DAP-PIR: Efficient Genetic Association Analysis through Pseudo Importance Resampling

Bo Wang[1] and Xiaoquan Wen[1*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

*Corresponding author(s). E-mail(s): xwen@umich.edu;
Contributing authors: wnbo@umich.edu;

Abstract

In genome-wide association studies (GWAS), identifying causal variants from a vast number of genetic variants is challenging. This paper presents DAP-PIR, a method combining Deterministic Approximation of Posteriors (DAP) with Pseudo-Importance Resampling (PIR), to enhance computational efficiency and improve the accuracy of fine-mapping in genetic studies. By selecting a limited number of high-probability models, DAP-PIR effectively narrows down the candidate models, maintaining accuracy in posterior estimation while reducing computational costs. Simulations and real data analyses show the superiority of DAP-PIR over existing methods, including SuSiE, particularly in managing high-dimensional genetic data.

Keywords: Genetic fine mapping, variable selection, importance sampling, variational inference

1 Introduction

Genome-wide association studies (GWAS) have successfully uncovered thousands of genetic variants associated with complex traits and diseases [1]. However, due to the intricate linkage disequilibrium (LD) structure in the human genome, many variants identified in GWAS are merely correlated with true causal variants rather than causative themselves [2]. Fine mapping analysis is therefore essential to refine these associations and identify specific genetic variants likely to be causal [3, 4].

To address this challenge, Bayesian variable selection regression (BVSR) methods are widely employed to account for the uncertainty in identifying candidate causal variants. Traditional methods like CAVIAR rely on exhaustive searches [5], while DAP-G employs a boosting approach to prioritize variants, which, although effective, can sometimes be slow in practice [6, 7]. Alternatively, the SuSiE method offers an efficient, variational approach to prioritize SNPs based on posterior inclusion probabilities (PIP) [8]. However, SuSiE can occasionally experience accuracy issues with PIP, particularly in complex LD structures.

Here, we propose DAP-PIR, which combines the strengths of both DAP and SuSiE to achieve efficient and accurate deterministic approximation of posteriors. By leveraging SuSiE's variational approximation results, DAP-PIR efficiently prioritizes models for deterministic posterior estimation. This approach ensures computational efficiency while providing a robust solution to fine mapping challenges in GWAS. Through both simulations and real data applications, we demonstrate DAP-PIR's advantages over existing methods in terms of both speed and accuracy, making it an effective tool for fine mapping causal variants in GWAS.

2 Results

2.1 Overview of the DAP-PIR algorithm

The DAP-PIR algorithm uses two steps to efficiently approximate the posteiror distributions both at the variant-level and cluster-level. In the first step, the algorithm utilizes the variational approximation of the posterior distribution in a lower dimensional space as the proposal distribution in the importance sampling. Then, the algorithm employs the pseudo importance resampling (PIR) to further reduce the sample space and avoiding the need to draw random samples, instead focus on the high-probability regions. In the second step, the algorithm conduct the deterministic approximation of posteriors (DAP) by calculating the posterior inclusion probabilities (PIPs) for each high-probability model obtained in the first step by marginalizing over the priror. The DAP-PIR algorithm outputs the PIP of each variant by marginalizing the posterior model probability, and also provides level-ρ credible sets, which is a subset of variants that has a specific probability of containing at least one causal variants.

2.2 Simulation studies

Accuracy of the DAP-PIR algorithm

To demonstrate the accuracy and efficiency of the DAP-PIR algorithm, we compare its posterior approxmiations to those obtained through exact calculations. We focus on a scenario with $p = 10$ *cis*-SNPs to reduce the computational burden of exact posterior calculations. We simulate genotypes from a standard normal distribution, and randomly select one to five variants as causal variants. Phenotypes are simulated by adding random noise to the genetic effects. For each configuration of causal variants, we simulate 1,000 times. Using these 5,000 generated datasets, we evaluate the performance of the DAP-PIR algorithm alongside exact posterior calculations and results

from the DAP-G algorithm, employing the same prior specifications for all methods. The SuSiE algorithm is also included to assess its posterior approximations. For the DAP-PIR algorithm, we use a threshold value $\lambda = 10^{-6}$.

We begin by exmaining the ratio of the estimated normalizing constant C^* (calculated by DAP-PIR and DAP-G) to the true normalizing constant C derived from exact calculations. A ratio C^*/C close to 1 indicates better exploration of the model space. As shown in Table 1, the DAP-PIR algorithm consistently achieves a high C^*/C ratio across varying numbers of causal variants. This highlights its ability to efficiently identify high-probability regions of the model space and provide a more accurate estimate of the normalizing constant compared to DAP-G.

A key factor contributing to DAP-PIR's performance is its ability to explore more models than DAP-G, whcih naturally results in a higher C^*/C ratio. Importantly, the small Root Mean Squared Error (RMSE) of DAP-PIR's PIPs confirms that it focuses effectively on the high-probability regions of the model space when including more models than DAP-G. Notably, although exploring more models than DAP-G, DAP-PIR still investigates far fewer models than the total $2^{10} = 1,024$ models required for exact calculations. This demonstrates that by focusing on high-probability regions, the DAP-PIR algorithm achieves accurate posterior approximations while remaining computationally efficient, particularly as p increases.

Additionally, the RMSE of the PIPs from the DAP-PIR algorithm is smaller than that of both the DAP-G and SuSiE algorithms, demonstrating better posterior approxiamtion. The PIP comparison plot (Figure 1) further confirms that the DAP-PIR algorithm yields more accurate variant PIP estimates than the other methods.

Table 1 Comparison of the performance metrics for DAP-PIR, DAP-G, and SuSiE across varying number of causal variants (S). Metrics include the mean ratio of the estiamted normalizing constant (C^*) to the true constant (C), the median number of models explored, and the RMSE of the PIPs. A higher C^*/C ratio and lower RMSE indicate better performance.

	Mean of C^*/C		Median Model Numbers		RMSE of approximate PIP		
S	DAP-G	DAP-PIR	DAP-G	DAP-PIR	DAP-G	DAP-PIR	SuSiE
1	0.922	0.999	13	114	8.05×10^{-3}	3.93×10^{-5}	1.42×10^{-1}
2	0.913	0.999	14	101	9.01×10^{-3}	4.44×10^{-5}	1.12×10^{-1}
3	0.915	0.999	14	91	9.71×10^{-3}	7.14×10^{-5}	8.88×10^{-2}
4	0.922	0.999	15	81	9.40×10^{-3}	1.11×10^{-4}	7.57×10^{-2}
5	0.928	0.998	16	74	9.87×10^{-3}	5.45×10^{-4}	6.02×10^{-2}

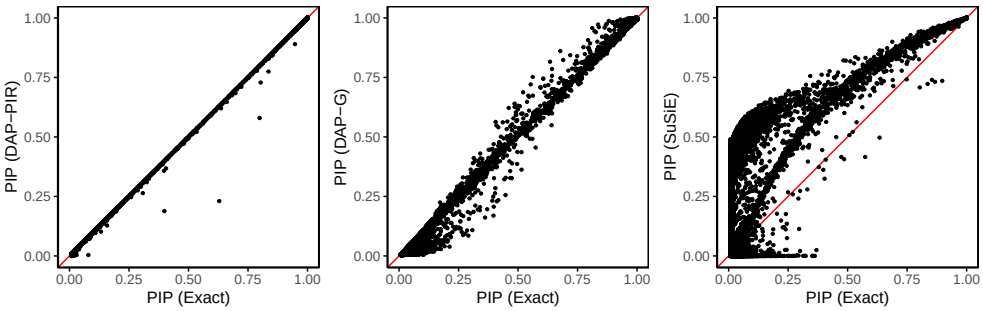

Fig. 1 Comparison of PIPs obtained using DAP-PIR, DAP-G, SuSiE against exact calculations for scenarios with one to five causal variants. The plots illustrate the accuracy of each method in estimating the PIP of each variant, while exact calculations are considered the ground truth.

Power comparions of fine mapping algorithms

We conduct several simulation studies to show the accuracy and efficiency of the DAP-PIR algorithm. Comparing the (1) power; (2) coverage; (3) PIP calibration; (4) computational efficiency among fine mapping methods.

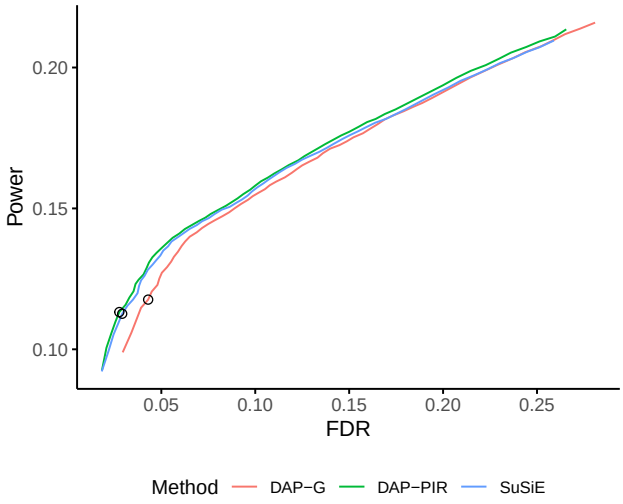

Fig. 2 Power versus FDR curves of results from the DAP-PIR, DAP-G, and SuSiE algorithms across all simulations. The plot is obtained by varying the PIP threshold, where the circle represents the threshold of 0.95. FDR is calculated as FP/(FP+TP), where FP and TP represent the number of false positives and true positives. The power is defined as TP/(TP+FN), where FN denotes the number of false negatives.

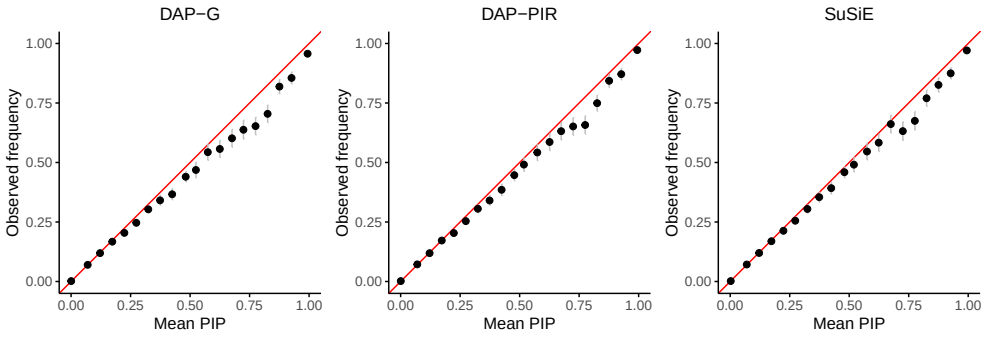

Fig. 3 PIP calibration plot across all simulations. PIPs are grouped into 20 evenly spaced bins from 0 to 1, and the observed frequency is calculated by the proportion of causal variants in the corresponding bin. A well calibrated method should yield points close to the diagnoal line. Gray error bars show 2 standard errors.

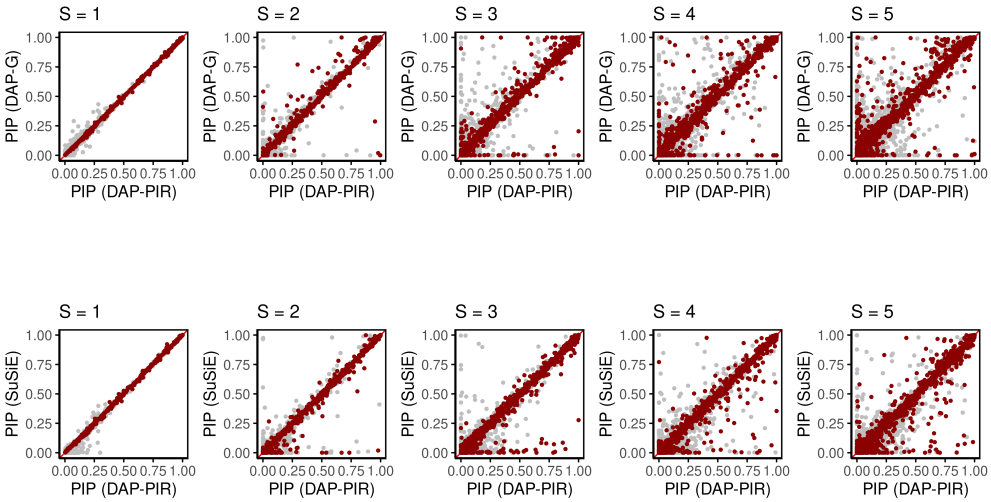

Fig. 4 Comparison of variant PIP across all simulations for varying causal variant number S and PVE = 0.2.

2.3 Real data results

We analyze some real data using various fine mapping methods.

3 Discussion

4 Methods

4.1 Bayesian variable selection model

We consider the context of association analysis for a single quantitative trait. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ denote an n-dimensional vector representing the trait of interest, $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_p)$ represent an $n \times p$ genotype matrix, where each column $\mathbf{g}_j$ is an n-dimensional vector of genotypes for the j-th variant. We assume that $\mathbf{y}$ and the columns of $\mathbf{G}$ are centered to have a mean of zero to avoid the need for an intercept term. The model is then defined as:

$$\mathbf{y} = \sum_{j=1}^{p} \beta_j \mathbf{g}_j + \mathbf{e},\ \mathbf{e} \sim N_n\left(0, \tau^{-1} I_n\right), \tag{1}$$

where β_j denotes the genetic effect of the j-th variant, and τ represents the precision. The goal of GWAS fine mapping is to prioritize causal variants rather than estimate their effect sizes, therefore we introduce an inclusion indicator $\gamma_j = I(\beta_j \neq 0)$ for each variant, where $\gamma_j = 1$ indicates that the j-th variant is causal. The primary aim is infer the posterior distribution of the causal status vector: $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$. While evaluating each possible model configuration $\boldsymbol{\gamma}$ is straightforward as it only requires integrating out β_j's and τ, summing over all 2^p model configurations becomes computationally intractable as p increases. This makes traditional Markov Chain Monte Carlo (MCMC) methods inefficient for fine mapping tasks involving even a moderate number of variants. Therefore, we utilize the efficient pseudo importance resampling method to explore the model space and approximate the posterior distribution of $\boldsymbol{\gamma}$.

4.2 Pseudo importance resampling

For a discrete random variable $\mathbf{X}$ with target distribution $p(\mathbf{x})$, the objective is to approxiamte the expectation $\mathbb{E}_p[f(\mathbf{X})]$ for a function $f(\mathbf{x}) = \mathbf{x}$. Due to the high-dimensional sample space, exact computation is impractical. We employ Importance Sampling to estimate this expectation by sampling from a proposal distribution $q(\mathbf{x})$ and assigning importance weights $w\left(\mathbf{x}^i\right) = p\left(\mathbf{x}^i\right)/q\left(\mathbf{x}^i\right)$ for samples $\mathbf{x}^i \sim q(\mathbf{x})$. The IS estimate is:

$$\tilde{\boldsymbol{\mu}}_{\mathrm{IS}} = \frac{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right) f\left(\mathbf{x}^i\right)}{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right)}. \tag{2}$$

To obtain a reliable estimate, $q(\mathbf{x})$ should closely approximate $p(\mathbf{x})$, particularly in regions where $f(\mathbf{x})\, p(\mathbf{x})$ has significant values.

We use Variational Approximation (VA) to construct an effective proposal distribution by minimizing the Kullback-Leibler (KL) divergence between $p(\mathbf{x})$ to $q(\mathbf{x})$. VA provides a tractable approxiamtion, balancing efficiency and accuracy. The Variational Importance Sampling (VIS) method integrates VA with IS, leveraging VA's computational efficiency while reducing bias through IS.

Sampling Importance Resampling (SIR) selects IS-generated samples proportional to their importance weights to approximate $p(\mathbf{x})$ [9, 10]. Given n samples

Algorithm 1 Variational Importance Sampling (VIS)

1. Target distribution: derive the analytical expression of $p'(\mathbf{x})$ up to a normalizing constant.
2. Proposal distribution: obtain the variational approximation $q(\mathbf{x})$ to $p(\mathbf{x})$ within the family $\mathcal{Q}$.
3. Sampling: for $i \in \{1, \ldots, n\}$,
   (a) draw $\mathbf{x}^i$ from the proposal distribution $q(\mathbf{x})$.
   (b) calculate the function value $f\left(\mathbf{x}^i\right)$.
   (c) calculate the importance weight $w\left(\mathbf{x}^i\right) = p'\left(\mathbf{x}^i\right) / q\left(\mathbf{x}^i\right)$.
4. Estimation: estimating $\mathbb{E}_p[f(\mathbf{X})]$ by $\tilde{\boldsymbol{\mu}}_{\mathrm{IS}} = \frac{\sum_{i=1}^n w\left(\mathbf{x}^i\right) f\left(\mathbf{x}^i\right)}{\sum_{i=1}^n w\left(\mathbf{x}^i\right)}$.

$\left\{\left(\mathbf{x}^i, w\left(\mathbf{x}^i\right)\right) : i = 1, \cdots, n\right\}$, SIR resamples m values from them according to their importance weights:

$$\Pr\left[\mathbf{x}^* = \mathbf{x}^k\right] = \frac{w\left(\mathbf{x}^k\right)}{\sum_{i=1}^n w\left(\mathbf{x}^i\right)}. \tag{3}$$

The resampled set $\left\{\mathbf{x}^{*j} : j = 1, \cdots, m\right\}$ has equal weights and represents the target distribution p. The SIR algorithm is outlines in Algorithm 2. The rationale behind SIR is based on the fact that as n/m tends to infinity, the m samples are drawn with probabilities given by:

$$p^*(\mathbf{x}) \propto q(\mathbf{x})\, w(\mathbf{x}) \propto q(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} = p(\mathbf{x}),$$

which shows that the SIR algorithm generates independent and identically distributed (i.i.d) samples from the target distribution $p(\mathbf{x})$, as desired. However, drawing such a large number of samples can be computationally expensive. To balance computational efficiency and performance, Rubin suggested using $n/m = 20$ as a practical ratio for resampling, which provides adequate performance without requiring excessive duplicates in the resampling [9].

To enhance efficiency, we introduce Pseudo Importance Resampling (PIR), which eliminates explicit resampling by shrinking the proposal distribution. PIR focuses on regions where $p(\mathbf{x})$ has meaningful contributions, reducing computational cost in high-dimensional settings.

In practice, if a value of $\mathbf{x}$ exhibits a near-zero probability density under the target distribution $p(\mathbf{x})$, its contribution to the final result becomes negligible. In such cases, the proposal distribution can be effectively reduced to zero for that region, excluding it from further consideration. This reduction in the number of regions to explore allows us to focus computational resources on areas where $p(\mathbf{x})$ significantly impacts the result. Prioritizing these high-priority regions improves efficiency in high-dimensional settings, making the process more computational feasible. The KL divergence penalizes cases where p is large while $q(\mathbf{x})$ is small, implying that when $q(\mathbf{x})$ is small, it is highly probable that $p(\mathbf{x})$ is also small. As a result, we can safely reduce the sample space by identifying and excluding values of $\mathbf{x}$ that are negligible under the

Algorithm 2 Sampling Importance Resampling (SIR)

1. Target distribution: the distribution $p(\mathbf{x})$ with the expression of $p'(\mathbf{x})$ up to a normalizing constant.
2. Proposal distribution: the variational approximation $q(\mathbf{x})$ to $p(\mathbf{x})$ within the family $\mathcal{Q}$.
3. Sampling: for $i \in \{1, \ldots, n\}$,
   (a) draw $\mathbf{x}^i$ from the proposal distribution $q(\mathbf{x})$.
   (b) calculate the importance weight $w\left(\mathbf{x}^i\right) = p'\left(\mathbf{x}^i\right) / q\left(\mathbf{x}^i\right)$.
4. Resampling: for $j \in \{1, \ldots, m\}$,
   (a) resample $\mathbf{x}^{*j}$ from the samples $\left\{\mathbf{x}^i : i = 1, \cdots, n\right\}$ with probabilities proportional to the importance weights.
   (b) include $\mathbf{x}^{*j}$ in the resampled set.
5. Result: the resampled set $\left\{\mathbf{x}^{*j} : j = 1, \cdots, m\right\}$ which is approximately distributed according to the target distribution $p(\mathbf{x})$.

proposal distribution. Even in cases where $p(\mathbf{x})$ is small but $q(\mathbf{x})$ is large (and thus not excluded), the contribution to the overall result remains minimal due to the small value of $p(\mathbf{x})$.

For a specific point $\mathbf{x}^{(k)}$ in the target distribution, the proposal distribution shrinkage is defined as:

$$q'\left(\mathbf{x}^{(k)}\right) = \begin{cases} 0, & \text{if } q\left(\mathbf{x}^{(k)}\right) < \epsilon, \\ q\left(\mathbf{x}^{(k)}\right), & \text{if } q\left(\mathbf{x}^{(k)}\right) \geq \epsilon, \end{cases}$$

where ϵ is a threshold value. The normalized proposal distribution is given by:

$$q^*\left(\mathbf{x}^{(k)}\right) = \frac{q'\left(\mathbf{x}^{(k)}\right)}{\sum_{i=1}^{N} q'\left(\mathbf{x}^{(i)}\right)} = \frac{q'\left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} q\left(\mathbf{x}^{(j)}\right)},$$

where $\mathbf{x}^{(j)}$ corresponds to points with non-zero proposal density, N is the total number of possible values, and M is the number of points with non-zero proposal density. During the SIR process, the probability of selecting a value $\mathbf{x}^{(k)}$ in the reduced sample

space is calculated as:

$$\begin{aligned}
\Pr\left(\mathbf{x}^* = \mathbf{x}^{(k)}\right) &= \frac{w\left(\mathbf{x}^{(k)}\right) \# \left(\mathbf{x}^{(k)}\right)}{\sum_{i=1}^{N} w\left(\mathbf{x}^{(i)}\right) \# \left(\mathbf{x}^{(i)}\right)} \\
&= \frac{w\left(\mathbf{x}^{(k)}\right) \# \left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} w\left(\mathbf{x}^{(j)}\right) \# \left(\mathbf{x}^{(j)}\right)} \\
&\xrightarrow{n\to\infty} \frac{\frac{p'\left(\mathbf{x}^{(k)}\right)}{q^*\left(\mathbf{x}^{(k)}\right)} q^*\left(\mathbf{x}^{(k)}\right) \times n}{\sum_{j=1}^{M} \frac{p'\left(\mathbf{x}^{(j)}\right)}{q^*\left(\mathbf{x}^{(j)}\right)} q^*\left(\mathbf{x}^{(j)}\right) \times n} \\
&= \frac{p'\left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} p'\left(\mathbf{x}^{(j)}\right)},
\end{aligned}$$

where $\#\left(\mathbf{x}^{(k)}\right)$ is the number of samples with value equal to $\mathbf{x}^{(k)}$. Thus, instead of drawing an large number of n samples and resampling an adequate number of m from them, we can directly approximate the probability of each value in the target distribution by focusing only on the M points with non-negligible probability. This process, based on SIR but avoiding the separate steps of sampling and resampling, is referred to as Pseudo Importance Resampling (PIR) outlined in Algorithm 3. PIR effectively reduces the sample space by shrinking the proposal distribution and concentrating on the regions that matter most, providing a good approximation to the target distribution.

Algorithm 3 Pseudo Importance Resampling (PIR)

1. Target distribution: the distribution $p(\mathbf{x})$ with the expression of $p'(\mathbf{x})$ up to a normalizing constant.
2. Proposal distribution: the variational approximation $q(\mathbf{x})$ to $p(\mathbf{x})$ within the family $\mathcal{Q}$.
3. Proposal distribution shrinkage: shrink $q(\mathbf{x})$ to obtain $q^*(\mathbf{x})$, reducing the sample space to M non-negligible partitions.
4. Pseudo resampling: for $j \in \{1, \ldots, M\}$,
   (a) calculate the probability score $p'\left(\mathbf{x}^{(j)}\right)$.
   (b) calculate the approximation $p'\left(\mathbf{x}^{(j)}\right) / \sum_{j=1}^{M} p'\left(\mathbf{x}^{(j)}\right)$.

4.3 Computation of Bayes factors over prior distributions

The posterior probability of one model configuration $\boldsymbol{\gamma}^{(i)}$ can be derived using Bayes' theorem:

$$p\left(\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(i)} \mid \mathbf{G}, \mathbf{y}\right) = \frac{\pi\left(\boldsymbol{\gamma}^{(i)}\right) \mathrm{BF}\left(\boldsymbol{\gamma}^{(i)}\right)}{\sum_{\boldsymbol{\gamma}' \in \Gamma} \pi\left(\boldsymbol{\gamma}'\right) \mathrm{BF}\left(\boldsymbol{\gamma}'\right)},$$

where Γ denotes all 2^p possible model configurations, $\pi\left(\boldsymbol{\gamma}^{(i)}\right)$ is the prior probability of model $\boldsymbol{\gamma}^{(i)}$, and $\mathrm{BF}\left(\boldsymbol{\gamma}^{(i)}\right)$ is the Bayes factor, which measures the marginal likelihood of $\boldsymbol{\gamma}$ evaluated at $\boldsymbol{\gamma}^{(i)}$.

Under the D_2 prior, the prior distribution for τ is given by:

$$\tau \sim \Gamma\left(\kappa/2, \lambda/2\right),$$

where the limiting form of the prior is obtained as $\kappa, \lambda \to 0$. And we assign a spike-and-slab prior for the effect size β_j:

$$\beta_j \mid \gamma_j \sim (1-\gamma_j)\delta_0 + \gamma_j N\left(0, \phi^2/\tau\right),$$

where δ_0 refers to Dirac's delta function, and ϕ^2 is the prior variance of the effect size of the causal variant. In the computation of the Bayes factor, we use $\phi^2 = 0.6^2$. For a specific value of ϕ^2 and a model configuration $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)$, the limiting of the Bayes factor is calculated by:

$$\lim_{\substack{\kappa\to 0\\ \lambda\to 0}} \mathrm{BF}(\boldsymbol{\gamma}) = \frac{1}{|\phi^{-2}I + \mathbf{X}^T\mathbf{X}|^{1/2}} \frac{1}{\phi} \left[\frac{\mathbf{y}^T\left(I - \mathbf{X}\left(\phi^{-2}I + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right]^{-n/2}, \quad (4)$$

where $\mathbf{X}$ represents genotypes indicated by non-zero entry in $\boldsymbol{\gamma}$. We can also use a grid of values for the prior variance of the effect size to capture a spectrum of effect sizes, which is particularly useful when the true effect size is unknown, and the Bayes factor can be calculated by averaging over the grid of values.

4.4 Posterior inference of the DAP-PIR algorithm

4.4.1 Calculation of variant-level PIPs

The posterior inclusion probability (PIP) for each variant is then calculated by marginalizing the posterior probabilities across all model configurations:

$$\mathrm{PIP}_j := \Pr\left(\gamma_j = 1 \mid \mathbf{G}, \mathbf{y}\right) = \sum_{\boldsymbol{\gamma}':\gamma_j=1} p(\boldsymbol{\gamma} = \boldsymbol{\gamma}' \mid \mathbf{G}, \mathbf{y}). \quad (5)$$

4.4.2 Procedure to constuct signal clusters

We propose a procedure to construct signal clusters from multiple single-effects detected by SuSiE and calculate signal-level PIPs, as detailed in Algorithm 4. This procedure requires genotype data $\mathbf{X}$, a genotype R^2 threshold τ, and SuSiE results as input. Alternatively, genotype data can be replaced with the LD information $\mathbf{R}$. From the SuSiE results, we use only the vector of posterior probabilities $\boldsymbol{\alpha}_l$. Additionally, we define a function $\mathrm{get\_r2}\left(s, \mathcal{Q}\right)$ to compute the pairwise genotype R^2 between a variant s and all elements in the set $\mathcal{Q}$.

First, we identify the indices of variables based on their posterior probabilities in the effect, which is identical to SuSiE's initial ranking. The variable with the highest

posterior probability is then included in the cluster. For the remaining $p-1$ variables, we sequentially evaluate them according to their indices and check if they are correlated with the elements in the cluster. A variable is included in the signal cluster if the minimum pairwise correlation with **all variables already in the cluster** exceeds the threshold. Once all variables have been processed, the cluster construction is finalized. This procedure is repeated for all L effects detected by SuSiE, resulting in L signal clusters.

For each signal cluster, we can compute its signal-level PIP by marginalizing the posterior proabilities of models with at least one variant in the signal cluster C:

$$\begin{aligned} \text{SPIP}_C &:= \Pr\left(\text{at least one } \gamma_j = 1,\ j \in C \mid \mathbf{G}, \mathbf{y}\right) && (6)\\ &= 1 - \Pr\left(\text{all } \gamma_j = 0,\ j \in C \mid \mathbf{G}, \mathbf{y}\right) && (7)\\ &= 1 - \sum_{\boldsymbol{\gamma}':\gamma_j=0,\ \forall j \in C} p(\boldsymbol{\gamma} = \boldsymbol{\gamma}' | \mathbf{G}, \mathbf{y}). && (8) \end{aligned}$$

The signal-level PIP SPIP_C is the probability of containing at least one causal variant in the signal cluster C.

Algorithm 4 Algorithm for Constructing Signal Clusters

Require: Genotype data $\mathbf{G}$ (or equivalently the LD information $\mathbf{R}$), genotype R^2 threshold τ, and posterior probabilities $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_L$ from SuSiE

Require: Function $\text{get\_r2}\,(s, \mathcal{Q})$ to compute the pairwise genotype R^2 between variant s and all variants in the set $\mathcal{Q}$

Ensure: Construct signal clusters $C_1, \cdots, C_L$ and calculate their signal-level PIPs $\text{SPIP}_{C_1}, \cdots, \text{SPIP}_{C_L}$ using genotype data and SuSiE results

Initialize clusters: $C_1, \cdots, C_L \leftarrow \varnothing$
for $l = 1, \cdots, L$ **do**
  Obtain indices $\boldsymbol{r} = (r_1, \ldots, r_p)$, where $\alpha_{l_{r_1}} > \alpha_{l_{r_2}} > \ldots > \alpha_{l_{r_p}}$ ▷ Rank variants
  $C_l \leftarrow \{r_1\}$ ▷ Add the variant with highest posterior probability
  for $i = 2, \cdots, p$ **do**
    if $\min(\text{get\_r2}(r_i, C_l)) \geq \tau$ **then**
      $C_l \leftarrow C_l \cup \{r_i\}$ ▷ Add variant r_i if correlated with the cluster
    end if
  end for
  Calculate SPIP_{C_l} for signal cluster C_l ▷ Compute signal-level PIP
end for

4.4.3 Recovering level-ρ credible sets from signal clusters

We can then recover the level-ρ credible set from the signal cluster C by adding variants one by one to the subset Ω in descending order of their posterior probabilities in the effect, until the cumulative sum of the posterior probabilities SPIP_Ω exceeds

the threshold ρ. For the subset of variants Ω, the posterior probability of Ω containing at least one causal variant is given by [11]:

$$\begin{aligned}\text{SPIP}_\Omega &:= \Pr\left(\text{at least one } \gamma_j = 1,\ j \in \Omega \mid \mathbf{G}, \mathbf{y}\right) && (9)\\ &= 1 - \Pr\left(\text{all } \gamma_j = 0,\ j \in \Omega \mid \mathbf{G}, \mathbf{y}\right) && (10)\\ &= 1 - \sum_{\boldsymbol{\gamma}':\gamma_j=0,\ \forall j\in\Omega} p(\boldsymbol{\gamma} = \boldsymbol{\gamma}'|\mathbf{G}, \mathbf{y}). && (11)\end{aligned}$$

The computation of SPIP_Ω is also trivial, when adding one variant k to Ω_0 and getting to the new Ω, the updated SPIP_Ω is computed by adding the posterior probability of models with $\gamma_k = 1,\ \gamma_j = 0,\ \forall j \in \Omega_0$:

$$\text{SPIP}_\Omega = \text{SPIP}_{\Omega_0} + \sum_{\boldsymbol{\gamma}':\gamma_k=1,\ \gamma_j=0,\ \forall j\in\Omega_0} p(\boldsymbol{\gamma} = \boldsymbol{\gamma}'|\mathbf{G}, \mathbf{y}). \quad (12)$$

To put it more explicitly, for current subset Ω_0, we can divide all model configurations into two partitions: $\boldsymbol{\gamma}_0$ with $\gamma_j = 0$, for all $j \in \Omega_0$, and $\boldsymbol{\gamma}_1$ with at least one $\gamma_j = 1,\ j \in \Omega_0$. The SPIP_{Ω_0} is computed as the summation of the model posterior probabilities in the partion $\boldsymbol{\gamma}_1$. After adding a new variant k, we can find models with $\gamma_k = 1$ in the partition $\boldsymbol{\gamma}_0$, and the updated SPIP_Ω is calculated by adding the posterior probabilities of these models with $\gamma_k = 1$ in the partition $\boldsymbol{\gamma}_0$, which is easy to find. This process is repeated until the cumulative sum of the posterior probabilities SPIP_Ω exceeds the threshold ρ, and the level-ρ credible set is obtained by the subset Ω.

The signal-level PIP is also the maximum of the posterior probabilities of the level-ρ credible sets: we can only recover at most the level-ρ credible set with $\rho \leq \text{SPIP}_C$ from the signal cluster C.

4.5 Simulations

4.5.1 Simulation with normal distribution

To evaluate the performance of DAP-PIR, we conduct a simulation study with a normal distribution for the genotype data. Specifically, we consider a scenario with $n = 500$ individuals and $p = 10$ SNPs to relive the computational burden of exact computation of the posterior. Each SNP g_{ij} is generated independently from a standard normal distribution. We randomly select one to five variants and specify them as causal variants with the setting:

$$\begin{aligned}g_{ij} &\sim \text{N}\left(0, 1\right), && (13)\\ y_i &= \sum_{j=1}^{10} \beta_j g_{ij} + \epsilon_i, && (14)\\ \epsilon_i &\sim \text{N}(0, \tau^{-1}), && (15)\\ \beta_j &\sim (1 - \gamma_j)\delta_0 + \gamma_j \,\text{N}\left(0, \sigma_0^2/\tau\right), && (16)\end{aligned}$$

where β_j represents the effect size of the j-th SNP, ϵ_i is the residual term with varaince τ^{-1}, and σ_0^2/τ is the prior effect size variance. The effect sizes β_j are determined by a binary indicator γ_j: if $\gamma_j = 0$, β_j is set to zero and indicates a non-causal SNP; if $\gamma_j = 1$, β_j is drawn from a normal distribution with mean zero and variance σ_0^2/τ, representing a causal SNP. We randomly assign S causal variants, and the indicator vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{10})$ is a 10-vector of binary variables such that only S entries are 1 and the other entries are 0. In the simulation, we set $\tau = 1$, $\sigma_0^2 = 0.6^2$, and choose $S \in \{1, 2, 3, 4, 5\}$. For each setting, we simulate 1000 times. Then we have simulated $1,000 \times 5 = 5,000$ datasets for further investigation.

4.5.2 Simulation on GTEx V8 data

Using the GTEx V8 data [12] with $n = 670$ individual, we select 1,000 genes and choose $p = 1000$ neighboring *cis*-SNPs of each gene. We simulate the phenotypes with various combinations of the number of effects, S, and proportion of variance explained (PVE) by genotypes, ϕ. The simulation settings are as follows:

$$y_i = \sum_{j=1}^{1000} \beta_j g_{ij} + \epsilon_i, \tag{17}$$

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{18}$$

$$\sigma^2 = \frac{1-\phi}{\phi}\mathrm{Var}(\mathbf{G}\boldsymbol{\beta}), \tag{19}$$

$$\beta_j \sim (1-\gamma_j)\delta_0 + \gamma_j \,\mathrm{N}\left(0, 0.6^2\right). \tag{20}$$

Specifically, for each gene, we randomly assign S variants to be causal, while their effect sizes are independently drawn from $\mathrm{N}(0, 0.6^2)$ and set the other effect sizes to zero. The variance of the error term σ^2 is given by $\frac{1-\phi}{\phi}\mathrm{Var}(\mathbf{G}\boldsymbol{\beta})$ and the simulated phenotypes follow $\mathrm{N}(\mathbf{G}\boldsymbol{\beta}, \sigma^2)$. We follow the SuSiE setting and generate data with pairwise combinations of $S \in \{1, 2, 3, 4, 5\}$ and $\phi \in \{0.05, 0.1, 0.2, 0.4\}$. Then we have simulated $1,000 \times 5 \times 4 = 20,000$ datasets for further investigation.

4.6 Real data application

5 Figures and tables

References

[1] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J.: 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics **101**(1), 5–22 (2017)

[2] Ardlie, K.G., Kruglyak, L., Seielstad, M.: Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics **3**(4), 299–309 (2002)

[3] Spain, S.L., Barrett, J.C.: Strategies for fine-mapping complex traits. Human molecular genetics **24**(R1), 111–119 (2015)

[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] Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E.: Identifying causal variants at loci with multiple signals of association. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 610–611 (2014)

[6] 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)

[7] Lee, Y., Luca, F., Pique-Regi, R., Wen, X.: Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. BioRxiv, 316471 (2018)

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

[9] Rubin, D.B.: The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. Journal of the American Statistical Association **82**(398), 543–546 (1987)

[10] Rubin, D.B.: Using the sir algorithm to simulate posterior distributions. In: Bayesian Statistics 3. Proceedings of the Third Valencia International Meeting, 1-5 June 1987, pp. 395–402 (1988). Clarendon Press

[11] Samaddar, A., Maiti, T., Los Campos, G.: Bayesian hierarchical hypothesis testing in large-scale genome-wide association analysis. Genetics, 164 (2024)

[12] Consortium, G., Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., *et al.*: The

genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science **348**(6235), 648–660 (2015)