well as a refinement procedure
219 to address the convergence problem of the algorithm.

220 As for the choices of hyper-parameters, for h2-D2, we set $a_1 = \dots = a_M = a =$
221 0.005 . When using LD matrices $\hat{\mathbf{R}}$ or $\hat{\mathbf{R}}_{\text{UKBB},3000}$, we set b according to Equation 8
222 and Equation 9. When using $\hat{\mathbf{R}}_{\text{UKBB},500}$ or $\hat{\mathbf{R}}_{\text{1KG}}$, we set $b = 2 \times 10^5 M a$ for scenario
223 1-3, and $100 M a$ for scenario 4, respectively. For SuSiE, we set options “refine=TRUE”
224 and “estimate_residual_variance=TRUE”. For both SuSiE and FINEMAP, we set the
225 number of single-effect vectors or the maximum number of causal variants equals the
226 true number of causal variants in each scenario (5 or 10).

227 Comparison of causal variant effect sizes and their posterior 228 mean in simulation studies

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

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

233 where k and $\beta_0 > 0$ are the coefficients to be estimated. We used least square method
234 to estimate these coefficients. The fitted curves are shown in Figure S4.

235 Prostate cancer GWAS data preprocessing

236 We applied h2-D2 to identify candidate causal variants of prostate cancer (PrCa)
237 using summary GWAS data from a large meta-analysis involving $N_1 = 79,148$ cases
238 and $N_0 = 61,106$ controls of European ancestry [33]. We excluded the major histo-
239 compatibility complex MHC region (chr6 25-33 M) from our analysis. The remaining
240 autosomal regions were partitioned into 3,717 nonoverlapping regions with approxi-
241 mately independent LD (Material and methods). There are 126 risk variants out of
242 3,717 regions (i.e., contain at least one SNP with $P < 5 \times 10^{-8}$). 275,768 unrelated
243 British individuals from UK Biobank database were used as reference panel.

244 We filtered out duplicated SNPs, SNPs that were not present in UKBB reference
245 panel, SNPs with a imputation $r^2 < 0.3$, with a standard error of marginal effect size
246 on the allelic scale $< 5 \times 10^{-3}$ or $> 10^{-2}$, with a $\text{MAF} < 0.01$ in UKBB reference panel,
247 or with a $\text{logit}(\text{MAF})$ difference between UKBB reference panel and meta-analysis

248 larger than 0.5. Since mismatched LD matrices were used, to avoid unreliable results,
249 for each pair of SNPs with an absolute correlation $|r| \geq 0.8$, we checked if the pattern
250 of LD and GWAS summary statistics is suspicious using DENTIST-S statistic [28]. If
251 DENTIST-S statistic was greater than or equal to 30, the less significant SNP would
252 be removed. After these quality control steps, 6, 446, 747 common SNPs were retained
253 in our analysis. Before fine-mapping, the variants were pruned such that all pairwise
254 correlation $|r| < 0.95$. A total of 1, 342, 667 tag SNPs were retained for fine-mapping.
255 We used h2-D2 with specific hyper-parameters ($a_1 = \dots a_M = 0.005$ and $b = 2 \times 10^5$)
256 to fine-map each region and identify 95% CSs. Each 95% CS includes a set of tag
257 SNPs with a joint credible level ≥ 0.95 , as well as the pruned SNPs that are in high
258 LD with them.

259 **Annotations of variants**

260 The gene-based annotations of variants and their associated genes were extracted from
261 the dbSNP database (build 151) with GRCh37.p13 as the reference assembly [34].
262 These annotations include: NSF (non-synonymous frameshift), NSM (non-synonymous
263 missense), NSN (non-synonymous nonsense), SYN (synonymous), U3 (in 3' UTR), U5
264 (in 5' UTR), ASS (in acceptor splice site), DSS (in donor splice-site), INT (in intron),
265 R3 (in 3' gene region), and R5 (in 5' gene region).

266 Prostate cancer-specific cis- and trans-eQTL data were obtained from PanCanQTL
267 database [35]. Cis-eQTLs from normal prostate tissues mapped in European-American
268 subjects were obtained from GTEx V8 database [36].

269 DNaseI peaks, ChIP-seq peaks of histone marks and transcription factor binding
270 sites in prostate-derived cell lines were obtained from Cistrome database [37]. Details
271 of downloaded data are shown in Table S3. The peak coordinates were converted from
272 hg38 to hg19 reference assembly using LiftOver. Variants located within these peaks
273 were selected using BEDTools.

274 Enhancer-promoter loops identified from Hi-C data in RWPE1, C42B, and 22Rv1
275 cell lines were obtained from Supplementary Table 5A-C of ref. [38]. Annotated
276 H3K27ac HiChIP loops in LNCaP cell line were obtained from Table S7 of ref. [39].
277 Variants located within the identified enhancers were selected using BEDTools.

278 **Pathway enrichment analysis**

279 Potential target genes of credible causal variants (CCVs) were derived by merging (i)
280 associated genes of CCVs annotated in dbSNP database (build 151); (ii) associated
281 genes of eQTLs in CCVs in PanCanQTL and GTEx V8 databases; (iii) genes whose
282 promoters interact with enhancers covering CCVs in Hi-C or H3K27ac HiChIP data.
283 Protein-coding genes were retained based on GENCODE v42 annotations mapped to
284 GRCh37 assembly.

285 Enrichment analyses for pathways from GO Biological Process [40] and WikiPath-
286 ways [41] were carried out using GeneCodis [42]. To remove redundant pathways, we
287 computed Dice coefficients for all pairs of pathways. If the Dice coefficient between
288 two pathways is larger than 0.3, only the more significant one was retained.

289 Results

290 Simulation results

291 We conducted simulation studies to evaluate the performance of h2-D2 and compared
292 it with other fine-mapping methods. In brief, we chose 100 regions on chromosome
293 2 (Table S1) and simulated quantitative traits for each region. We considered four
294 scenarios with varying sample sizes, local heritabilities, and numbers causal variants.
295 To examine the influence of LD matrix accuracy on the fine-mapping performance, we
296 computed four LD matrices from different reference panels with varying sample sizes
297 for each block. Details are provided in [Material and methods](#).

298 We compared h2-D2 with two state-of-art fine-mapping methods, FINEMAP
299 [10, 11] and SuSiE-RSS [12]. On the SNP level, we evaluated the performance of vari-
300 able selection using the area under the precision-recall curve (AUPRC), which was
301 computed based on the credible level of each SNP for h2-D2 or the marginal poste-
302 rior inclusion probability (PIP) of each SNP for SuSiE and FINEMAP. In addition,
303 we assessed the accuracy of effect size estimation using the sum of squared error
304 (SSE) of β based on its posterior mean. When using in-sample LD matrices, h2-D2
305 consistently outperformed SuSiE and FINEMAP in terms of both AUPRC and SSE
306 across all scenarios (Figure 1A,B). As expected, all methods exhibited degraded per-
307 formance as the accuracy of the LD matrices decreased. In most cases, h2-D2 still
308 demonstrated superior performance. Additionally, h2-D2's credible levels were bet-
309 ter calibrated than PIPs of SuSiE and FINEMAP, particularly when inaccurate LD
310 matrices were used (Figure S1). The performance of SuSiE was close to that of h2-
311 D2. However, FINEMAP had significantly larger SSE and performed much worse in
312 Scenario 3 where the true number of causal variants was 10.

313 To gain further insights into the differences among the three methods, we compared
314 the AUPRC for each simulated dataset between h2-D2 and the other two methods
315 (Figure S2). While the AUPRC values were generally close for all three methods across
316 most datasets, h2-D2 exhibited significantly better performance in certain datasets. By
317 visualizing the fine-mapping results of these datasets, we noticed that in many cases
318 if there was a non-causal SNP having moderate LD with one or more causal SNPs
319 and having a stronger marginal association than causal SNPs, SuSiE and FINEMAP
320 tended to select that non-causal SNP instead of the causal ones. Figure S3 provides
321 two examples illustrating this issue. This phenomenon may be attributed to the step-
322 wise selection nature of SuSiE and the shotgun stochastic search algorithm employed
323 by FINEMAP. Once a marginally significant variant is included in the model, it is dif-
324 ficult for discrete-mixture prior-based methods to remove it, i.e., the algorithms are
325 more prone to be trapped into suboptimal solutions. It appears that the refinement
326 step of SuSiE cannot always alleviate this problem. On the other hand, continu-
327 ous shrinkage prior-based methods allow for the continuous updating of coefficients,
328 enabling smoother transitions among different local modes, and making the Markov
329 Chain Monte Carlo (MCMC) algorithm to explore the space of causal configurations
330 more extensively.

331 We also compared differences among the three methods in effect size estimation. We
332 grouped the variants into causal and non-causal categories and analysed the prediction

333 error for each group ([Material and methods](#), Figures S4 and S5). Although SuSiE and
334 h2-D2 produced similar estimation of causal variant effect sizes, h2-D2 had smaller
335 prediction errors for the non-causal variant effect sizes, suggesting that h2-D2 had
336 lower FDR than SuSiE. While FINEMAP demonstrated the lowest SSE for non-causal
337 variant effect sizes, it grossly underestimated causal variant effect sizes, presumably
338 from excessive shrinkage, resulting in larger SSE compared with SuSiE and h2-D2.

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

353 Finally, we compared runtime of the three methods (Figure 1H). The computa-
354 tional complexity of h2-D2 is proportional to M^2 (where M is the number of variants)
355 and the number of MCMC iterations (n_{MCMC}), while the computational complexity
356 of SuSiE is proportional to M^2 and the maximum number of single effects L . When
357 $n_{MCMC} = 10000$ and $L = 5$, the runtime of h2-D2 were approximately three times as
358 long as the runtime of SuSiE. The computational complexity of FINEMAP is primarily
359 determined by the maximum number of causal variants and the number of iterations,
360 so the runtime of FINEMAP didn't significantly vary with the number of variants.

361 Fine-mapping causal variants of prostate cancer

362 We applied h2-D2 to identify candidate causal variants of prostate cancer (PrCa) using
363 summary GWAS data from a large meta-analysis of European ancestry [33] ([Material
364 and methods](#)). Overall, we identified 164 CSs at 95% level (Table S2), containing
365 4,706 credible causal variants (366 tags). Among these CSs, 93 overlapped with the
366 106 CSs in autosomal risk loci reported by ref. [43] and 86 overlapped with the CSs
367 identified by ref. [39]. Out of the 3,717 regions analysed, 92 regions contained a single
368 CS, while 23 regions contained multiple CSs. 6 CSs were detected within non-risk
369 regions. The region with the highest number of CSs was chr8 127708268-128658961,
370 where 15 CSs were detected. This finding is consistent with previous research that
371 chr8q24 region harbors multiple loci associated with PrCa susceptibility [44]. The
372 sizes of the CSs ranged from 1 to 282 variants, with a median size of 12 variants.
373 There were 22 CSs containing only a single variant, including some well-established
374 causal variants of PrCa, such as rs77559646, which disrupts *ANO7* mRNA splicing
375 and protein expression [45], and rs61752561, which affects glycosylation and function
376 of prostate-specific antigen [46] (Table 1).

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_C_A	rs6545977	0.50	7.35×10^{-46}	1		
chr2:241912029-243041411	2.242135265_G_A	rs77559646	0.02	9.93×10^{-21}	1	<i>ANO7</i>	NSM ⁹ ,INT ¹⁰
chr3:169194244-170170389	3.170083629_C_G	rs61436251	0.21	1.76×10^{-63}	1	<i>SKIL</i>	INT
chr5:1279701-1551138	5.1287194_G_A	rs2853677	0.58	0.02	1	<i>TERT</i>	INT
chr5:1279701-1551138	5.1288547_T_C	rs2853676	0.73	8.86×10^{-12}	1	<i>TERT</i>	INT
chr5:1279701-1551138	5.1292118_G_A	rs71595003	0.03	1.78×10^{-16}	0.96	<i>TERT</i>	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	<i>IRX4</i>	INT,Cis eQTL (GTEx)
chr6:159951830-161847113	6.160581374_A_G	rs651164	0.69	2.15×10^{-36}	1	<i>SOD2,ACAT2,TCPI,MRPL18</i>	Enhancer (H3K27ac) HiChIP, LNCaP
chr6:159951830-161847113	6.160581502_T_C	rs4646283	0.14	1.31×10^{-5}	0.99	<i>SOD2,ACAT2,TCPI,MRPL18</i>	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	<i>TMM23B,MSMB,NCOA4</i>	INT,R5 ¹¹ , Cis eQTL (GTEx)
chr11:124697216-125111546	11.125054793_C_T	rs138466039	0.01	2.01×10^{-11}	1	<i>PKNOX2</i>	INT,Cis eQTL (GTEx)
chr12:12101106-12922339	12.12871099_T_G	rs2066827	0.24	2.31×10^{-9}	1	<i>CDKN1B</i>	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	<i>MMP14</i>	R5
chr17:7251713-8007416	17.7571752_T_G	rs78378222	0.01	1.73×10^{-9}	1	<i>TP53</i>	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	<i>KLK3</i>	NSM,INT
chr22:42872086-43649657	22.43500212_G_T	rs5759167	0.50	5.55×10^{-71}	1		

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

²Variant ID in the format {chr}-{pos}-{ref.seq}-{alt.seq}.

³dbSNP (build 151, GRCh37/hg19) rsID.

⁴Alternative allele frequency of controls in meta-analysis.

⁵Meta-analysis p-value.

⁶h2-D2 redible level.

⁷Putative target gene(s) of the variant.

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

⁹Non-synonymous missense.

¹⁰In Intron.

¹¹In 5' gene region.

¹²In 3' gene region.

¹³In 3' UTR.

377 In our analysis, we identified some novel independent association signals that have
378 not been previously reported. One such example is chr11 68810837-69542062, where
379 four 95% CSs were detected (Figure 2A, Figure S6A). CS:11-88-1 is represented by
380 rs12275055 ($P = 3.7 \times 10^{-98}$), which is known to have pleiotropic associations with
381 multiple cancer types [47]. This SNP acts as an eQTL in multiple tissues for *TPCN2*,
382 which plays a role in autophagy progression and extracellular vesicle secretion in cancer
383 cells [48]. The location of CS:11-88-2 overlaps with CS:11-88-1. Hi-C data from the
384 normal prostate cell line RWPE1 indicated that several SNPs within CS:11-88-2 are
385 located in an enhancer region that looping to the promoter of the cell cycle related
386 gene *CCND1* [38]. Furthermore, an interaction between the *TPCN2* promoter and the
387 *CCND1* promoter was detected by H3K27ac HiChIP in the LNCaP prostate cancer cell
388 line [39]. These findings suggest a possible mechanism involving a three-way interaction
389 between an enhancer harboring the causal SNPs in CS:11-88-1 and CS:11-88-2, the
390 *TPCN2* promoter, and the *CCND1* promoter. We also identified two other CSs, CS:11-
391 88-3 and CS:11-88-4, near the gene *CCND1*. Within CS:11-88-3, 4 out of 17 variants
392 are located in the 5' flanking region of *CCND1*. In CS:11-88-4, the most likely causal
393 variant is the lead SNP rs3212870 ($P = 1.5 \times 10^{-3}$), which is located intronic in
394 *CCND1*. The associations between CS:11-88-4 and PrCa have not been previously
395 reported, because of the weak marginal associations, which can be explained by the
396 moderate LD between CS:11-88-1, CS:11-88-2, and CS:11-88-4 (Figure S6A). Another
397 interesting example is chr4 73256856-74885359 (Figure 2B, Figure S6B), where we
398 identified a novel CS, CS:4-88-3, insignificantly associated with PrCa (minimum $P =$
399 6.5×10^{-4}). The lead SNP in CS:4-88-3, rs72649118, is a non-synonymous missense
400 SNP of *RASSF6*, a member of the RASSF family of tumor suppressors [49].

401 **Functional enrichment of prostate cancer credible causal** 402 **variants**

403 We used the hypergeometric tests to investigate the enrichment of credible causal
404 variants (CCVs) in specific genomic features, including prostate-specific DNaseI hyper-
405 sensitivity sites, ChIP-seq peaks of transcription factors and histone marks ([Material](#)
406 [and methods](#), Table S3). We observed significant enrichment of CCVs in active reg-
407 ulatory regions (defined by H3K27ac and H3K4me1 marks), active gene promoters
408 (defined by H3K4me3 and H3K9ac marks), actively transcribed gene bodies (defined
409 by H3K36me3 and H3K79me2 marks), and DNaseI hypersensitivity sites (Figure 3A).
410 CCVs were also significantly enriched in the binding sites of various transcription fac-
411 tors (Figure 3B, Table S3), including *AR* (androgen receptor), *NR3C1* (glucocorticoid
412 receptor), *ASH2L*, and *FOXA1*.

413 To formally evaluate the relationship between the biological functions associated
414 with SNPs and their contributions to the risk of PrCa, we fitted a linear model for the
415 logarithm of per-SNP heritability (i.e., the posterior mean of squared effect size) of all
416 1, 342, 667 tag SNPs using the following functional annotations of SNPs as predictors:
417 (i) 11 gene-based annotations extracted from the dbSNP database (build 151) [34];
418 (ii) cis- and trans-eQTLs within PrCa tissues from the TCGA database [35]; (iii) cis-
419 eQTLs within normal prostate tissues from the GTEx v8 database [36]; (iv) DNaseI

420 hypersensitivity sites, ChIP-seq peaks of 48 transcription factors and 9 histone modifi-
421 cations from normal prostate or prostate cancer cell lines, obtained from the Cistrome
422 Data Browser [37]; (v) enhancer elements identified by Hi-C data and H3K27ac ChIP-
423 seq peaks in normal prostate (RWPE1) and prostate cancer (C42B and 22Rv1) cell
424 lines [38]; (vi) enhancer elements predicted by H3K27ac HiChIP in the prostate cancer
425 cell line LNCaP [39]. In addition, $\log(f(1-f))$ and $\log(\text{LD score})$ were included as
426 covariates, where f is the minor allele frequency of SNP. This analysis revealed that
427 cis-eQTL (TCGA) ($P_{\text{adj}} = 5.4 \times 10^{-137}$), cis-eQTL (GTEx) ($P_{\text{adj}} = 8.2 \times 10^{-52}$), and
428 enhancer (H3K27ac HiChIP, LNCaP) ($P_{\text{adj}} = 1.8 \times 10^{-50}$) were the most significant
429 three annotations associated with per-SNP heritability (Figure 3C, Table S4). Trans-
430 eQTL (TCGA) ($P_{\text{adj}} = 1.5 \times 10^{-8}$) exhibited the largest effect size (0.36). Notably,
431 HDAC1 (histone deacetylase 1) binding site was the only significant functional anno-
432 tation with a negative effect on per-SNP heritability. These findings suggested that
433 genetic variants influencing gene expression levels and enhancer activity play a crucial
434 role in the development and progression of PrCa.

435 Putative target genes of prostate cancer credible causal variants

436 To identify potential target genes of CCVs, we integrated various sources of infor-
437 mation, including gene-based annotations from the dbSNP database (build 151),
438 eQTL data, and enhancer-promoter interaction data from Hi-C and HiChIP experi-
439 ments (Material and methods). As a result, we identified 385 protein-coding genes as
440 potential target genes of CCVs across all 95% CSs (Figure 4A, Table S2).

441 We further conducted pathway enrichment analysis to gain insights into the biolog-
442 ical functions and processes associated with these putative target genes. Our analysis
443 revealed significant over-representation of these genes in 52 non-redundant pathways
444 at an FDR of 0.05 (Figure 4B, Table S5). Notable enriched pathways included prostate
445 gland development, DNA damage response (only ATM dependent), positive regula-
446 tion of transcription by RNA polymerase II, and regulation of mitotic cell cycle. The
447 enrichment of putative target genes in cellular response to BMP (bone morphogenetic
448 protein) stimulus, collagen catabolic process, and definitive hemopoiesis pathways may
449 be attributed to the involvement of these processes in PrCa bone metastasis [50–
450 52]. Furthermore, putative target genes were also over-represented in toxin transport
451 pathway. Although previous studies have reported associations between PrCa and sev-
452 eral genes in toxin transport pathway, such as *SLC22A1–A3* [53, 54], the relationship
453 between PrCa and this pathway is not well elucidated and needs further investigation.

454 Discussion

455 In this article, we present h2-D2, a fine-mapping method that utilizes a continuous
456 global-local shrinkage prior. As an extension of R2-D2, h2-D2 is designed for GWAS
457 data where the phenotype values are standardized. Unlike existing fine-mapping meth-
458 ods that rely on discrete mixture priors, h2-D2 does not impose a constraint on the
459 maximum number of causal variants and allows for the exploration of a wider range
460 of causal configurations. In addition, h2-D2 does not rely on assumptions regard-
461 ing the distribution of causal variant effect sizes, compatible with infinitesimal effect

462 assumption for non-causal variants, which has been adopted by some recent works in
463 fine-mapping [55, 56]. These features ensure the applicability and flexibility of h2-D2
464 in various scenarios.

465 We develop an efficient MCMC algorithm for h2-D2 to sample from the posterior
466 distribution. We utilize several strategies to accelerate the mixing of MCMC chains,
467 allowing for a more extensive exploration of the model space. Simulation studies show
468 that h2-D2 is less likely to get trapped into local optima and performs better in variable
469 selection than discrete-mixture-prior-based methods including SuSiE and FINEMAP.
470 This may be due to the property of continuous priors that the coefficients are updated
471 continuously and the transitions among local modes can be smoothly. Our results also
472 highlight the importance of using accurate LD matrices derived from adequately large
473 reference panels, which concurs with previous discoveries [12, 57].

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

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

493 As a future direction, fine-mapping resolutions may be improved by integrating
494 functional annotations into the h2-D2 prior. Stratified LD score regression-based meth-
495 ods like PolyFun [58] are well suited to be incorporated with h2-D2, since h2-D2
496 prior is imposed on the per-SNP heritability directly. Furthermore, h2-D2 can also be
497 extended to multi-trait fine-mapping. Given the widespread existence of pleiotropy,
498 fine-mapping multiple traits simultaneously has the potential to enhance the power
499 of identifying shared causal variants among traits [59–61]. Jointly analyzing multiple
500 traits may provide valuable insights into the genetic architecture underlying complex
501 diseases and traits, and improve our understanding of the shared genetic basis between
502 different phenotypes.

503 Data and code availability

504 Prostate cancer summary data are available from the PRACTICAL Consor-
505 tium (http://practical.icr.ac.uk/blog/?page_id=8164). Enhancer-promoter loops iden-
506 tified from Hi-C data in RWPE1, C42B, and 22Rv1 cell lines are available
507 at [https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-12079-8/
508 MediaObjects/41467_2019_12079_MOESM7_ESM.xlsx](https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-019-12079-8/MediaObjects/41467_2019_12079_MOESM7_ESM.xlsx). Annotated H3K27ac HiChIP
509 loops in LNCaP cell line are available at [https://ars.els-cdn.com/content/image/1-s2.
510 0-S0002929721004195-mm3.csv](https://ars.els-cdn.com/content/image/1-s2.0-S0002929721004195-mm3.csv). The software h2D2 is available at [https://github.
511 com/xiangli428/h2D2](https://github.com/xiangli428/h2D2). Scripts and data related to PrCa fine-mapping analysis are
512 available at <https://github.com/xiangli428/PrCaFineMapping>.

513 Acknowledgments

514 This work was supported, in part, by the Hong Kong Research Grants Council (RGC)
515 Early Career Scheme 2021/22 (project number 27305221).

516 Declaration of interests

517 The authors declare no competing interests.

518 Web resources

519 UK Biobank, <https://www.ukbiobank.ac.uk/>
520 1000 Genomes on GRCh38, <https://www.internationalgenome.org/>
521 Prostate cancer summary data, http://practical.icr.ac.uk/blog/?page_id=8164
522 dbSNP (build 151) with GRCh37.p13 as reference assembly, [https://ftp.ncbi.nih.
523 gov/snp/organisms/human_9606_b151_GRCh37p13/](https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/)
524 PancanQTL, http://gong_lab.hzau.edu.cn/PancanQTL/
525 GTEx V8, <https://gtexportal.org/home/>
526 Cistrome, <http://cistrome.org/>
527 plink, <https://zzz.bwh.harvard.edu/plink/>
528 BEDTools, <https://bedtools.readthedocs.io/en/latest/>
529 LDetect, <https://bitbucket.org/nygresearch/ldetect/src/master/>
530 bigsnpr, <https://privefl.github.io/bigsnp/>
531 GeneCodis, <https://genecodis.genyo.es/>
532 FINEMAP, <http://christianbenner.com/>
533 SuSiE, <https://github.com/stephenslab/susieR>

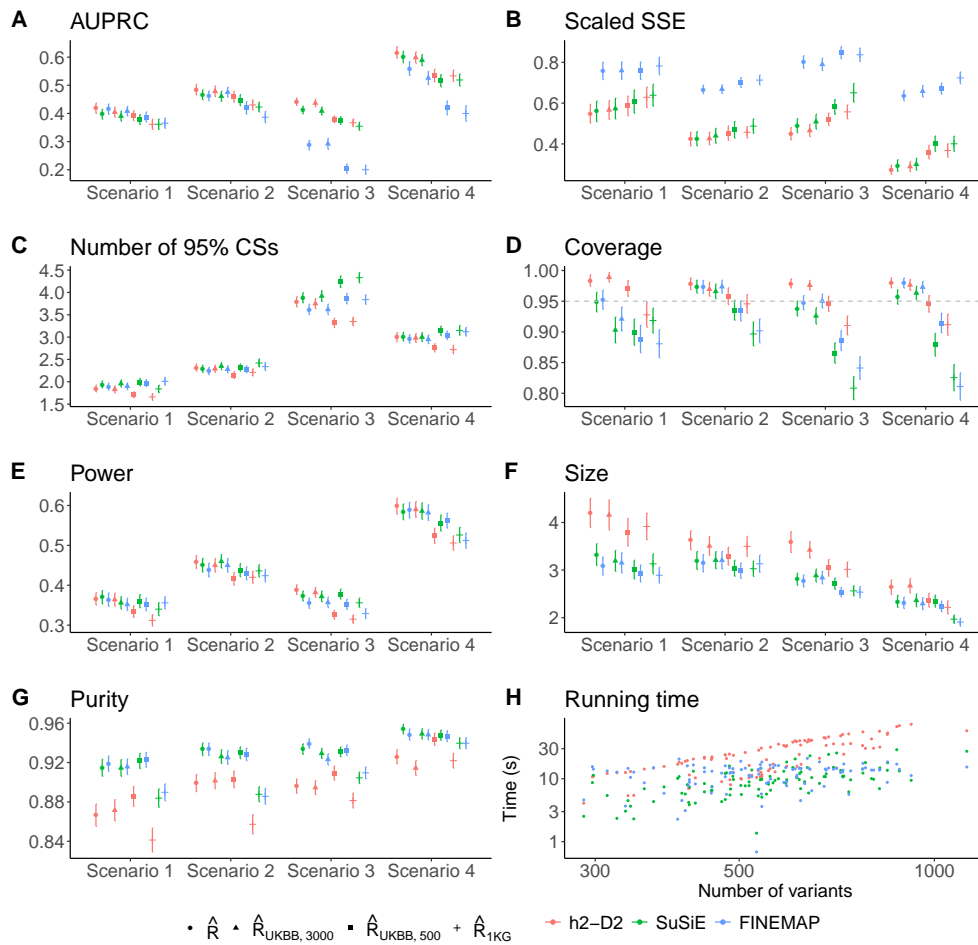


Figure 1 Performance comparison of h2-D2, SuSiE and FINEMAP on simulated data. In A-G, all values are the average ones across 100 datasets, with standard errors indicated by the error bars. (A) Area under the precision-recall curve (AUPRC) based on the credible level of each SNP for h2-D2 or the marginal posterior inclusion probability (PIP) of each SNP for SuSiE and FINEMAP. (B) Sum of squared error (SSE) of β 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, $K = 5$.

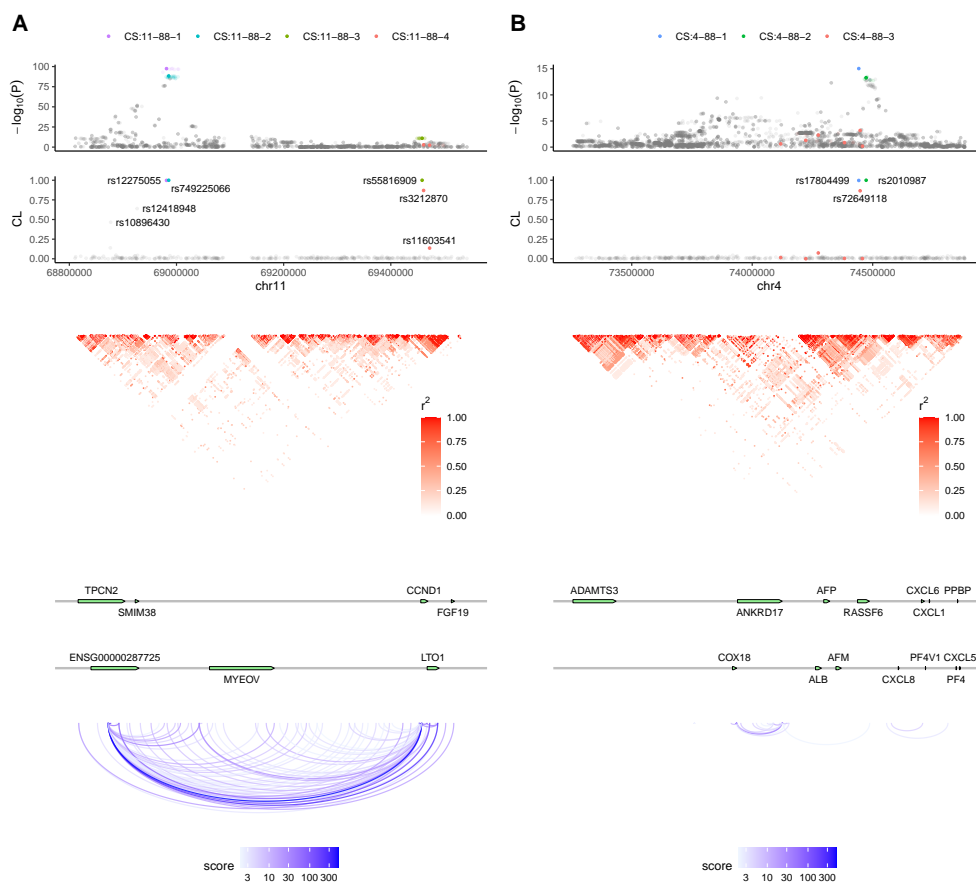


Figure 2 Fine-mapping results of two genomic regions in prostate cancer data analysis. (A) chr11 68810837-69542062; (B) chr4 73256856-74885359. The top panel depicts the marginal associations of variants ($-\log_{10}(P)$) from the GWAS meta-analysis data. The second panel illustrates the credible levels 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 [39].

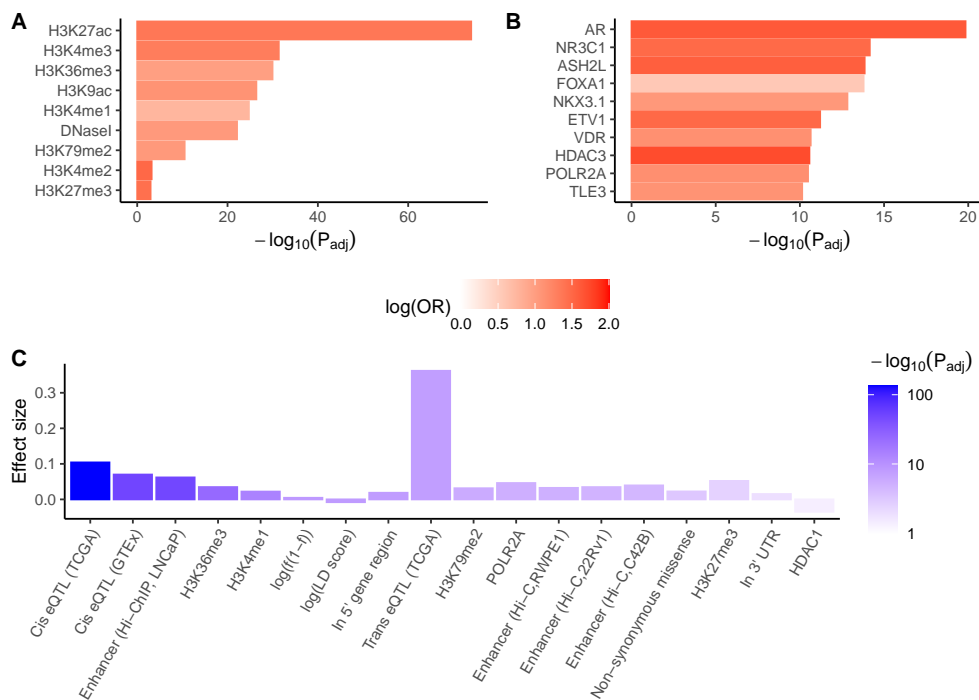


Figure 3 (A,B) Enrichment of credible causal variants in prostate-specific (A) histone marks and DNaseI hypersensitivity sites (B) top 10 transcription factor binding sites. Hypergeometric test P values are adjusted using the Benjamini-Hochberg (BH) method. (C) A linear regression model is fitted for 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 $P_{adj} < 0.05$.

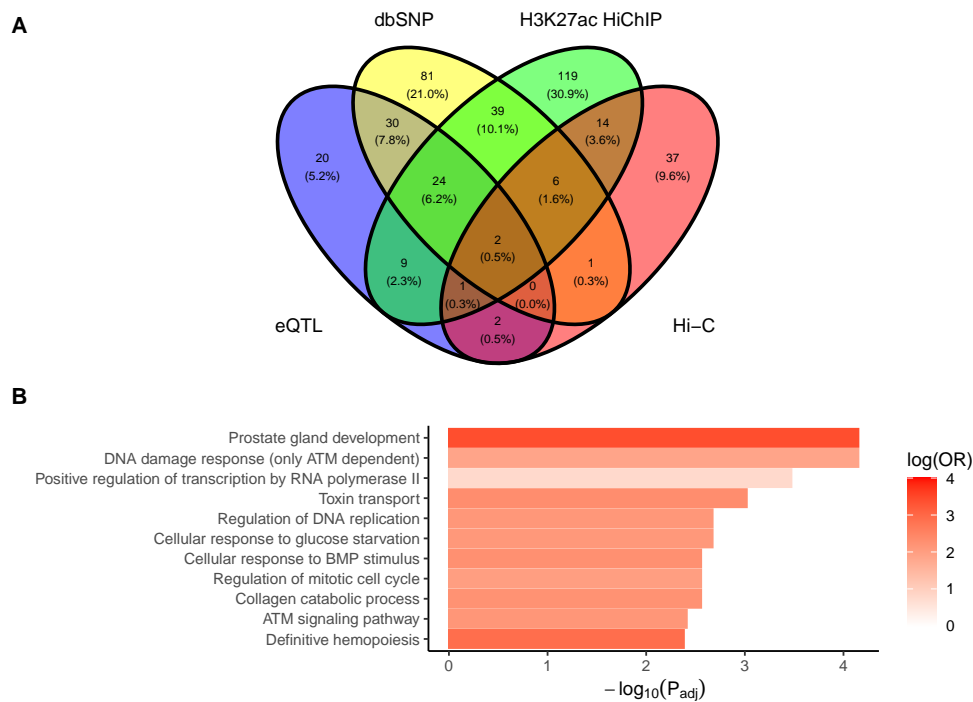


Figure 4 (A) Venn diagram showing the numbers of putative target genes inferred from different sources of information. (B) Enrichment of putative target genes in pathways from Gene Ontology Biological Processes and WikiPathways. Hypergeometric test P values are adjusted using the BH method. Pathways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

References

- [1] Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malan-gone, 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 candi-date causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**(8), 491–504 (2018)
- [5] 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 approx-imate bayesian method using marginal test statistics. *Genetics* **200**(3), 719–736 (2015)
- [6] 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)
- [7] 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 associa-tion signals for 14 loci in 3 common diseases. *Nature genetics* **44**(12), 1294–1301 (2012)
- [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 associa-tion 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] 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)
- [27] 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)
- [28] 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)
- [29] Berisa, T., Pickrell, J.K.: Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**(2), 283 (2016)
- [30] 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)
- [31] 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)
- [32] 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)
- [33] 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)
- [34] 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)
- [35] 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)
- [36] Consortium, G.: The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**(6509), 1318–1330 (2020)
- [37] 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)
- [38] 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)
- [39] 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)
- [40] 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)
- [41] 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)
- [42] 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)
- [43] 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)
- [44] 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)
- [45] 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)
- [46] 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)
- [47] 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)
- [48] Sun, W., Yue, J.: Tpc2 mediates autophagy progression and extracellular vesicle

- secretion in cancer cells. *Experimental cell research* **370**(2), 478–489 (2018)
- [49] Allen, N., Donniger, 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)
- [50] 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)
- [51] 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)
- [52] 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)
- [53] 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)
- [54] 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)
- [55] 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)
- [56] 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)
- [57] 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)
- [58] 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)

- [59] 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)
- [60] 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)
- [61] 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)

During the preparation of this work the authors used ChatGPT in order to improve the expression. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.