Comparison of Multiple Regression Methods on Data Sets with Varying Degrees of Correlation

by

Lijia Wang

Presented to the Department of Statistics
in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science

University of Chicago
Chicago, IL

April 2021

1. Abstract

Multiple regression is a popular method to identify the relationship between predictor and response variables, which allows for its application on prediction and variable selection problems. However, variable selection with common multiple regression methods is frequently problematic on data sets where variables are highly correlated. Since these methods cannot distinguish between multiple highly correlated predictors, they frequently choose one arbitrarily as the true effect variable. This property could undermine the power of multiple regression methods in variable selection. This issue also limits the potential application of these multiple regression methods to contexts where it is important to identify the causal variable, such as finding a gene that causes increased disease risk. We aim to compare and identify the most accurate variable selection method for data of different degrees of correlation.

We evaluate the variable selection ability of two multiple regression methods on simulated data sets with varying degrees of correlation. The Sum of Single Effects (SuSiE) model is designed for settings where variables are highly correlated and detectable effects are sparse, while Multiple Regression Adaptive Shrinkage (Mr.ASH) is predicted to perform better on data sets with less highly correlated variables. Our results show that when true effects are sparse, Mr.ASH has higher accuracy in variable selection when the maximum correlation of the data set is low, but SuSiE outperforms Mr.ASH at all levels of correlation. We also show that Mr.ASH initialization with LASSO can make up the power difference between Mr.ASH and SuSiE at moderate correlation.

2. Introduction

Multiple linear regression is a fundamental model used for prediction, forecasting, or explaining variance in the response variable when there are several explanatory variables, and is applied to a wide range of topics [4]. One of its common application in statistical genetics is to model the effect of genetic variants, such as single nucleotide polymorphisms (SNPs), on a phenotype of interest. For example, with the recent growth in quantity and sample size of genome wide association studies (GWAS), identifying genetic variants with a significant effect on phenotypes has become an important step to narrowing down the observations obtained from GWAS studies and enabling biologists to develop targeted gene therapies [1]. Many methods, such as step-wise regression (forward selection, backward elimination), calculation of information criteria (Aikaike information criterion, Bayesian information criterion), and penalized likelihood (LASSO, and Elastic Net regression), take different approaches towards selecting a subset of the most significant variables associated with change in observation [5], [12].

Currently, variable selection through multiple linear regression faces numerous challenges. With the increasing size of genomics data such as in GWAS and expression quantitative trait loci (eQTL) studies, we face the problem where the number of predictors, such as SNPs, frequently exceed the number of observations (samples) [6]. This leads to over-fitting and the inclusion of many nonessential predictors in the final model. To solve this problem, many penalized regression methods, such as LASSO, are capable of selecting for variables with nonzero effect sizes with a penalty function [9]. However, in scenarios where variables are highly correlated, variable selection methods such as LASSO arbitrarily select one of the highly correlated variables present [10]. While the method still provides a good model fit, this arbitrary selection limits the potential of these methods when applied to problems such as genetic fine-mapping. In human genotype data, many SNPs are highly correlated due to linkage disequilibrium (LD), and the true causal SNPs could be left out by methods such as the LASSO: when causal SNP are in high LD with non-causal SNPs, the causal SNPs can potentially be discarded while the non-causal SNPs are arbitrarily selected and included in the model. In most studies in genetics, the researcher would like to identify the true causal SNPs so that they can be targeted for future experiments, either for understanding disease mechanism or for developing targeted gene therapy. This implies a need for variable selection methods that can select as many of the correct causal variants while making as few false discoveries as possible.

Although it is difficult to distinguish highly correlated variables, one approach around this problem involves assigning the proper uncertainty for each of the variables selected [10]. The "Sum of Single Effects" (SuSiE) model generates credible sets (CS) for selected variables that are highly correlated to each other while contributing to the same effect. By selecting sets instead of variables for each effect, this approach avoids the issue of leaving out the potentially causal variable and keeping variables that are in high correlation with it instead [10]. Another quantity SuSiE provides for each variable is the posterior inclusion probability (PIP), which summarizes the likelihood of this variable being in the true model. Although PIPs do not provide the same amount of information as the CSs, the PIP values can still summarize the posterior distribution for each variable. This means that, in general, we expect to see higher PIP values for variables that are more likely to be included in the true model [7][11].

Multiple Regression Adaptive Shrinkage (Mr.ASH) employs an empirical Bayes (EB) approach for large-scale multiple linear regression, and is comparable in computational speed and often superior in prediction accuracy than fast penalized regression methods such as LASSO [6]. Similar to penalized regression methods such as LASSO, Mr.ASH also arbitrarily chooses one of the highly correlated variables when they are selected from the data. However, we can estimate the PIP from Mr.ASH calculations, which allows it to be directly comparable to SuSiE. Since our goal is to evaluate the power of variable selection methods on data sets of

varying degrees of correlation, PIP values will be a good metric to rank variables, allowing us to set PIP thresholds to determine the set of selected variable for both methods.

Our experiments aim to evaluate and compare SuSiE and Mr.ASH on their ability to perform variable selection on data sets with different levels of correlation. We use SNP genotypes from the Genotype-Tissue Expression project v8 data set as our genotype matrix [2]. Phenotypic observation of gene expression is then simulated after fixing number of true effect variables and proportion of variance explained (PVE) by these true effects. Different levels of LD between variants are achieved by trimming the variables in high correlation with each other. We subsequently fit SuSiE and Mr.ASH on the resulting data sets and compare their accuracy in variable selection by comparing their power at low false discovery rate (FDR).

An additional goal of our experiment is to find out if there exists a threshold of correlation, above which SuSiE outperforms Mr.ASH. Since SuSiE is designed for highly correlated and sparse effects, we hypothesize that SuSiE would outperform Mr.ASH when variables in the data set are in high LD. When variables are in low LD, we expect SuSiE to have a similar performance compared to Mr.ASH, if not worse. By identifying the degree of correlation that each of these respective methods performs best at, we hope our results can be used to select the method most appropriate for distinct data set features.

3. Background

3.1. **Simple Multiple Linear Regression Model.**

We model the relationship between genetic variants and the observed phenotype as a multiple regression:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \\ \boldsymbol{e} &\sim N_n(0, \sigma^2 I_n), \end{aligned} \tag{3.1}$$

where $\boldsymbol{y}$ is our n-vector observed data, $\boldsymbol{X}$ is the $n \times p$ matrix of n observations of p variables, $\boldsymbol{b}$ is a p-vector of predictors coefficients, and $\boldsymbol{e}$ is an n-vector of error terms.

When simulating the data, we take $\boldsymbol{X}$ matrix values directly from the processed GTEx data sets. For each simulation, we fix the number of true effect variables as well as the proportion of variance explained (PVE) of the true predictors. This simulation can be interpreted as making the additive effects of predictors sum to the heritability of our observed phenotype. This implies that when PVE is fixed, the more true effects we simulate, the smaller the effect size of each true effect variable, and the more difficult it is for variable selection methods to identify the correct variables for a given number of samples.

In our experiment, we use posterior inclusion probability (PIP) to rank variables and determine if variables are selected to be included in the model by the two variable selection methods. If a variable is assigned a

PIP above the threshold, we conclude that the data favors the inclusion of this variable in the regression. The PIP is defined as:

$$\text{PIP}_j := \Pr(b_j \neq 0 | \boldsymbol{X}, \boldsymbol{y}) = 1 - \Pr(b_j = 0 | \boldsymbol{X}, \boldsymbol{y}), \tag{3.2}$$

where b_j refers to the coefficient of the j-th variable in the regression. The larger the PIP_j, the more likely it is for the j-th variable to be included in the true model. In the following sections, we will discuss how PIPs are estimated in SuSiE and Mr.ASH models, the shortcomings of PIPs as a ranking measurement, and comparable alternatives to PIP measurements.

3.2. **Sum of Single Effects Regression (SuSiE).**

Under the SuSiE model in Wang et. al. (2020), the prior distribution for the effect vector $\boldsymbol{b}$ is as follows:

$$\begin{aligned} \boldsymbol{b} &= \sum_{l=1}^{L} \boldsymbol{b}^{(l)} \\ \boldsymbol{b}^{(l)} &= \boldsymbol{\gamma}^{(l)} \beta^{(l)} \\ \boldsymbol{\gamma}^{(l)} &\sim \text{Mult}(1, \boldsymbol{\pi}) \\ \beta^{(l)} &\sim \mathcal{N}(0, \sigma_0^{2(l)}). \end{aligned} \tag{3.3}$$

Here, $\beta^{(l)}$ is a scalar representing the size of the "single effect", $\boldsymbol{\gamma}^{(l)} \in \{0,1\}^p$ is a p-vector of indicator variables, and $\text{Mult}(1, \boldsymbol{\pi})$ is the multinomial distribution on class counts when 1 sample is drawn with class probabilities given by $\boldsymbol{\pi}$. Each vector $\boldsymbol{\gamma}^{(l)}$ has exactly one non-zero entry, therefore each vector $\boldsymbol{b}^{(l)}$ also only have one component that is non-zero, representing the "single effect". L is specified as the maximum number of effects allowed in the model. Since some of the $\boldsymbol{b}^{(l)}$ vectors will have the same non-zero coordinate, the number of non-zero elements in $\boldsymbol{b}$ might be less than L. The sum of single effects, $\boldsymbol{b}$, is the sum of all $\boldsymbol{b}^{(l)}$ vectors.

In our experiment, we aim to obtain the PIP for each predictor j using the SuSiE regression. Under the SuSiE model, each predictor j has a corresponding coefficient $b_j := \sum_{l=1}^{L} b_j^{(l)}$. Assuming that $b_j^{(l)}$ are independent across l, the PIPs can be calculated as:

$$\widehat{\text{PIP}}_j = 1 - \prod_{l \in \mathcal{L}} (1 - \boldsymbol{\alpha}_j^{(l)}), \tag{3.4}$$

where $\mathcal{L} := \{I : \sigma_0^{2(l)} > 0\}$, and $\boldsymbol{\alpha}^{(l)} = (\alpha_1^{(l)}, \alpha_2^{(l)}, \ldots, \alpha_p^{(l)})$ is a p-vector of PIPs, defined as:

$$\alpha_j = \Pr(\gamma_j = 1 | \boldsymbol{X}, \boldsymbol{y}, \sigma^2, \sigma_0^2). \tag{3.5}$$

To obtain the $\widehat{\text{PIP}}_j$ for each variable j, we need to estimate $\boldsymbol{\alpha}$ and σ_0^2 by fitting each SER model. This process is carried out using the Iterative Bayesian stepwise selection (IBSS) algorithm [10]. For each component $l = 1, 2, \ldots, L$, the algorithm calculates the expected residual without the l-th single effect, fits the single effect regression (SER):

$$\text{SER}(\boldsymbol{X}, \boldsymbol{y}; \sigma^2, \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2), \tag{3.6}$$

and returns a vector of $(\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$. Fitting the SuSiE model returns $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\mu}_1^{(1)}, \boldsymbol{\sigma}_1^{2(1)}, \ldots, \boldsymbol{\alpha}^{(L)}, \boldsymbol{\mu}_1^{(L)}, \boldsymbol{\sigma}_1^{2(L)}$. By maximizing the likelihood in the SER model, we can also estimate the hyperparameter $\sigma_0^{2(l)}$ for each component l. The SuSiE program directly outputs PIPs for downstream comparisons.

The benefit of using SuSiE in variable selection is that it is capable of assigning the appropriate PIP values to highly correlated variables. For example, if one of the two highly correlated predictors, x_1 and x_2, have a large effect, SuSiE will place them in the same credible set, and assign a PIP of 0.5 to both of them. Our study, however, does not explore the credible set generation aspect of SuSiE's variable selection abilities in depth. We are interested in predictors that were assigned a large PIP by SuSiE and Mr.ASH, and in testing if PIP is a good metric to judge the power of these variable selection methods.

3.3. **Multiple Regression Adaptive Shrinkage (Mr.ASH).**

Mr.ASH utilizes the same multiple regression model outlined in equation (3.1). However, Mr.ASH assumes a prior distribution where the scaled regression coefficients, b_j/σ are independent and identically distributed (i.i.d.) from some distribution with density g [6]:

$$b_j | g, \sigma^2 \overset{i.i.d.}{\sim} g_\sigma(\cdot). \tag{3.7}$$

Specifically, Kim et. al.(2020) assume priors g are scale mixtures of normals:

$$b_j | g, \sigma^2 \overset{i.i.d.}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\cdot; 0, \sigma^2 \sigma_k^2), \tag{3.8}$$

where k denote the mixtures, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ denotes a set of mixture proportions, and $\sigma_1^2, \ldots, \sigma_K^2$ denotes the component variances of a normal mixture. Kim et. al. detailed the assumptions as:

$$\begin{aligned} \Pr(\gamma_j = k | g) &= \pi_k \\ b_j | g, \gamma_j = k &\sim \mathcal{N}(\cdot | 0, \sigma^2 \sigma_k^2), \end{aligned} \tag{3.9}$$

where $\gamma_j \in \{1, \ldots, K\}$ indicates which mixture component gives rise to b_j.

The goal is to fit the multiple linear regression model outlined by equation (3.1) and (3.7), and to obtain the PIPs for each predictor, as we have done with the SuSiE model. We will subsequently use the PIPs estimated by both methods to perform variable selection. Mr.ASH fits the model using a Variational Empirical Bayes (VEB) approach, first obtaining the estimate of prior density and error variance, $\hat{g}$ and $\hat{\sigma}^2$, and then generating a posterior distribution [6] :

$$p(b_j|\boldsymbol{X}, \boldsymbol{y}, \hat{g}, \hat{\sigma}^2) = \sum_{k=1}^{K} \phi_{jk} \mathcal{N}(b_j; \mu_{jk}, s_{jk}^2), \tag{3.10}$$

for $j = 1, \ldots, p$, $k = 1, \ldots, K$, and ϕ_k refers to the posterior mixture assignment probability for each mixture component.

Unlike SuSiE, the Mr.ASH package implementation does not directly output the PIP calculations. However, it is very straightforward to obtain the PIPs for variants from the posterior calculation outputs. Using the definition of PIP in equation (3.2) and the posterior mixture assignment probability ϕ_k obtained by fitting the VEB model, we can estimate the PIP as:

$$\begin{aligned} \widehat{\text{PIP}}_j &= 1 - \Pr(\gamma_j = 1|\boldsymbol{X}, \boldsymbol{y}, g, \sigma^2) \\ &= 1 - \phi_{j1}, \end{aligned} \tag{3.11}$$

where we specify $\sigma_1^2 = 0$ when fitting the Mr.ASH model.

3.4. **PIP vs. Local False Sign Rate.**

One of the problems with using PIP directly as the likelihood identified variable is in the true model is that it is calculated using the local false discovery rate (*lfdr*) of each variable. *lfdr* only evaluates the probability of whether an effect is exactly zero [3], [8]:

$$lfdr_j := \Pr(b_j = 0|\boldsymbol{X}, \boldsymbol{y}). \tag{3.12}$$

From equation (3.2) and the definition in equation (3.12), we can see that PIPs are estimated as:

$$\text{PIP}_j = 1 - lfdr_j. \tag{3.13}$$

lfdr as a ranking measure is highly sensitive to model assumption. Assuming $b_j = 0$ for all j, if a variable that is not a true predictor has a small but non-zero effect, the model could assign that variable a small *lfdr* and therefore assigning that variable a large PIP. Since SuSiE assumes that true effects are sparse, it is capable of disregarding variables with effects closer to zero, and selecting variables with larger effects when we use a high PIP threshold. However, since Mr.ASH does not have an assumption on the sparsity of effects,

it is prone to assigning large PIPs to a large number of variables, which tend to increase the number of false discoveries.

Stephens (2017) proposed a more conservative metric, "local False Sign Rate (*lfsr*)", to estimate the probability of assigning a wrong sign to the effect j [8]:

$$lfsr_j := \min[\Pr(b_j \geq 0|\boldsymbol{X}, \boldsymbol{y}), \Pr(b_j \leq 0|\boldsymbol{X}, \boldsymbol{y})]. \tag{3.14}$$

We know previously that obtaining a small $lfdr_j$ implies we are confident that the effect j is nonzero. Similarly, obtaining a small $lfsr_j$ implies we are confident in the sign of effect j [8]. It follows that $lfsr_j \geq lfdr_j$, meaning *lfsr* is a more conservative measure of significance than *lfdr* [8].

Because of SuSiE's assumption of sparse effects, the *lfdr* approximation and *lfsr* approximation for SuSiE are similar. We can therefore continue to use PIP estimation as our ranking measure, as described in equation (3.4). For Mr.ASH, however, we use $1 - lfsr_j$ instead of PIP_j for estimation of inclusion probability for effect j. We can estimate *lfsr* in Mr.ASH by plugging the posterior distribution generated in equation (3.10) into definition of *lfsr* in equation (3.14):

$$\widehat{lfsr}_j = \min[\phi_{j1} + \sum_{k=2}^{K} \phi_{jk}\Phi(\frac{0 - \mu_{jk}}{s_{jk}}), \phi_{j1} + 1 - \sum_{k=2}^{K} \phi_{jk}\Phi(\frac{0 - \mu_{jk}}{s_{jk}}))]. \tag{3.15}$$

In short, we apply the same PIP cutoffs in SuSiE to $1 - lfsr$ values in Mr.ASH, and compare the power achieved by the two methods at different cutoffs.

3.5. **LASSO Initialization for Mr.ASH.**

One of the issues Kim et. al. discovered about Mr.ASH is that when it is applied to highly correlated variables, initializing $\bar{\boldsymbol{b}} = \text{E}(\boldsymbol{b})$ using the solution from LASSO after cross-validation can perform better than the null initialization, where $\bar{b}_j = 0$ for all j. LASSO is a penalized regression model where predictors $\hat{\beta}_j$ are selected by minimizing:

$$||y_i - \sum_j \beta_j x_{ij}||_2^2 + \lambda||\beta_j||_1. \tag{3.16}$$

The parameter $\hat{\lambda}$ is commonly chosen using cross-validation (CV), where one divides the data set into K folds, trains on all but the k-th part of the data, and validate the model using the k-th part, iterating over $k = 1, 2, \ldots, K$, and $\hat{\lambda}$ is the value that minimizes the cross-validation error [9]. As we addressed before, LASSO is not a good method for variable selection in high correlation settings because it selects the best subset k by choosing k largest coefficients and setting the rest to zero [9]. We know that linear regression models cannot tease apart coefficient assignment to variables when they are highly correlated. When two

predictors are highly correlated, LASSO will assign a large coefficient to one of them and not the other, therefore only one of them will be included in the model, when there is a possibility the other predictor is the true causal variable. However, the benefit of initializing Mr.ASH with LASSO is that Mr.ASH will be able to quickly obtain an initialized set of $\bar{b}_j$ coefficients, which Kim et. al. showed to improve prediction accuracy of Mr.ASH, particularly when predictors have a correlation $r^2 > 0.6$ [6]. In our experiment, we want to further investigate if LASSO initialization will similarly lead to an improvement in variable selection accuracy when predictors are highly correlated, or if the problem of arbitrarily choosing one of the highly correlated variables will persist in LASSO initialized Mr.ASH fitting.

4. Methods

The main purpose of our study is to compare the performance of different variable selection methods on predictors with varying degrees of correlation. For simplicity, we generated $\boldsymbol{X}$ matrices of different degrees of correlation through LD pruning. Linkage disequilibrium in genotype data refers to the non-random association of two alleles at different loci due to recombination, and is generally calculated as squared correlation r^2. To control degrees of correlation between the variables and test our variable selection methods, we perform LD pruning at different levels of r^2 so that the remaining variables would not have correlation exceeding the specified LD threshold. We subsequently fit SuSiE and Mr.ASH models and compare the power of their variable selection abilities using a variation of the precision-recall plot.

4.1. **Generating Genotype Matrix.**

Our $\boldsymbol{X}$ matrix of genotype data was taken from the GTEx Analysis Release V8 [2], with n-rows corresponding to individuals and p-columns corresponding to genetic variants, such as single nucleotide polymorphisms (SNPs). In our simulations, we only selected $n = 300$ genes for the majority of our analysis, each with genotypes from 838 individuals. For each gene, we first take a subset of 5000 SNPs and perform LD pruning: a correlation matrix is calculated for all SNPs in the subset; if the r^2 between a pair of SNPs exceeds the specified LD threshold, one of the SNPs is removed at random. After such LD pruning, the variables will have r^2 no larger than the LD threshold at which the data set is pruned.

As observed in Figure 1, there is a large difference in number of SNPs remaining across different samples when pruned at the same LD threshold. This would cause a problem because in cases where there are fewer SNPs to start out with, the variable selection method is more likely to pick the correct causal variant by chance alone. To make sure the pruned data sets are all the same size, we perform the simulations with a two-step process: first, we select a subset of genes greater than n, and pruned all samples with the lowest LD threshold (e.g. $r^2 = 0.4$). By plotting a histogram of the number of remaining SNPs (Figure 2), we can select

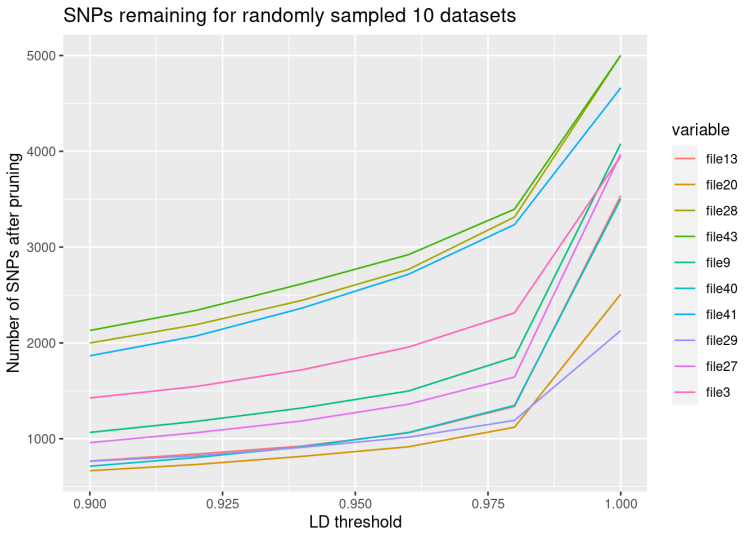

FIGURE 1. **Number of SNP remaining after LD pruning** the number of SNPs remaining after LD pruning at thresholds 0.9 to 1.0 for 10 randomly selected genes

a reasonable lowest SNP count such that there are more than $n = 300$ genes that satisfy the lowest SNP count after pruning. These genes that satisfy the minimum SNP count at the harshest pruning threshold will then be pruned at all r^2 thresholds $r^2 \in \{0.4, 0.5, 0.6, \ldots, 1.0\}$ and fitted with SuSiE and Mr.ASH. The case $r^2 = 1.0$ correspond to the scenario where some SNPs in the data set may be in perfect LD, so no LD pruning was performed on the data set. Since pruning at a higher LD threshold will always result in a larger number of SNPs remaining, we sample from the leftover SNPs randomly so that all pruned genotype matrices have the same number of SNPs. For example, in Figure 2, we observe that more than 300 genes had greater than 200 SNPs remaining, therefore we select 200 as the minimum SNP count for subsequent LD pruning at all levels of LD.

4.2. **Generate Gene Expression Vector.**

We simulate the gene expression data $\boldsymbol{y}$ for $n = 300$ genes. $\boldsymbol{y}$ is simulated under the multiple regression model as described in (3.1). To generate each $\boldsymbol{y}$ vector, we fix the number of true effects S and the proportion of variance in $\boldsymbol{y}$ explained by $\boldsymbol{X}$ by the same process detailed in Wang et. al. (2020) [10]:

(1) Sample the indices of the S effect variables, $\mathcal{S}$, uniformly at random from $\{1, \ldots, p\}$. These are the simulated "true effect variables".

(2) For each $j \in \mathcal{S}$, independently draw $b_j \sim N(0, 1)$, and for all $j \notin \mathcal{S}$, set $b_j = 0$.

(3) To achieve our specified PVE $= 0.2$, we solve for the σ^2 by solving the equation $\text{PVE} = \frac{\text{Var}(\boldsymbol{Xb})}{\sigma^2 + \text{Var}(\boldsymbol{Xb})} = 0.2$, where $\text{Var}(\cdot)$ denotes sample variance.

(4) For each sample $i = 1, \ldots, n$, we simulate $y_i \sim N(x_{i1}b_1 + x_{i2}b_2 + \cdots + x_{ip}b_p, \sigma^2)$

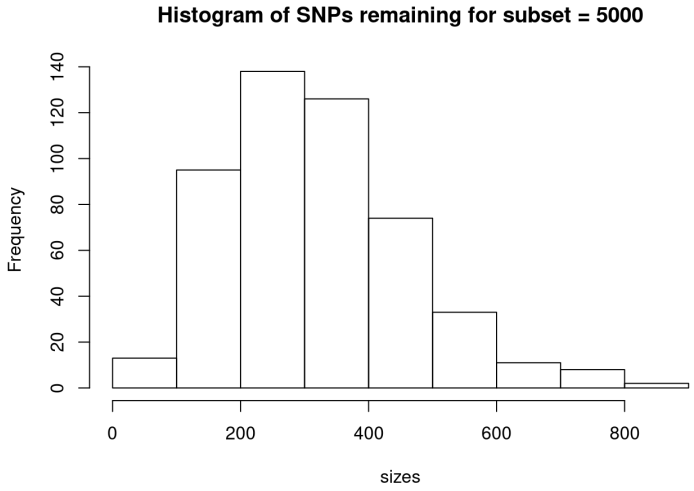

FIGURE 2. **Histogram of number of SNPs remaining** 500 genes were pruned at LD threshold $r^2 = 0.4$, 394 had more than 200 SNPs remaining pruning at the lowest LD threshold.

Gene expression simulation is performed on data sets at all LD thresholds $r^2 \in \{0.4, 0.5, \ldots, 0.9, 1.0\}$ for all numbers of true effect variables $S \in \{1, \ldots, 5\}$.

4.3. **Model Fitting.**

For our analysis with SuSiE, we specify the maximum number of single effects in each regression $L = 10$, and we estimate both the residual variance σ^2 and prior variance σ_0^2. Note here that the L specified is greater than the maximum number of effects simulated, which is $S = 5$. This is a convention adopted in the Wang et. al. (2020) and is not problematic. SuSiE distributes the extra effects broadly among many variables, therefore inflating the PIPs of many variables slightly but should not be significant enough to reach the PIP thresholds we are interested in (PIP ≥ 0.9) [10].

Mr.ASH model is fitted in two different ways: with and without LASSO cross-validated initialization. Kim et. al. (2020) specified that the solution obtained from the VEB model is dependent on initialization of $\boldsymbol{b}$, especially in scenarios where columns of $\mathbf{X}$ are highly correlated (correlation $r^2 > 0.6$) [6]. In our experiment with initialized Mr.ASH, we first obtain cross-validated LASSO estimates $\boldsymbol{b}$, and then initialize Mr.ASH using these LASSO estimates for the subsequent PIP calculation, as described in Kim. et. al.

5. RESULTS

5.1. **Varying Number of Effects and LD.**

We first want to see if the number of effect variables have a large influence on the performance of SuSiE and Mr.ASH. Since we wanted to test both methods on data sets with sparse effects, we limited the number

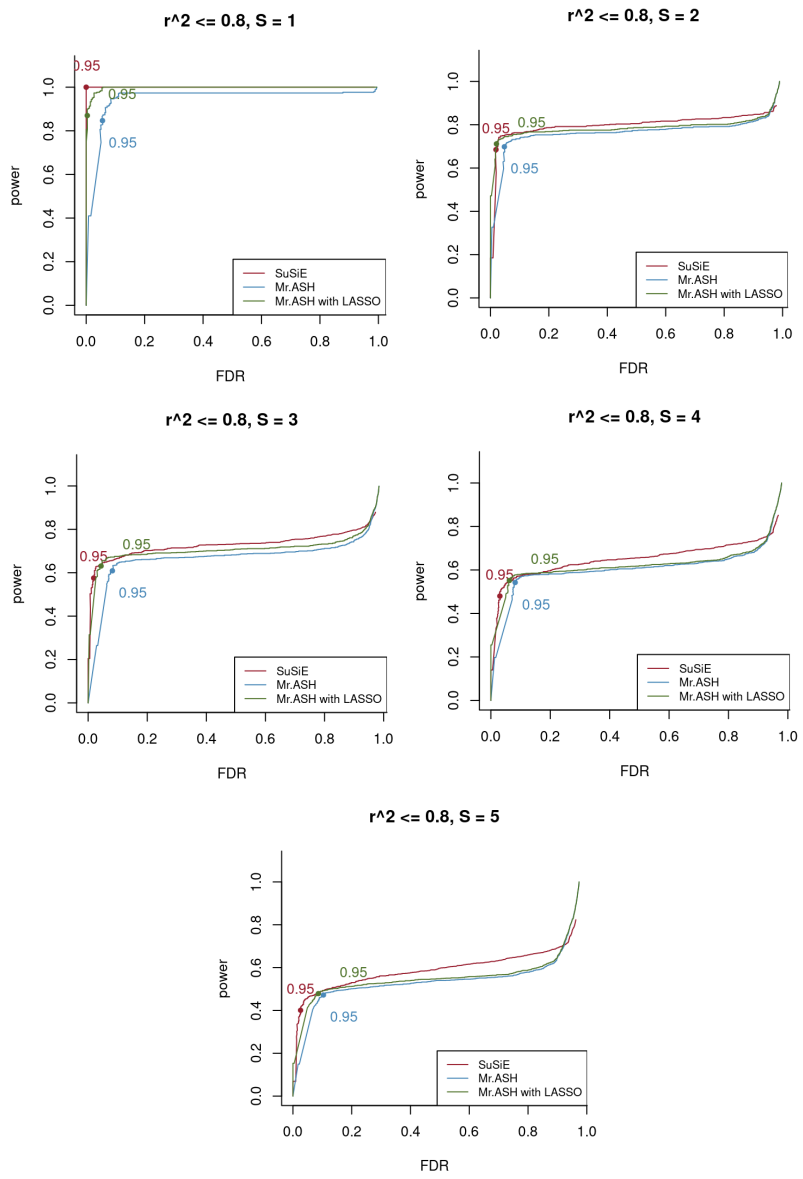

FIGURE 3. **Effect of Variable Number on SuSiE and Mr.ASH** Gene expression vector $\boldsymbol{y}$ is simulated with varying number of true effect variables $\mathbf{b}_j \neq 0$. SuSiE and Mr.ASH are used to fit data sets pruned at all LD levels. Only $r^2 = 0.8$ is displayed here, the rest of the figures are included in the appendix.

of true effects simulated to $S \leq 5$. Figure 3 shows that both SuSiE and Mr.ASH achieved very high power at low FDR for all number of effect variables used in our simulations, as both curves increased quickly in power at high FDR, and leveled out as FDR increases. Both SuSiE and Mr.ASH achieve lower maximum power when more effect variables are simulated in the data, but their relationship remains the same in all five settings. Figure 3 also demonstrates that SuSiE consistently outperforms Mr.ASH except at high FDR. Since SuSiE and Mr.ASH achieve consistent results for all numbers of effect variables, we can pool identifications

across different numbers of true variables in our subsequent analysis. When trials with different number of variables are pooled (Figure 4), we observe that Mr.ASH starts to have worse identifications when $r^2 > 0.7$, but are a lot closer to that of SuSiE at lower LD thresholds. At high LD, such as $r^2 = 0.8$, we can see that Mr.ASH identifications at cutoff PIP = 0.95 had higher FDR while achieving the same power as the respective PIP cutoff in SuSiE.

5.2. **Initialization of Mr.ASH.**

Kim et. al. (2020) pointed out that Mr.ASH is sensitive to initialization conditions when the correlation between variables are larger than $r^2 = 0.6$ [6]. Since we observe a similar trend where Mr.ASH's performance became worse than SuSiE at an LD threshold greater than 0.6, we hypothesized that it could be due to its sensitivity to initialization conditions. We therefore also initialize $\bar{\boldsymbol{b}}$ with LASSO cross validation (CV). Figure 4 shows that when no pruning took place, SuSiE achieved best variable selection performance among all three methods, reaching relatively high power at low FDR. From $r^2 = 0.7$ to $r^2 = 0.9$, Mr.ASH initialized with LASSO shows slight improvement compared to its uninitialized counterpart, achieving a performance almost as good as SuSiE. At $r^2 \leq 0.6$, there is almost no difference between the three methods.

In our discussion of the methods, we mentioned that the benefit of using SuSiE and Mr.ASH (as opposed to using methods like LASSO alone) is that we can easily obtain the PIP or $1 - lfsr$ for each variable selected, and can interpret variables assigned higher PIPs or$1 - lfsr$ values as variables with a larger probability of being true effects. We can see that a high PIP threshold, such as 0.95, is indeed located at the top left corner of the power vs. FDR curve, corresponding to a high power and low FDR, which is what we want our variable selection standard to achieve. This justifies the validity and convenience of using PIP or $1 - lfsr$ as a metric to rank and select variables.

5.3. **Comparison in non-sparse settings.**

In all of our experiments so far, Mr.ASH rarely appears to perform better than SuSiE except at high FDR. This result is not surprising, since our variables have high correlation and sparse effect, both of which are conditions SuSiE was developed to be suitable for. We proceed to compare SuSiE and Mr.ASH's variable selection ability on effects that are less sparse ($S = 20$), since SuSiE's single effect assumption might not be as suitable in that scenario.

Figure 5 demonstrates the variable selection ability of SuSiE and Mr.ASH when PVE = 0.2, as we have described in section 4.2 and used in all previous experiments. By increasing the number of effects from 1 through 5 to 20, we effectively decrease the average effect size, obtaining a new average effect size per variable of 0.01. It is not surprising to see that SuSiE and Mr.ASH both perform significantly worse here than in the more sparse scenario, when $S \leq 5$. Since each individual effect is decreased significantly, it is natural

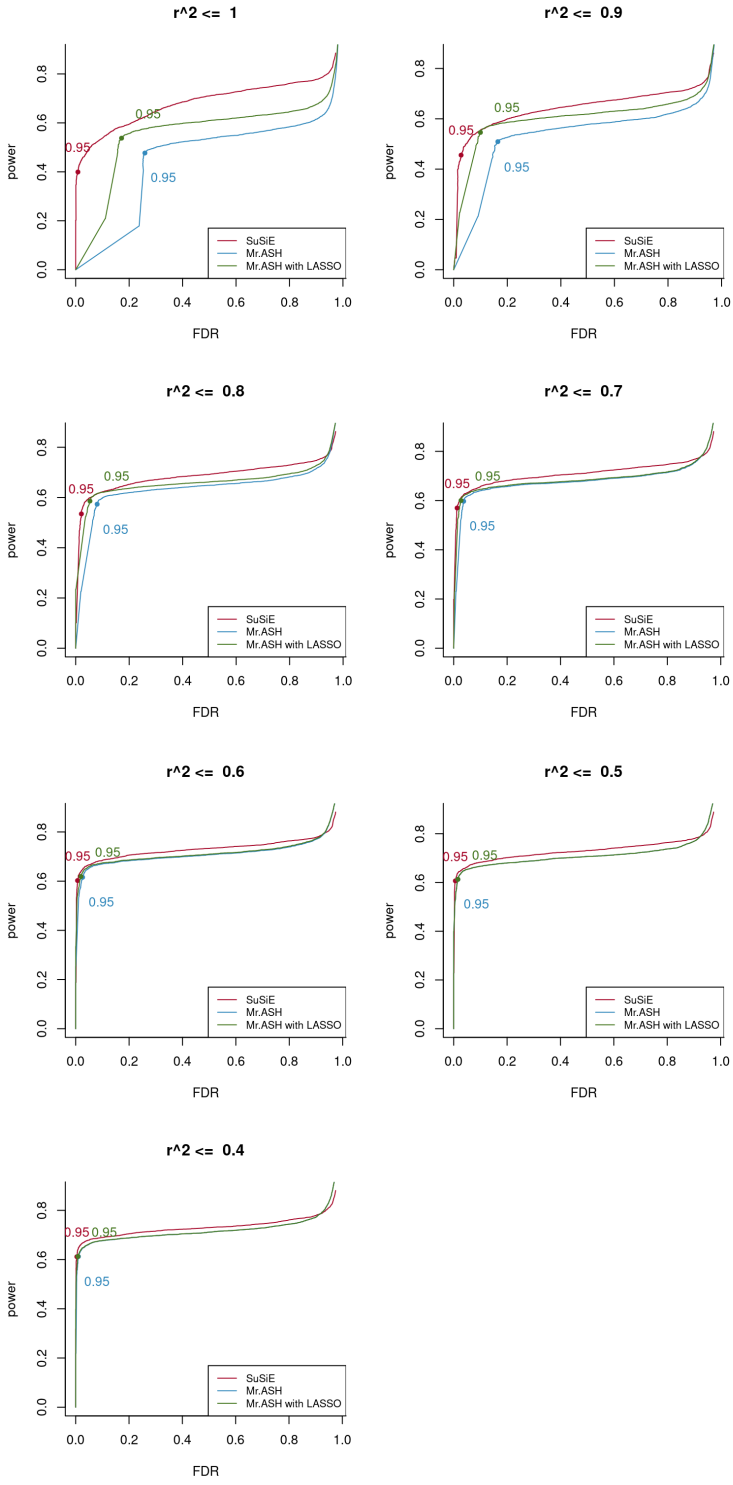

FIGURE 4. **Evaluation of posterior inclusion probabilities(PIPs)** power vs FDR for data sets of the specified LD. PIP thresholds of 0.95 are marked by circles for each method. Power is calculated as $\frac{\text{TP}}{\text{TP+FN}}$, and FDR is calculated as $\frac{\text{FP}}{\text{TP+FP}}$, where TP, FN, FP stands for count of True Positive identifications, False Negatives, and False Positives, respectively.

that regression methods have a harder time identifying these small true effects. Figure 5 shows that all three methods achieve low power at low FDR in the precision-recall curves in comparison to our observation in Figure 4. There also seems to be less difference between initialized and non-initialized Mr.ASH identifications for all levels of LD. Comparing Mr.ASH to SuSiE, Mr.ASH has marginally better power than SuSiE at low FDR (FDR < 0.4). SuSiE achieved better performance at higher FDR (FDR ≥ 0.4), but that is a region we are less interested in.

By holding PVE constant, we are essentially decreasing effect size and increase the effect density at the same time. It is ambiguous which factor actually caused the significant drop in power in both methods. Therefore, we simulate a non-sparse scenario where there are more true effects but effects have similar sizes as our simulations in the sparse cases. We increase the PVE 4 fold, corresponding to our 4-fold increase in number of true effects, so that the average size per effect is still held at 0.04. In Figure 6, we observe that other than the $r^2 = 1$ case, Mr.ASH outperformed SuSiE at all LD thresholds, and there is a larger difference between the initialized and uninitialized Mr.ASH methods at $r^2 \geq 0.7$. These observations support our hypothesis that Mr.ASH has better variable selection performance than SuSiE when signals are non-sparse, since Mr.ASH regression does not have any assumption on signal sparsity. At high LD threshold $r^2 = 1$ and very low LD threshold $r^2 = 0.4$, the difference in performance between SuSiE and Mr.ASH seems to be smaller than when data has moderate LD (r^2 threshold between 0.5 and 0.9). Additionally, initialization plays a bigger role in Mr.ASH's variable selection power when effects are non-sparse and large compared to when effects are sparse or when effects are small. In particular, when LD threshold is between $r^2 = 0.7$ to $r^2 = 0.9$, initialized Mr.ASH achieved significantly higher power compared to both SuSiE and uninitialized Mr.ASH at low FDR. This further supports Kim et. al.'s statement and our previous observation of the importance of LASSO initialization of Mr.ASH at moderate LD.

As we discussed in section 3.2, one of the important parameters for SuSiE is the number of effects L used in the model, and that we should set this parameter to be greater or equal to the number of true effects. In Figure 6, we fit the SuSiE model with $L = 20$, which is exactly the number of true effects simulated. However, we also test the scenario where we specify an L greater than the number of true effects. In Figure A.2, we fit the SuSiE model with $L = 40$. An L larger than the number of true effects does not improve the SuSiE model fit; there is in fact very little difference between the results from the two model fits. This result is expected, since Wang et. al. (2020) discussed that SuSiE effectively estimates the number of effects and is therefore robust to overstating L [10].

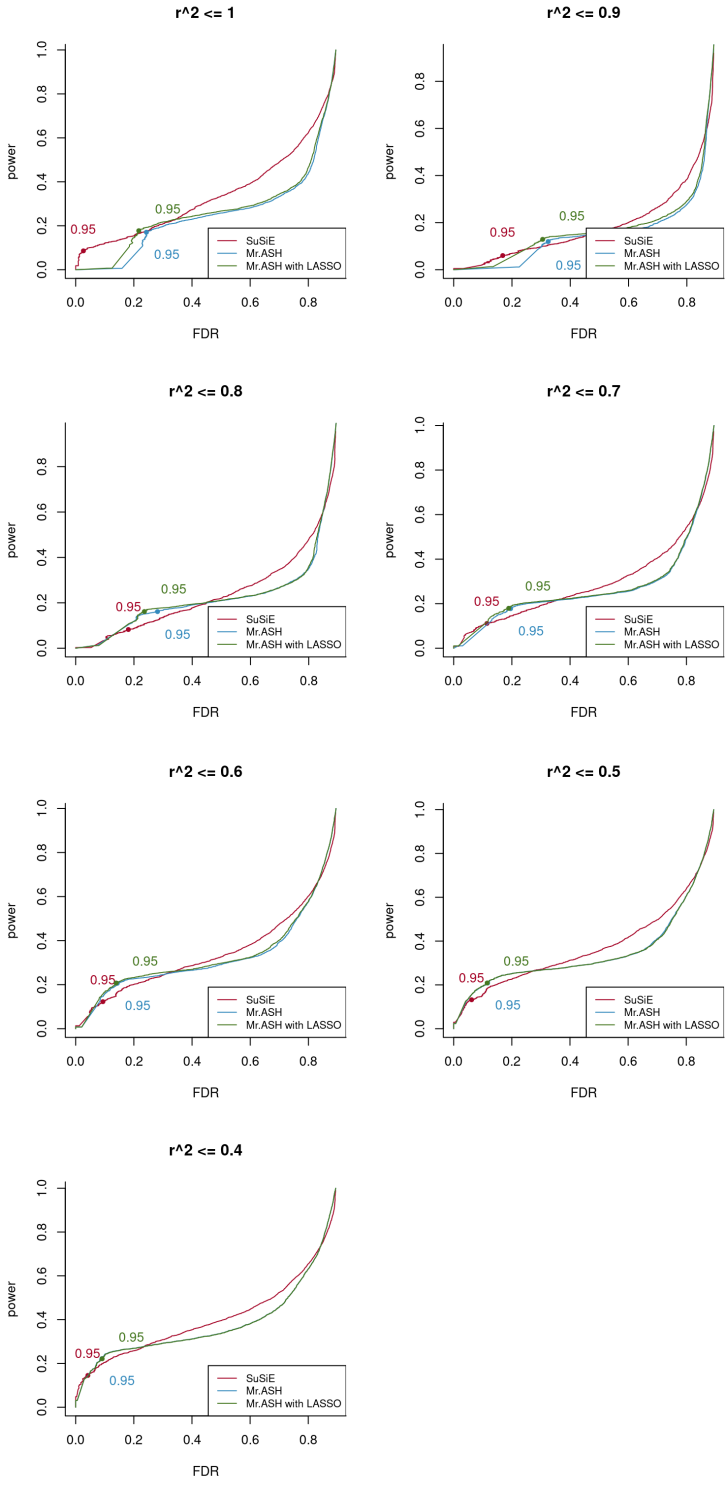

FIGURE 5. **Evaluation of posterior inclusion probabilities(PIPs) when effects are small and non-sparse** power vs FDR for data sets with number of true effects $S = 20$, at different levels of LD, PVE = 0.2. PIP thresholds of 0.95 are marked by filled circles for each method.

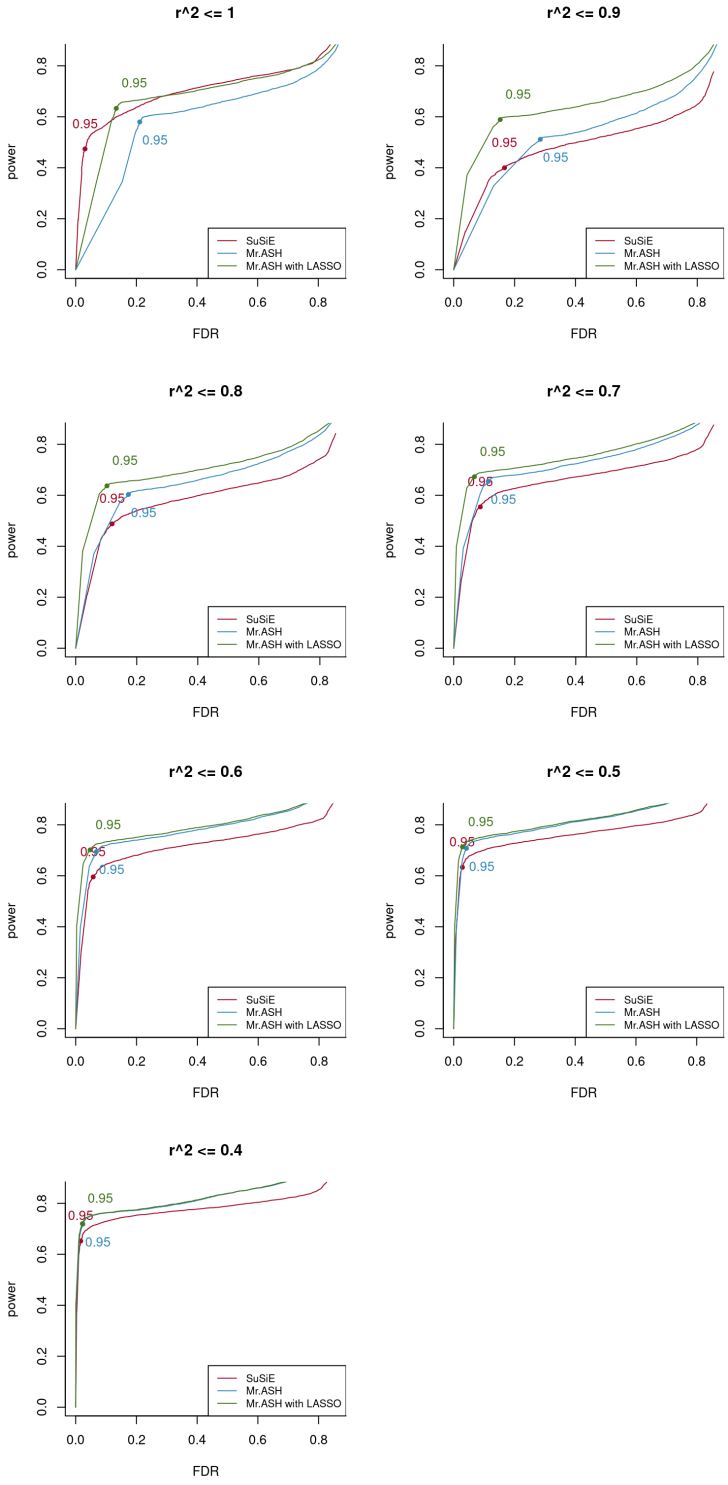

FIGURE 6. **Evaluation of posterior inclusion probabilities(PIPs) when effects are non-sparse** power vs FDR for data sets with number of true effects $S = 20$ and PVE $= 0.8$. SuSiE is fitted with parameter $L = 20$. PIP thresholds of 0.95 are marked by filled circles for each method.

6. Discussion

The central question we are addressing is how to determine the optimal variable selection method for data sets with predictors of varying degrees of correlation. Specifically, we hypothesize that SuSiE will outperform Mr.ASH on data sets where true effects are sparse and variants are in high correlation. Our results confirm that in a sparse setting, SuSiE not only outperforms Mr.ASH when variables are highly correlated, but also demonstrated high variable selection power at lower correlations.

To simulate the varied levels of correlation between genetic variants, we trim the data using LD pruning so that the remaining SNPs are equal in number and their correlation do not exceed the LD threshold. In the sparse setting, change in number of true effects results in very little difference between SuSiE and Mr.ASH identifications (Figure A.1). SuSiE, uninitialized Mr.ASH, and LASSO initialized Mr.ASH appear all have worse performance when LD threshold is $r^2 = 1$ or $r^2 = 0.9$ compared to $r^2 < 0.9$ cases. This is reasonable because variables that have $r^2 > 0.9$ can essentially be regarded as perfectly correlated, and no method can reliably distinguish perfectly correlated variables. We mention briefly in section 3.2 that SuSiE is capable of placing these highly correlated variables into credible sets with high purity ($r^2 \geq 0.95$). While that is not the focus of this study, it is an appropriate method to identify and include variables that are true effects or in high correlation to true effects, and avoid the problem of arbitrarily leaving out true effects like LASSO does.

For the purpose of comparing the power of variable selection methods, it is therefore more informative to investigate their behavior on data sets with variables in moderate LD, particularly from $r^2 = 0.7$ to $r^2 = 0.9$ as observed in our results. In this range of LD, SuSiE had a marginally better performance than Mr.ASH even when Mr.ASH is initialized. This agrees with our initial expectation that SuSiE would perform better when variants are in high LD (Figure 4). Our observations also confirmed our hypothesis that SuSiE would perform no worse than Mr.ASH at moderate to low LD: when $r^2 \leq 0.6$, there is very little difference between the performance of the three methods.

Further, our results demonstrate that PIP or 1-*lfsr* is a good ranking measure for variable selection. For all of our experiments, the PIP $= 0.95$ or $1 - lfsr = 0.95$ cutoffs are located very close to the top left of the power-FDR curves, corresponding to a high power and low FDR in variable selection in both methods. However, one limitation of using PIP as our variable selection cutoff is that not all variable selection methods produce a ranking measurement directly comparable to PIP (e.g. LASSO). If we would like to extend the comparison to methods such as LASSO, we need to find a generalized method that can rank variables by the probability they are in the true model, similar to PIP and $1 - lfsr$ values for SuSiE and Mr.ASH. This

will allow us to extend our comparison to more commonly applied variable selection methods and evaluate their variable selection ability on data sets of varying features.

Lastly, we briefly explore the performance of SuSiE and Mr.ASH on non-sparse data sets, such as when 20 out of 200 SNPs considered are simulated to be true effect variables. When effects are small, SuSiE and Mr.ASH both demonstrated significantly worse performance compared to the previous scenario where effects are larger. If effect sizes are kept the same as the experiments in sparse scenario, Mr.ASH achieves higher power than SuSiE at low FDR when effects are less sparse. This suggests that sparsity of the effects might also be an important factor in determining which method is more suitable for a given data set, especially when variables are in moderate LD (LD threshold $0.7 \leq r^2 \leq 0.9$), where we tend to observe the greatest difference between Mr.ASH, LASSO initialized Mr.ASH, and SuSiE.

In conclusion, we found that both methods demonstrate superior performance under assumptions for which they are designed. SuSiE is designed specifically for data sets with sparse effects and highly correlated variables, therefore when true effects are sparse, SuSiE consistently outperforms Mr.ASH. Mr.ASH is designed to be a flexible multiple regression method that can be applied to data sets of all features, therefore it demonstrates power comparable to SuSiE in all but the highest r^2 thresholds. However, in scenarios where true effects are less sparse, Mr.ASH shows better performance than SuSiE, as it has no assumption on effect sparsity. One possible future experiment is to perform a systematic comparison of SuSiE and Mr.ASH's variable selection performance with data of varying sparsity. This may allow us to find a concrete threshold of "sparsity" necessary to satisfy SuSiE's assumption and achieve equal or superior performance than Mr.ASH.

References

[1] RV Broekema, OB Bakker, and IH Jonkers. "A practical view of fine-mapping and gene prioritization in the post-genome-wide association era". In: *Open biology* 10.1 (2020), p. 190221.

[2] GTEx Consortium et al. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (2017), p. 204.

[3] Bradley Efron. "Microarrays, empirical Bayes and the two-groups model". In: *Statistical science* (2008), pp. 1–22.

[4] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[5] Georg Heinze, Christine Wallisch, and Daniela Dunkler. "Variable selection–a review and recommendations for the practicing statistician". In: *Biometrical journal* 60.3 (2018), pp. 431–449.

[6] Youngseok Kim et al. "A Flexible Empirical Bayes Approach to Multiple Linear Regression and Connections with Penalized Regression". Unpublished Manuscript. 2020.

[7] Younseok Kim. "Bayesian Shrinkage Methods for High-dimensional Regression". PhD thesis. University of Chicago, 2020.

[8] Matthew Stephens. "False discovery rates: a new deal". In: *Biostatistics* 18.2 (2017), pp. 275–294.

[9] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[10] Gao Wang et al. "A simple new approach to variable selection in regression, with application to genetic fine mapping". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5 (2020), pp. 1273–1300.

[11] Wei Wang. "Applications of Adaptive Shrinkage in Multiple Statistical Problems". PhD thesis. University of Chicago, 2017.

[12] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

Appendix A. Figure 3 Cont'd: Precision-Recall Curves

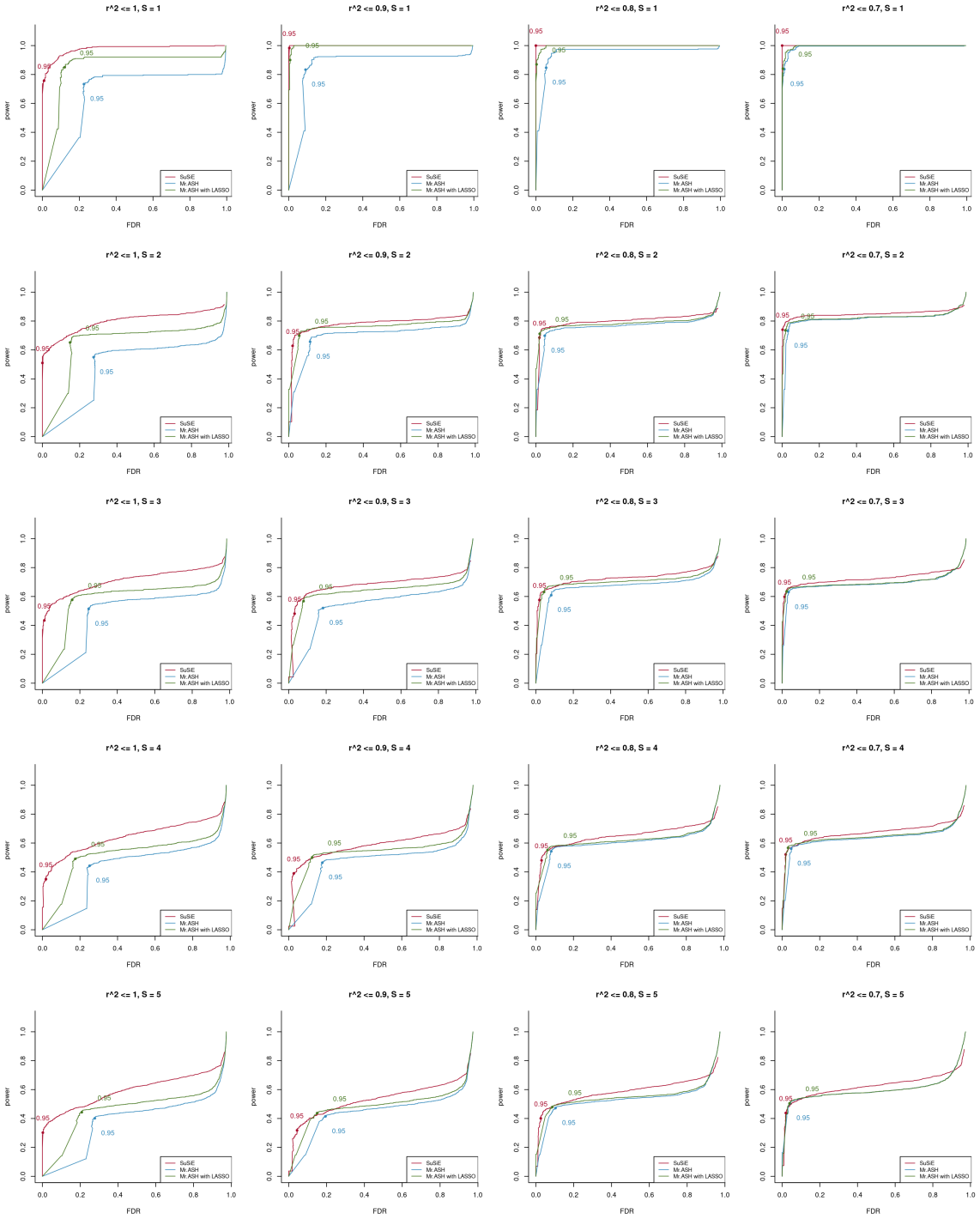

Figure A.1. **Effect of Number of True Effect Variables on SuSiE and Mr.ASH** Gene expression vector $\boldsymbol{y}$ is simulated with S number of true effect variables $\mathbf{b}_j \neq 0$. SuSiE and Mr.ASH are used to fit data sets pruned at all LD levels.

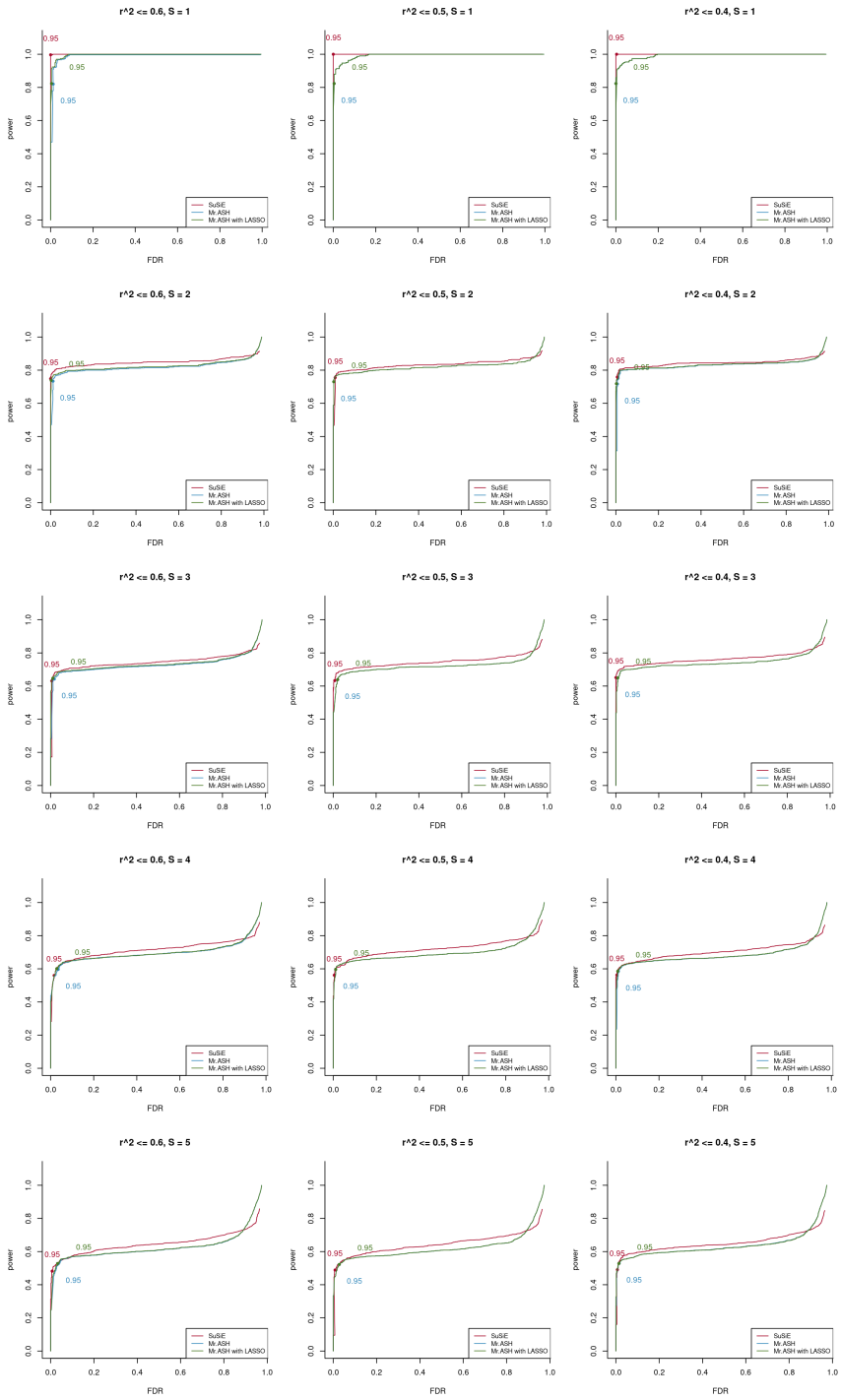

FIGURE A.1. **Effect of Number of True Effect Variables on SuSiE and Mr.ASH (cont'd)**

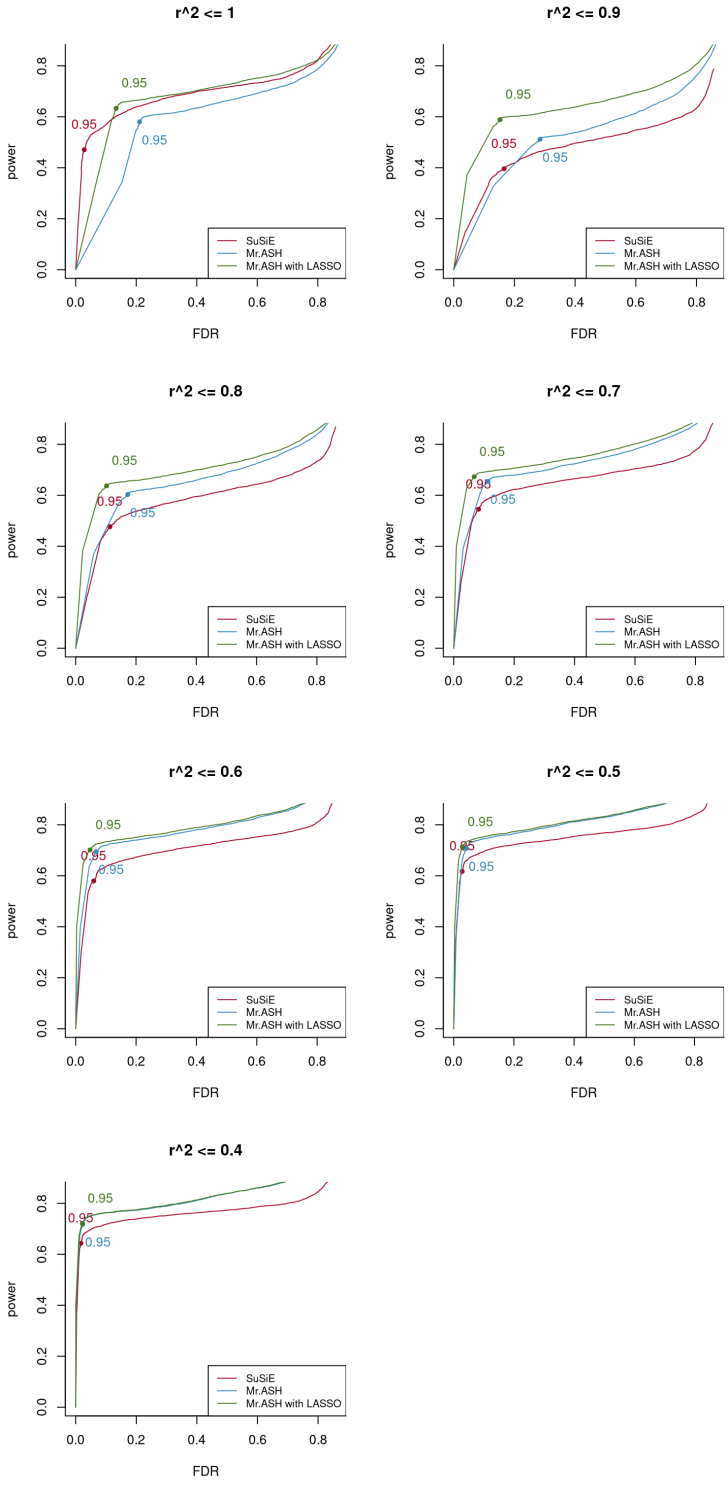

FIGURE A.2. **Evaluation of posterior inclusion probabilities(PIPs) when effects are non-sparse** power vs FDR for data sets with number of true effects $S = 20$ and PVE $= 0.8$. SuSiE is fitted with parameter $L = 40$ PIP thresholds of 0.95 are marked by filled circles for each method.