SuSiE-ASH: A sum of single effects regression and adaptive shrinkage model as a unified framework for fine-mapping and TWAS

Alex McCreight[1,2], Yanghyeon Cho[1,2], William R.P. Denault[1*], and Gao Wang[1,3*]

[1]Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, New York, NY

[2]Department of Biostatistics, Columbia University, New York, NY

[3]Department of Neurology, Columbia University, New York, NY

Abstract

Many Bayesian variable selection methods, such as Sum of Single Effects Regression (SuSiE), are widely used in genetic association studies to identify causal variants influencing molecular phenotypes like gene expression. These methods oftentimes assume a sparse genetic architecture, implying that only a few variants have large effects on the outcome while the remaining have none. This assumption can lead to a high false discovery rate (FDR) in settings where multiple variants contribute to the phenotype, including oligogenic and infinitesimal effects. We introduce "SuSiE-ASH (Random Effects)", a novel method that integrates flexible adaptive shrinkage priors into the SuSiE framework. SuSiE-ASH (RE) offers a flexible alternative to SuSiE by modeling both strong sparse-effect variants and a variety of oligogenic and infinitesimal effects. We conducted extensive simulations reflecting various expression quantitative trait loci (eQTL) settings. Our results demonstrate that SuSiE-ASH (RE) reduces FDR by an absolute difference of 27.7% compared to SuSiE, while simultaneously increasing the recall.

Keywords: fine-mapping, Bayesian variable selection, TWAS, variational inference

1 Introduction

Transcriptome-wide association studies (TWAS) have identified thousands of loci associated with gene expression across various tissues and conditions. However, oftentimes these identified variants are not causal themselves; rather, they are in linkage disequilibrium (LD) with the true causal variants. LD presents a significant challenge in analyzing TWAS results as it obscures which genetic variants actually contribute to gene expression. To address this, significant efforts over the past few decades have focused on developing methods to refine associations and identify truly causal variants (Maller et al. [2012]).

Among these methods, Bayesian fine-mapping methods (Wang et al. [2020], Wakefield [2009], Hormozdiari et al. [2014], Benner et al. [2016], Wen et al. [2016], Lee et al. [2018]) have emerged as powerful tools for refining TWAS associations. They are especially powerful as they provide posterior inclusion probabilities (PIPs), which quantify the likelihood that a given variant is causal, and they generate credible sets, a manageable subset of variants with ρ-level confidence of containing at least one causal variant. However, one limitation of these methods is that many of them assume a sparse genetic architecture.

A sparse genetic architecture assumes that only a small number of genetic variants have large effects on gene expression, while the rest are assumed to have no effect. While this assumption offers computational efficiency and theoretical simplicity, it is often unrealistic and fails to capture the true relationship of non-sparse genetic backgrounds. In such scenarios, many variants individually have small effects on gene expression, yet collectively they can contribute a significant proportion of phenotypic variation (Fisher [1918], Yang et al. [2010], Barton et al. [2017]). This includes oligogenic effects, where each variant contributes a fraction of a phenotypic variance compared to sparse effects, and infinitesimal effects, where individual variants contribute miniscule amounts of variance, yet together account for a substantial proportion of the total. Ignoring these contributions can lead to inflated false discovery rate (FDR) and reduced accuracy in identifying true causal variants (Cui et al.

[2024]).

To address the limitations of current fine-mapping methods in handling non-sparse genetic architectures, we propose a novel method, SuSiE-ASH (Random Effects). This approach integrates Multiple Regression with Adaptive Shrinkage Priors (mr.ASH; Kim et al. [2024]) into the Sum of Single Effects regression (SuSiE; Wang et al. [2020]) framework. By combining the strengths of SuSiE's single effect regression model with the flexibility of the adaptive shrinkage priors, SuSiE-ASH (RE) is designed to accurately model both sparse and non-sparse genetic effects.

We evaluate SuSiE-ASH (RE) in the context of expression quantitative trait loci (eQTL) studies, where the genetic regulation of gene expression often involves a mixture of sparse, oligogenic, and infinitesimal effects (Lloyd-Jones et al. [2017], Grundberg et al. [2012]). Through extensive simulations using UKBB data, we demonstrate that SuSiE-ASH (RE) achieves significantly better FDR compared to SuSiE, while simultaneously improving recall.

2 Methods

2.1 Model

SuSiE-ASH (RE) is based upon the following model:

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a centered $n \times 1$ vector of phenotype, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ is a standardized $n \times p$ matrix of genotypes for p genetic variants in a genomic region of interest, with $\boldsymbol{x}_j$ being the j-th column of $\boldsymbol{X}$, the p-vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ represent strong spare effect and oligogenic and infinitesimal effects, respectively, which are independent of each other, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Here, we construct $\boldsymbol{\beta}$ by summing multiple single-effect vectors to model the strong sparse component Wang et al. [2020]. We assume that precisely L variants have a non-zero effect

on the outcome:

$$\boldsymbol{\beta} \sim \sum_{\ell=1}^{L} \boldsymbol{\beta}^{(\ell)}, \tag{2}$$

$$\boldsymbol{\beta}^{(\ell)} = (b^{\ell} \odot \boldsymbol{\gamma}^{\ell}), \tag{3}$$

$$\boldsymbol{\gamma}^{\ell} \sim Mult(1, \boldsymbol{p}), \tag{4}$$

$$b^{\ell} \sim N(0, \sigma_{0\ell}^2), \tag{5}$$

where $\boldsymbol{\beta}^{(\ell)}$ denotes the ℓ-th single-effect vector, $\boldsymbol{\gamma}^{\ell} = (\gamma_1^{\ell}, \ldots, \gamma_p^{\ell})$ is a p-vector indicating the location of the causal SNP in the ℓ-th single effect, with $\boldsymbol{p} = (p_1, \ldots, p_p)^T$ representing the prior weight that sum to 1, and b^{ℓ} is a scalar representing the causal effect size in the ℓ-th single effect. Note that the single-effect regression (SER) model is a special case of the above-specified model when $L = 1$.

Then we construct $\boldsymbol{\theta}$ using an adaptive shrinkage prior for the scaled coefficients θ_j/σ (Kim et al. [2024]; Stephens [2017]) to model the remaining oligogenic and infinitesimal effects:

$$\frac{\theta_j}{\sigma} \sim \sum_{k=1}^{K} \pi_k N(0, \sigma_k^2), \tag{6}$$

where $\pi = (\pi_1, ...\pi_K)$ represents the mixture proportions which are each non-negative and collectively sum to one, and $\sigma_1^2, ..., \sigma_K^2$ are a non-negative, increasing, pre-specified grid of component variances such that $0 \leq \sigma_1^2, ..., \sigma_K^2 < \infty$ with σ_1^2 set to 0.

2.2 Method

SuSiE-ASH (RE) is a novel approach that improves fine-mapping by combining SuSiE (Wang et al. [2020]) for sparse variable selection and mr.ASH (Kim et al. [2024]) for adapative shrinkage estimation. The key idea is to iteratively update the strong sparse effect using SuSiE, and then apply mr.ASH to the residuals to approximate the remaining oligogenic and infinitesimal effects size and their variance parameters.

Note that both SuSiE and mr.ASH adopt the variational approximation (VA) method (Blei et al. [2017]) to approximate the posterior distribution under their respective models. By assuming a fully factorized variational approximation, they simplify the optimization of the evidence lower bound (ELBO) over joint prior variables, making it tractable. This tractability is achieved by employing the coordinate ascent algorithm (Bertsekas [1999]), which converts the complex joint optimization problem into a series of simpler tasks, such as fitting single-effect regression (SER) models or normal mean (NM) models.

In SuSiE-ASH (RE), we have also embraced this approach, assuming that the approximation of the joint posterior q for $\boldsymbol{\beta}$ is factorized as:

$$q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \prod_{\ell=1}^{L} q_{\boldsymbol{\beta}^\ell}(\boldsymbol{\beta}^\ell). \tag{7}$$

Our proposed Algorithm 1 is iteratively optimizing variational approximations $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ by maximizing the following ELBO under SuSiE-ASH (RE) (1):

$$\begin{aligned} F(\boldsymbol{\beta}; \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det\Lambda + E_q\bigg[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\Lambda(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\bigg] \\ + E_{q_{\boldsymbol{\beta}}}\bigg[\log\frac{g_{\boldsymbol{\beta}}(\boldsymbol{\beta})}{q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}\bigg], \end{aligned} \tag{8}$$

where $\boldsymbol{\sigma}_{0b}^2 = (\sigma_{01}^2, \ldots, \sigma_{0L}^2)$, $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$, $\Lambda = (\nu^2\boldsymbol{X}\boldsymbol{X}^T + \sigma^2\boldsymbol{I})^{-1}$ with $\nu^2 = \mathrm{Var}(\theta_j) = \sum_{k=1}^{K}\pi_k\sigma_k^2$ for all j, and $g_{\boldsymbol{\beta}}$ is the prior distribution of $\boldsymbol{\beta}$. Then, we update the variance components $(\boldsymbol{\pi}, \sigma^2)$ and approximate the best linear unbiased predictor (BLUP) for $\boldsymbol{\theta}$ by fitting mr.ASH to the residuals. For completeness, the following steps illustrate how Algorithm 1 is implemented within an iteration loop:

Updating the strong spare effect $\boldsymbol{\beta}$

Conditioned on variance components σ^2, $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2$ and $\boldsymbol{\pi}$, we update each $q_{\boldsymbol{\beta}^\ell}$, $\ell = 1, \ldots, L$ by fitting the following single-effect regression model:

$$\bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell} := E_{q_{\boldsymbol{\beta}}}\Big[\boldsymbol{y} - \sum_{\ell' \neq \ell}\boldsymbol{X}\boldsymbol{\beta}^{\ell'}\Big] = \boldsymbol{X}\boldsymbol{\beta}^\ell + \tilde{\epsilon}, \tag{9}$$

where $\tilde{\epsilon} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \sim N(0, \Lambda^{-1})$. Denote the posterior distribution for $\boldsymbol{\beta}^\ell = b^\ell \boldsymbol{\gamma}^\ell$ under the SER model (Equation 9) as:

$$\boldsymbol{\gamma}^\ell \mid \boldsymbol{X}, \bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}, \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi} \sim \text{Mult}(1, \boldsymbol{\alpha}^\ell), \tag{10}$$

$$b^\ell \mid \boldsymbol{X}, \bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}, \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi}, \gamma_j^\ell = 1 \sim N(m_{1j}^\ell, \sigma_{1j}^{2,\ell}), \tag{11}$$

where $\boldsymbol{\alpha}^\ell = (\alpha_1^\ell, \ldots, \alpha_p^\ell)$ is the vector of the posterior inclusion probabilities (PIPs). Note that the posterior distribution for the ℓ-th single effect can be obtained by maximizing the following simpler ELBO:

$$\begin{aligned} F_\ell(\boldsymbol{\beta}^\ell; \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi}) &= -\frac{1}{2} E_{q^\ell}\big[-2\bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}^T \Lambda \boldsymbol{X}\boldsymbol{\beta}^\ell + (\boldsymbol{X}\boldsymbol{\beta}^\ell)^T \Lambda \boldsymbol{X}\boldsymbol{\beta}^\ell \big] + E_{q^\ell}\big[\log \frac{g_{\beta^\ell}(\boldsymbol{\beta}^\ell)}{q_{\beta^\ell}(\boldsymbol{\beta}^{\boldsymbol{\ell}})}\big] + c_\ell \\ &= \sum_{j=1}^p \alpha_j^\ell (\bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}^T \Lambda \boldsymbol{x}_j) m_{1j}^\ell - \frac{1}{2} \sum_{j=1}^p \alpha_j^\ell \boldsymbol{x}_j^T \Lambda \boldsymbol{x}_j (m_{1j}^{2,\ell} + \sigma_{1j}^{2,\ell}) \\ &\quad + \sum_{j=1}^p \frac{\alpha_j^\ell}{1} \left[1 + \log \frac{\sigma_{1j}^{2,\ell}}{\sigma_{0\ell}^2} - \frac{m_{1j}^{2,\ell} + \sigma_{1j}^{2,\ell}}{\sigma_{0\ell}^2} \right] + \sum_{j=1}^p \alpha_j^\ell \log \frac{p_j}{\alpha_j^\ell} + c_\ell, \end{aligned}$$

where c_ℓ is the sum of terms that are not dependent on q^ℓ. Then, by taking partial derivatives of $F_\ell(\boldsymbol{\beta}^\ell; \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi})$ with respect to parameters $\boldsymbol{\alpha}^\ell$, $\boldsymbol{m}_1^\ell = (m_{11}^\ell, \ldots, m_{1p}^\ell)$, and $\boldsymbol{\sigma}_1^{2,\ell} = (\sigma_{11}^{2,\ell}, \ldots, \sigma_{1p}^{2,\ell})$, the explicit formulas for this update are expressed as:

$$\alpha_j^\ell = \frac{p_j \exp\left(\log \frac{\sigma_{1j}^\ell}{\sigma_{0\ell}} + \frac{m_{1j}^{2,\ell}}{2\sigma_{1j}^{2,\ell}} \right)}{\sum_{j'=1}^p p_{j'} \exp\left(\log \frac{\sigma_{1j'}^\ell}{\sigma_{0\ell}} + \frac{m_{1j'}^{2,\ell}}{2\sigma_{1j'}^{2,\ell}} \right)} \tag{12}$$

and

$$m_{1j}^\ell = \bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}^T \Lambda \boldsymbol{x}_j \times \sigma_{1j}^{2,\ell} \text{ and } \sigma_{1j}^{2,\ell} = \left[\boldsymbol{x}_j^T \Lambda \boldsymbol{x}_j + 1/\sigma_{0\ell}^2 \right]^{-1}. \tag{13}$$

For the brevity, we introduce the following function that returns arguments of the posterior distribution of $\boldsymbol{\beta}^\ell$:

$$\text{SER}(\bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}, \boldsymbol{X}; \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi}) = (\boldsymbol{\alpha}^\ell, \boldsymbol{m}_1^\ell, \boldsymbol{\sigma}_1^{2,\ell}). \tag{14}$$

Updating variance components and the BLUP for $\boldsymbol{\theta}$

To update the variance components, $\boldsymbol{\pi}$ and σ^2, one could theoretically rely on the marginal distribution over $\boldsymbol{\beta}$. However, this approach is computationally demanding, particularly for large datasets. As a practical alternative, we leverage the output of the mr.ASH model, conveniently implemented using the existing `R` function `mr.ash.alpha`. This update involves fitting the mr.ASH model to the residuals after removing the updated sparse effect, $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)$, from $\boldsymbol{y}$, denoted as $\bar{\boldsymbol{r}}_{\boldsymbol{\theta}} = \boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$. Here, each $\bar{\beta}_j$ is computed as $\bar{\beta}_j = \sum_{\ell=1}^{L} \alpha_j^{(\ell)} m_{1j}^{\ell}$.

In a similar manner to SuSiE, mr.ASH employs a coordinate-ascent mean field variational inference: (1) the variational distribution of $\boldsymbol{\theta}$ is factorized as $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \prod_{j=1}^{p} q_{\boldsymbol{\theta}_j}(\theta_j)$; (2) the coordinate ascent update for each θ_j, $j = 1, \ldots, p$, is given by computing a posterior distribution under the following normal mean model:

$$\bar{\boldsymbol{r}}_{\theta_j} := E_{q_{\boldsymbol{\theta}_{-j}}}\left[\bar{\boldsymbol{r}}_{\boldsymbol{\theta}} - \sum_{j' \neq j} \boldsymbol{x}_{j'}\theta_{j'}\right] = \bar{\boldsymbol{r}}_{\boldsymbol{\theta}} - \sum_{j' \neq j} \boldsymbol{x}_{j'}\bar{\mu}_{j'} = \boldsymbol{x}_j\theta_j + \epsilon, \tag{15}$$

where $(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p) \sim q_{\boldsymbol{\theta}_{-j}}$. Then, by Kim et al. [2024], the posterior distribution for $\boldsymbol{\theta}_j$, denoted as q_{θ_j} under the normal mean model is given by:

$$q(\theta_j|\tilde{\theta}_j; \sigma^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2) = \sum_{k=1}^{K} \phi_{1jk} N(\mu_{1jk}, s_{1jk}^2), \tag{16}$$

where

$$\mu_{1jk} = \frac{\sigma^2\sigma_k^2}{\sigma^2 + \sigma^2\sigma_k^2}\tilde{\theta}_j, \tag{17}$$

$$s_{1jk}^2 = \frac{(\sigma^2)^2\sigma_k^2}{\sigma^2 + \sigma^2\sigma_k^2}, \tag{18}$$

$$\phi_{1jk} = \frac{\pi_k L_{jk}}{\sum_{k=1}^{K} \pi_k L_{jk}}, \tag{19}$$

with $\tilde{\theta}_j = (\boldsymbol{x}_j^T\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^T\bar{\boldsymbol{r}}_{\theta_j}$ and $L_{jk} = N(\tilde{\theta}_j; 0, \sigma^2 + \sigma^2\sigma_k^2)$. For future reference, we define the function $\mathrm{NM}^{\mathrm{post}}$, which returns the estimated parameters for the posterior distribution of θ_j under the normal mean model (Equation 15):

$$\mathrm{NM}^{\mathrm{post}}\left(\bar{\boldsymbol{r}}_{\theta_j}, \boldsymbol{X}; \sigma^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2\right) = (\boldsymbol{\mu}_{1j}, \boldsymbol{s}_{1j}^2, \boldsymbol{\phi}_{1j}), \tag{20}$$

where $\boldsymbol{\mu}_{1j} = (\mu_{1j1}, \ldots, \mu_{1jK})$, $\boldsymbol{s}^2_{1j} = (s^2_{1j1}, \ldots, s^2_{1jK})$, and $\boldsymbol{\phi}_{1j} = (\phi_{1j1}, \ldots, \phi_{1jK})$.

Given Equations 17–19, the mixture proportion $\boldsymbol{\pi}$ can be updated as:

$$\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_K),$$

where $\tilde{\pi}_k = \frac{1}{p}\sum_{j=1}^{p} \phi_{1jk}$. The residual variance can be updated as:

$$\tilde{\sigma}^2 = \frac{\|\bar{\boldsymbol{r}}_\theta - \boldsymbol{X}\bar{\boldsymbol{\theta}}\|^2 + \sum_{j=1}^{p}\sum_{k=2}^{K} \phi_{1jk}(z_j + 1/\sigma_k^2)(\mu_{1jk}^2 + s_{1jk}^2) - \sum_{j=1}^{p} z_j \bar{\mu}_j^2}{n + p \times (1 - \pi_1)},$$

where $\|\boldsymbol{u}\| = \sqrt{\boldsymbol{u}^T\boldsymbol{u}}$ denotes the L_2-norm of vector $\boldsymbol{u} \in \mathcal{R}^n$, $z_j = \boldsymbol{x}_j^T\boldsymbol{x}_j$, and $\bar{\boldsymbol{\theta}} = (\bar{\mu}_1, \ldots, \bar{\mu}_p)$, with $\bar{\mu}_j = \sum_{k=1}^{K} \phi_{1jk}\mu_{1jk}$, which is the approximation of the BLUP for $\boldsymbol{\theta}$ by Equation 16.

Remark The convergence criterion for iterations is based on the maximum difference between the posterior inclusion probabilities of the previous and current iterations.

Algorithm 1 SuSiE-ASH (RE) Algorithm

Require: $\mathbf{y}, \mathbf{X}, L, K, \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi}$

Initial estimates $\bar{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\theta}}$

while Not converged **do**

  for $\ell = 1$ to L **do**

    $\bar{\mathbf{r}}_{\boldsymbol{\beta},\ell} = \boldsymbol{y} - \mathbf{X}\sum_{\ell' \neq \ell} \bar{\boldsymbol{\beta}}^{\ell'}$

    $(\boldsymbol{\alpha}^\ell, \boldsymbol{m}_1^\ell, \boldsymbol{\sigma}_1^{2,\ell}) \leftarrow \mathrm{SER}(\bar{\boldsymbol{r}}_{\boldsymbol{\beta},\ell}, \boldsymbol{X}; \sigma^2, \boldsymbol{\sigma}_{0b}^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2, \boldsymbol{\pi})$

    $\bar{\boldsymbol{\beta}}^\ell \leftarrow (\boldsymbol{\alpha}^\ell)^T \boldsymbol{m}_1^\ell$

  end for

  for $j = 1$ to p **do**

    $\bar{\boldsymbol{r}}_{\theta_j} = \boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}} - \sum_{j' \neq j} \boldsymbol{x}_{j'} \bar{\mu}_{j'}$

    $(\boldsymbol{\mu}_{1j}, \boldsymbol{s}_{1j}^2, \boldsymbol{\phi}_{1j}) \leftarrow \mathrm{NM}^{\mathrm{post}}(\bar{\boldsymbol{r}}_{\theta_j}, \mathbf{X}; \sigma^2, \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2)$

    $\bar{\mu}_j \leftarrow \sum_{k=1}^K \phi_{1jk}\mu_{1jk}$

  end for

  $\boldsymbol{\pi} \leftarrow (\sum_{j=1}^p \phi_{1j1}, \ldots, \sum_{j=1}^p \phi_{1jK})/p$

  $\sigma^2 \leftarrow \frac{\|\bar{\boldsymbol{r}}_\theta - \boldsymbol{X}\bar{\boldsymbol{\theta}}\|^2 + \sum_{j=1}^p \sum_{k=2}^K \phi_{1jk}(z_j + 1/\sigma_k^2)(\mu_{1jk}^2 + s_{1jk}^2) - \sum_{j=1}^p z_j \bar{\mu}_j^2}{n + p \times (1 - \pi_1)}$

end while

Output: $\boldsymbol{\alpha}^1$, $\boldsymbol{m}_1^1$, $\boldsymbol{\sigma}_1^{2,1}$, ...,$\boldsymbol{\alpha}^L$, $\boldsymbol{m}_1^L$, $\boldsymbol{\sigma}_1^{2,L}$, $\bar{\mu}_1, \ldots, \bar{\mu}_p$, $\boldsymbol{\pi}$, σ^2

3 Results

3.1 Simulation Design

We conducted extensive simulations to evaluate the performance of SuSiE-ASH (RE) under non-sparse genetic architectures, reflecting realistic scenarios where genetic variation is driven by a mixture of sparse, oligogenic, and infinitesimal effects (Lloyd-Jones et al. [2017]). These simulations used genotype data from UK Biobank. We randomly sampled 200 LD blocks across chromosomes 1-22 and derived the associated genotype matrices to serve as the basis for our simulations. Each genotype matrix was required to have between 5000 and 8000 variants, a minor allele frequency greater than 1% and a missing rate below 5%. We further refined these matrices by randomly sampling only 10000 individuals, removing variants with zero or near-zero variance, and using mean imputation for any missing data. Furthermore, acknowledging that each variant in our simulation influences the outcome, we define causal variants as those contributing 0.5% of total phenotypic variance.

3.2 Phenotype Simulation

To model a complex genetic architecture, we generated phenotypes incorporating sparse ($\boldsymbol{\beta}$), oligogenic ($\boldsymbol{\phi}$), and infinitesimal ($\boldsymbol{\theta}$) genetic effects based on the following linear model:

$$\mathbf{y} = \mathbf{X}\left(\beta_{\text{sentinel}} + \sum_{j\in\mathcal{S}} \beta_j + \sum_{k\in\mathcal{O}} \phi_k + \sum_{l\in\mathcal{I}} \theta_l\right) + \boldsymbol{\epsilon}, \tag{21}$$

where $\mathbf{y}$ is the phenotype vector, $\mathbf{X}$ is the standardized genotype matrix, β_{sentinel} is the effect size of the sentinel SNP, $\mathcal{S}$ is the set of indices for other sparse SNPs, $\mathcal{O}$ is the set of indices for oligogenic SNPs, $\mathcal{I}$ is the set of indices for infinitesimal SNPs, and $\boldsymbol{\epsilon}$ is the noise vector, such that $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$.

3.3 Effect Size Distributions

The effect sizes were drawn from the following distributions:

$$\beta_{\text{sentinel}} \sim N(0, \sigma^2_{\text{sentinel}}) \tag{22}$$

$$\beta_j \sim N\left(0, \frac{\sigma^2_{\text{other\_sparse}}}{n_{\text{other\_sparse}}}\right), \forall j \in \mathcal{S} \tag{23}$$

$$\phi_k \sim \begin{cases} N(0, \sigma^2_{\phi_1}), & \text{with probability } \pi_1, \\ N(0, \sigma^2_{\phi_2}), & \text{with probability } \pi_2, \end{cases}, \forall k \in \mathcal{O} \tag{24}$$

$$\text{where } \pi_1 + \pi_2 = 1 \text{ and } \sigma^2_{\phi_1} > \sigma^2_{\phi_2} > 0 \tag{25}$$

$$\theta_l \sim N\left(0, \frac{\sigma^2_\theta}{n_{\text{infinitesimal}}}\right). \tag{26}$$

3.4 Heritability

The total narrow-sense heritability (h^2_{total}) represents the proportion of phenotypic variance explained by all genetic factors. This total heritability was partitioned among the genetic components as follows:

$$h^2_{\text{total}} = h^2_{\text{sparse}} + h^2_{\text{oligogenic}} + h^2_{\text{infinitesimal}} \tag{27}$$

$$h^2_{\text{sparse}} = h^2_{\text{total}} \times 0.65 \tag{28}$$

$$h^2_{\text{oligogenic}} = h^2_{\text{total}} \times 0.20 \tag{29}$$

$$h^2_{\text{infinitesimal}} = h^2_{\text{total}} \times 0.15, \tag{30}$$

The sparse heritability was further divided between the sentinel SNP and other sparse SNPs

$$h^2_{\text{sentinel}} = h^2_{\text{sparse}} \times 0.7 \tag{31}$$

$$h^2_{\text{other\_sparse}} = h^2_{\text{sparse}} - h^2_{\text{sentinel}} \tag{32}$$

We scaled all effects to achieve their desired heritability proportions.

3.5 Simulation Parameters

We generated simulated data sets under many different parameter configurations. Specifically, we generated data sets using all pairwise combinations of $h^2_{\text{total}} = \{0.3, 0.5\}$, $|\mathcal{O}| = \{10, 20, 30\}$, and $\boldsymbol{\pi} = \{(0.5, 0.5)^T, (0.75, 0.25)^T, (0.25, 0.75)^T, (0.9, 0.1)^T\}$ where $\boldsymbol{\pi} = (\pi_1, \pi_2)^T$. We simulated a single replicate for each genotype matrix and for each combination of h^2_{total}, $|\mathcal{O}|$, and $\boldsymbol{\pi}$ resulting in a total of $200 \times 2 \times 3 \times 4 = 4800$ data sets.

3.6 Simulation Results

In our analysis, we compare the FDR and Recall across various SuSiE model variants, as illustrated in Figure 1. The SuSiE-ASH (RE) model using a default grid parametrization achieves the best FDR at 5.7%, at the cost of being too conservative. So, while this model achieves a high level of precision, it only captures about half of all truly causal variants. Conversely, the original SuSiE model exhibits the highest FDR at 51.6%, while capturing about 71.7% of causal variants. So, while SuSiE has drastically improved recall over SuSiE-ASH (RE; default grid), over half of its 95% credible sets contain false positives, raising major concerns about the reliability of all its discoveries.

To address this trade off, between high versus low FDR and Recall, the SuSiE-ASH (RE) model using the quadratic grid parametrization struck a more balanced performance with an FDR of 23.9% and recall of 75%, achieving a significant reduction in false positives while capturing more causal variants than SuSiE. While the FDR is still above the conventional 0.05 threshold, this model represents a promising compromise between minimizing false discoveries and maintaining reasonable recall. SuSiE-Inf (Cui et al. [2024]) was also able to significantly reduce the number of false positives with an FDR of 9.4%, while maintaining a moderate recall of 63.2%, but due to the conservative nature of the model, it is unable to capture the majority of the oligogenic effects making it much less flexible than SuSiE-ASH (RE).

In addition to FDR and Recall, we also evaluated model prediction performance using root mean-squared error (RMSE), as shown in Figure 2. Across all the models, SuSiE-Inf achieved the lowest RMSE (0.814), closely follwed by SuSiE-ASH (RE) using a default grid parametrization at 0.818 and the original SuSiE model at 0.821. The SuSiE-ASH (RE) model using a quadratic grid, demonstrated the highest RMSE at 0.85. Thus, while the quadratic grid parametrization improves the balance of FDR and recall, it slightly sacrifices overall prediction accuracy in effect size estimation compared to the other models.

4 Discussion

In this study, we propose SuSiE-ASH (RE), which demonstrates significant improvements in FDR and slight improvements in recall compared to the original SuSiE model. Our primary contribution lies in the flexibility of our model that we gain from introducing adaptive shrinkage priors into the SuSiE framework. SuSiE-ASH (RE) will enable researchers to balance the amount of true positives and false positives according to their specific objectives, making SuSiE-ASH (RE) a versatile tool for TWAS. For instance, it can be configured to be highly conservative, minimizing the number of false positives, or to allow a higher number of false positives in exchange for capturing many more causal variants than the original SuSiE algorithm. Another minor contribution is the simplification of computations by leveraging the residuals updated in each iteration, thereby avoiding the computationally intensive marginal likelihood for variance components, while retaining the accuracy required for their updates.

The choice of our variance grid for our ASH component is instrumental when using SuSiE-ASH (RE) and more work will be done to further investigate its optimal settings. Currently, we can alter the number of components in the grid, the spacing of the components (e.g., whether they are equally spaced or dense towards zero), the frequency of grid updates (e.g., updating at each iteration until convergence or only during the first iteration), and the

selection of an upper bound for the grid. Future work will explore these aspects in greater details to further improve the balance between FDR and Recall across multiple different simulation settings.

References

N. H. Barton, A. M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73, Dec 2017. doi: 10.1016/j.tpb.2017.06.001. URL https://doi.org/10.1016/j.tpb.2017.06.001. Epub 2017 Jul 11.

C. Benner, C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 01 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw018. URL https://doi.org/10.1093/bioinformatics/btw018.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

R. Cui, R. A. Elzur, M. Kanai, J. C. Ulirsch, O. Weissbrod, M. J. Daly, B. M. Neale, Z. Fan, and H. K. Finucane. Improving fine-mapping by modeling infinitesimal effects. *Nature Genetics*, 56(1):162–169, Jan 2024. doi: 10.1038/s41588-023-01597-3. URL https://doi.org/10.1038/s41588-023-01597-3. Epub 2023 Nov 30.

R. Fisher. The correlation between relatives on the supposition of mendelian inheritance. trans. roy. soc. 1918.

E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka,

V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector, and M. T. H. E. R. M. Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10): 1084–1089, Oct 2012. doi: 10.1038/ng.2394. URL `https://doi.org/10.1038/ng.2394`. Epub 2012 Sep 2.

F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, Oct 2014. doi: 10.1534/genetics.114.167908. URL `https://doi.org/10.1534/genetics.114.167908`. Epub 2014 Aug 7.

Y. Kim, W. Wang, P. Carbonetto, and M. Stephens. A flexible empirical bayes approach to multiple linear regression and connections with penalized regression. *Journal of Machine Learning Research*, 25(185):1–59, 2024.

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

L. R. Lloyd-Jones, A. Holloway, A. McRae, J. Yang, K. Small, J. Zhao, B. Zeng, A. Bakshi, A. Metspalu, M. Dermitzakis, G. Gibson, T. Spector, G. Montgomery, T. Esko, P. M. Visscher, and J. E. Powell. The genetic architecture of gene expression in peripheral blood. *American Journal of Human Genetics*, 100(2):228–237, Feb 2017. doi: 10.1016/j.ajhg.2016.12.008. URL `https://doi.org/10.1016/j.ajhg.2016.12.008`. Epub 2017 Jan 5. Erratum in: Am J Hum Genet. 2017 Feb 2;100(2):371. doi: 10.1016/j.ajhg.2017.01.026.

J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M. M. Howson, A. Auton, S. Myers, A. Morris, M. Pirinen, M. A. Brown, P. R. Burton, M. J. Caulfield, A. Compston, M. Farrall, A. S. Hall, A. T. Hattersley, A. V. S. Hill, C. G. Mathew, M. Pembrey,

J. Satsangi, M. R. Stratton, J. Worthington, N. Craddock, M. Hurles, W. Ouwehand, M. Parkes, N. Rahman, A. Duncanson, J. A. Todd, D. P. Kwiatkowski, N. J. Samani, S. C. L. Gough, M. I. McCarthy, P. Deloukas, P. Donnelly, and T. W. T. C. C. Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294–1301, Dec 2012. ISSN 1546-1718. doi: 10.1038/ng.2435. URL `https://doi.org/10.1038/ng.2435`.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

J. Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, 33(1):79–86, Jan 2009. doi: 10.1002/gepi.20359. URL `https://doi.org/10.1002/gepi.20359`.

G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens. 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.

X. Wen, Y. Lee, F. Luca, and R. Pique-Regi. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *American Journal of Human Genetics*, 98(6):1114–1129, Jun 2016. doi: 10.1016/j.ajhg.2016.03.029. URL `https://doi.org/10.1016/j.ajhg.2016.03.029`. Epub 2016 May 26.

J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

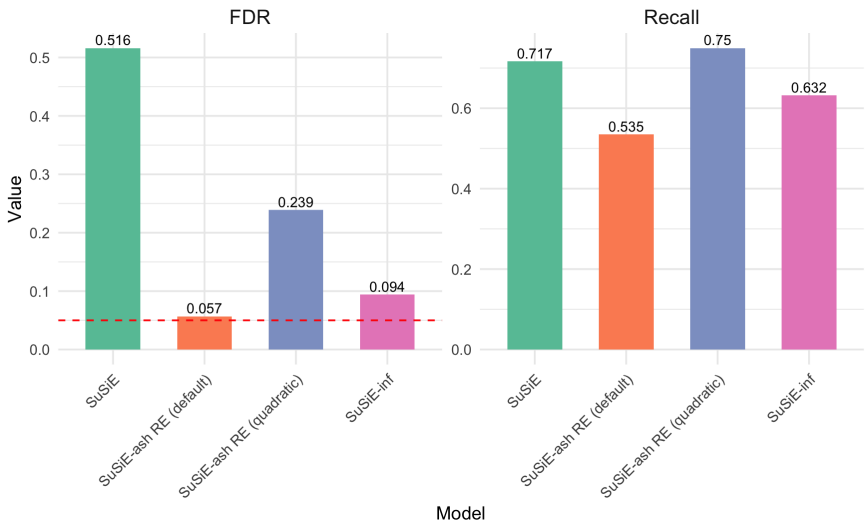

Fig. 1: Comparison of different SuSiE model variants by FDR, defined as the proportion of 95% credible sets without a causal variant, and Recall, defined as the proportion of causal variants captured

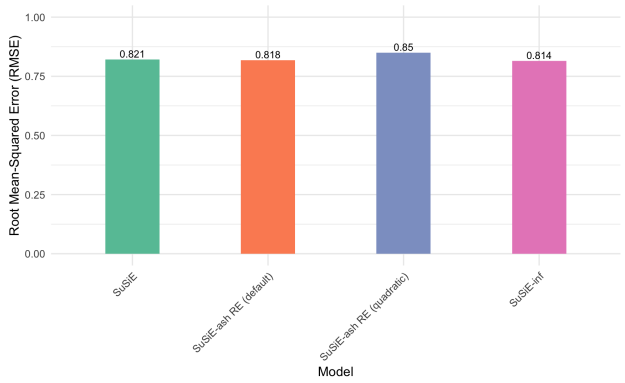

Fig. 2: Comparison of different SuSiE model variants by RMSE