Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–32

Spatial Knockoff Bayesian Variable Selection in Genome-Wide Association Studies

Justin J. Van Ee*, Diana Gamba†, Jesse R. Lasky†,
Megan L. Vahsen‡, and Mevin B. Hooten§

Abstract. A primary objective in many fields is identifying the relevant predictors of a response from a large collection of variables. In genome-wide association studies, for example, variable selection methods have been adapted for identifying single nucleotide polymorphisms (SNPs) linked to phenotypic variation. The mechanisms of inheritance, evolution, recombination, and gene expression produce predictable structures in genotype, phenotype, and their associations. Guided by these mechanisms, we develop a scalable Bayesian variable selection regression model that unifies several recent advances in variable selection. Using a restricted regression approach, we demonstrate a computationally stable method for estimating SNP inclusion probabilities simultaneously with sets of basis functions that account for population structure. Motivated by the spatial arrangement of genes and their regulators, we also accommodate the non-uniform distribution among evolutionarily relevant SNPs by modeling their inclusion probabilities jointly with a Markov random field. We modify our Bayesian variable selection regression model to control the false discovery rate using hidden Markov knockoff variables that account for linkage disequilibrium and population structure in genomic data. In a simulation study, we demonstrate that our spatial Bayesian variable selection regression model controls the false discovery rate and increases power when the relevant variables are clustered. We conclude with a genome-wide association study of flowering time, a polygenic trait, measured across globally distributed accessions of *Arabidopsis thaliana* and find the discoveries of our method concentrate near described flowering time genes.

Keywords: Markov random field, reduced rank Gaussian process, hidden Markov model, Genome-wide association studies, restricted regression, *Arabidopsis thaliana*.

1 Introduction

High-dimensional variable selection has emerged as one of the prevailing statistical challenges in the big data revolution (Saeys et al., 2007). In large scale applications, identifying all the relevant variables is infeasible, and a greater emphasis is given to controlling the proportion of selected variables that are spuriously associated with the response (Benjamini and Hochberg, 1995; Brzyski et al., 2017; Sesia et al., 2021). A confounding factor in many variable selection problems is that measurements of the response are

*Department of Statistics, Colorado State University, Fort Collins, CO, justin.vanee@gmail.com
†Department of Biology, Pennsylvania State University, University Park, PA
‡Department of Wildland Resources and the Ecology Center, Utah State University, Logan, UT
§Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX

 https://doi.org/10.1214/25-BA1556

often correlated as a result of latent structure in the sample (Price et al., 2006, 2010). Lack of independence among sample units decreases power and can increase the number of spurious associations (Sul et al., 2018). The relevant variables may also be structured and quantifying their importance independently can forfeit gains in power (Vannucci et al., 2010; Li and Zhang, 2010).

In genome-wide association studies (GWAS), variable selection methods have been adapted for identifying single nucleotide polymorphisms (SNPs) associated with phenotypic variation (Lu et al., 2021). Variable selection in GWAS is generally the first step in a larger pipeline for identification of causal variants (Sesia et al., 2021). Follow up analyses involve linking the selected SNPs to genes and regulatory pathways (Wang et al., 2010; Brodie et al., 2016; Schaid et al., 2018; Wang et al., 2020), although some methods have been proposed for doing both steps simultaneously (Zhang et al., 2014; Lee et al., 2023; He et al., 2024). Genes or regulatory regions linked to the selected SNPs are prioritized for genome editing to understand and validate the functional role of each SNP (Spisák et al., 2015). These subsequent analyses are costly, and GWAS provides a statistically robust framework for SNP prioritization. Following previous work (Guan and Stephens, 2011; Sesia et al., 2019, 2021), we adopt the terminology 'relevant' to describe the true discoveries of GWAS methods to emphasize some discoveries, though not necessarily spurious, may not be linked to phenotype functionally.

The vast majority of existing GWAS methods rely on marginal approaches where each SNP is tested independently for its association with the phenotype. One of several methods is then used to transform the marginal test statistics and control the number of false discoveries or positives (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003; Brzyski et al., 2017; Wang and Ramdas, 2022). Assuming the sample of genotypes consists of unrelated individuals that share the same population background (i.e., descended from a common distant relative), the preceding approach is valid and in many cases will yield high power (Kang et al., 2010). In practice, individuals used in many current GWAS analyses are related or drawn from different ancestries, and relevant SNPs are often confounded with neutral genetic variation that does not influence phenotype but mimics the genetic correlations of the relevant SNPs. To attenuate identification of these spurious associations, contemporary GWAS methods also include genotype random effects to account for population structure and familial relatedness in a mixed model for SNP effects (Yu et al., 2006; Kang et al., 2010; Zhou and Stephens, 2012; Sul et al., 2018).

While marginal testing approaches have been instrumental for discovering thousands of genotype-phenotype associations (Donnelly, 2008), they have limitations. Many traits are polygenic and likely associated with large collections of mutations (Risch and Merikangas, 1996). The individual effect of each SNP within collections is generally small (Barton et al., 2017), and testing SNPs independently can overlook collections of small effects or SNPs that are marginally uncorrelated with the phenotype (Visscher et al., 2012). Marginal approaches also ignore the correlation structure among SNPs which complicates inference regarding the discoveries. Many of the selected SNPs are likely not relevant themselves but rather highly correlated with a causal variant Buzdugan et al. (2016). The selected SNPs, although individually strong predictors of phenotype,

may collectively have weak predictive performance because of high mutual correlations (Guan and Stephens, 2011).

A number of GWAS methods for simultaneously analyzing multiple SNPs have been proposed (Guan and Stephens, 2011; Li et al., 2011; Lu et al., 2015; Buzdugan et al., 2016; Sesia et al., 2019, 2020; Gu and Yin, 2021; Sesia et al., 2021). These methods estimate genotype-phenotype associations simultaneously for all SNPs and hence quantify the conditional relevance of each SNP. Modeling the SNPs jointly reduces the overall residual variation and enables identification of relevant SNPs that are marginally uncorrelated with phenotype, both of which can improve power for polygenic traits (Frommlet et al., 2012). Estimating the conditional relevance of each SNP also provides more evidence that the selected SNPs have causal effects on phenotype (Buzdugan et al., 2016). In contrast to the single SNP approaches previously described, most multiple SNP approaches do not use genotype random effects in a mixed model, but rather account for population structure or familial relatedness by preprocessing the phenotype. Preprocessing steps include regressing the phenotype of interest against principal components that describe population structure (Price et al., 2006, 2010), or clustering genotypes into groups of unrelated individuals with homogeneous ancestry for testing within subgroups (Pritchard et al., 2000). While these preprocessing steps have been shown to reduce spurious associations and improve power, a more cohesive and theoretically justifiable approach would simultaneously estimate SNP effects along with measures of population structure and familial relatedness.

An advantage of the Bayesian variable selection regression (BVSR) approach is the flexibility to incorporate prior information into the selection of variables (Fridley, 2009; Stephens and Balding, 2009). Such models have gained particular attention in biology where the inclusion indicators of covariates were assumed to have latent structure and have been modeled as Markov random fields (Li and Zhang, 2010; Vannucci et al., 2010). Stingo et al. (2011) developed a BVSR model that incorporated the relationship between genes and their membership to biochemical pathways to improve understanding of their expression levels on a phenotype of interest. Zhang et al. (2014) extended this framework to GWAS with a hierarchical model that use Bayesian variable selection for selection of genes with additional layer of selection for gene linked SNPs. Li and Zhang (2010) proposed modeling inclusion indicators of SNPs in GWAS with an Ising prior based on the linear arrangement of SNPs within the genome. Analyses of simulated data from these studies showed that modeling the selection indicators jointly with a Markov random field improved power and predictive performance relative to models that did not account for latent spatial structure among the relevant predictors. These preliminary studies along with functional genomic evidence that suggests relevant SNPs cluster in gene regulatory hotspots (Stern and Orgogozo, 2009) motivate the aggregation of spatial signals in GWAS to increase power.

We develop a BVSR model that uses knockoff variables (Barber and Candès, 2015) to achieve false discovery rate (FDR) control in finite samples. Candès et al. (2018) first proposed the knockoff variables in the context of BVSR, but extensions and applications to real datasets are currently limited (Gu and Yin, 2021). Guan and Stephens (2011) developed a computationally efficient BVSR model for high-dimensional variable selection that produced better power and predictive performance compared with

standard marginal variable approaches and least absolute shrinkage and selection operator (LASSO) regression. We extend the BVSR model proposed by Guan and Stephens (2011) with three notable additions. First, we implement a restricted regression framework for estimating SNP importance while accounting for population structure in the sampled individuals. Second, we improve power to detect the relevant variants by modeling the SNP inclusion probabilities jointly with a Markov random field. Third, we modify existing BVSR approaches to control the FDR using knockoff variables tailored for GWAS. We integrate each of these extensions into one unified model that has comparable computational efficiency to its precursors. Our approach, like other GWAS methods, is phenomenological in that we use generalized linear modeling to detect linear genotype-phenotype associations that are undoubtedly non-linear and interactive (Zuk et al., 2012). Nonetheless, each of our extensions is guided by a mechanistic understanding of the evolution of genotype and phenotype that can improve our ability to detect casual variants with GWAS.

Across several simulation studies, we find our spatial BVSR model improves power to detect relevant variants relative to models that do not incorporate spatial dependence in the variable selection procedure. To assess the performance of our spatial BVSR model in a realistic setting, we analyzed the genetic factors influencing flowering time in 1058 wild accessions of *Arabidopsis thaliana*. Because of range expansions, bottlenecks, and multiple reintroductions to its invaded ranges, wild populations of *Arabidopsis thaliana* have complex population structure with high levels of admixture (Alonso-Blanco et al., 2016; Shirsekar et al., 2021) that make variable selection challenging. We focused on flowering time for our analysis because this trait has a polygenic architecture (Zan and Carlborg, 2018) with 282 described candidate genes (Brachi et al., 2010). Flowering time plays a central role in ecological adaptation and has been extensively studied for *Arabidopsis thaliana* (Shindo et al., 2005; Brachi et al., 2010; Alonso-Blanco et al., 2016; Zan and Carlborg, 2018). Traditional single SNP analyses of *Arabidopsis thaliana* flowering time have shown limited success, only revealing a few significant associations (Shindo et al., 2005; Brachi et al., 2010; Alonso-Blanco et al., 2016). We motivate our method in the context of GWAS for *Arabidopsis thaliana*, but also highlight the broader applications of our method where appropriate.

2 Methods

2.1 Bayesian Variable Selection Regression

Consider the generalized linear model (GLM)

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \phi], \tag{1}$$

$$f(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{u}, \tag{2}$$

where we use the bracket notation, $[\cdot]$, to denote probability distributions (Gelfand and Smith, 1990), $\boldsymbol{y}$ is a vector of phenotypes observed across n individuals with mean $\boldsymbol{\mu}$, ϕ is a set of additional parameters, possibly empty (e.g., Poisson regression), in the data model, $\boldsymbol{X}$ is a $n \times p$ matrix of covariates including an intercept, $\boldsymbol{G}$ is a $n_g \times n_u$ genotype

matrix containing the number of alleles each genotype has at a particular locus for the n_u SNPs in the sample of n_g genotypes, and $\boldsymbol{Z}$ is a matrix of 1s and 0s that links the n_g genotypes in the sample to the n observed phenotypes. As parameterized, the GLM assumes an additive effect at each loci, but we could also test for dominant and recessive effects at each loci if we added another n_u covariates (Li et al., 2011). Interactive effects could also be included but obtaining posterior inference is infeasible for comtemporary GWAS datasets (Zuk et al., 2012).

We assumed $n_u >> n$ and that only a fraction of the n_u sequenced SNPs have a non-zero effect on phenotype. Furthermore, we assumed effect sizes of relevant SNPs are small, with each relevant SNP explaining less than one percent of the total variation in $\boldsymbol{y}$ (Visscher et al., 2012; Barton et al., 2017). We sought to identify the relevant SNPs while controlling the number of SNPs we falsely conclude are important.

We adopted a BVSR approach and specified spike-and-slab priors (Mitchell and Beauchamp, 1988) for the SNP effects,

$$u_j \sim \begin{cases} \mathcal{N}(0, \sigma_a^2), & \text{for } \nu_j = 1 \\ 0, & \text{for } \nu_j = 0 \end{cases}, \tag{3}$$

$$\nu_j \sim \text{Bernoulli}(\pi_j). \tag{4}$$

The indicator variables, ν_j, act as "switches" for adding and removing SNPs from the model with $\nu_j = 0$ indicating that the phenotype $\boldsymbol{y}$ is independent of SNP j conditional on the other SNPs (i.e., $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{g}_j | \boldsymbol{G}_{-j}$, where $\boldsymbol{g}_j$ denotes the jth column of the genotype matrix and $\boldsymbol{G}_{-j}$ is the genotype matrix with the jth column removed). The posterior means of the ν_j are called posterior inclusion probabilities and help assess whether a SNP is a relevant predictor of phenotype. The BVSR approach is appealing because in GWAS many SNPs are assumed to be in non-coding and non-regulatory genomic regions. Conditional on the inclusion of the relevant variants, these SNPs are not associated with phenotype and setting $u_j = 0$ is theoretically justifiable (Fridley, 2009). Removing these irrelevant SNPs from the model both reduces runtime (Lu et al., 2015) and improves precision for the relevant SNP effects (Guan and Stephens, 2011).

Guan and Stephens (2011) introduced a flexible prior for the variance of the SNP effects, σ_a^2, that applies greater shrinkage for more complex models with many non-zero SNP effects. We let s_j^2 represent the sample variance of SNP j in allelic state and $h \sim \mathcal{U}(0, 1)$. Guan and Stephens (2011) induced a conditional prior distribution for σ_a^2, $[\sigma_a^2 | \boldsymbol{\nu}, \boldsymbol{s}^2, h]$, by defining

$$\sigma_a^2(\boldsymbol{\nu}, \boldsymbol{s}^2, h) = \frac{h}{1-h} \frac{1}{\sum_{j:\nu_j=1} s_j^2}. \tag{5}$$

Regardless of how many SNPs are estimated as non-zero, the adaptive prior holds the proportion of variance explained by all SNPs constant (Guan and Stephens, 2011). The prior is also heavy-tailed, an attractive property for minimally informative variance parameters in hierarchical models (Gelman, 2006), with density proportional to $f(h) = \frac{1}{(1+h)^2}$. We adopted the prior specification of Guan and Stephens (2011) for the variance of the non-zero SNP effects, σ_a^2.

2.2 Accounting for Population Structure

An implicit assumption of equation (2) is that all n_g genotypes constitute a random draw of unrelated individuals from the same population background (Kang et al., 2010). When the genotyped individuals are related or drawn from multiple ancestries, the observed phenotypes are no longer independent, and the BVSR model may select many SNPs that are spuriously associated with $\boldsymbol{y}$ (Sul et al., 2018). In single SNP methods, individual SNP effects are estimated jointly with a set of n_g genotype random effects that correct for population structure (Kang et al., 2010; Zhou and Stephens, 2012; Sul et al., 2018) and familial relatedness (Yu et al., 2006). For the multiple SNP methods, a common approach is to first decorrelate the observations by regressing $\boldsymbol{y}$ against a set of basis functions that describe population structure (Guan and Stephens, 2011; Li et al., 2011; Lu et al., 2015; Sesia et al., 2019). The residuals from the regression then replace $\boldsymbol{y}$ in equation (1).

A limitation of the step-wise procedure used for multiple SNP methods is that it does not propagate the uncertainty associated with estimating the basis function coefficients into the estimates of the other model parameters. We simultaneously estimate the environmental, SNP, and population structure effects by extending the original GLM, equation (2), to the following generalized linear mixed model (GLMM),

$$f(\boldsymbol{\mu}) = \boldsymbol{R\theta} + \boldsymbol{X\beta} + \boldsymbol{ZGu}, \tag{6}$$

where $\boldsymbol{R}$ is a matrix of n_R basis functions that provide a low-dimensional representation of population structure. Low-dimensional representations can be obtained from a spectral decomposition of either the genotype or kinship matrix (Price et al., 2006, 2010). The choice of basis function type and number will depend on the ancestries and relatedness of individuals in the sample. To clarify the connection between our reduced rank approach and the typical mixed model approach, we specified $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_\theta \boldsymbol{I})$. If we assume the phenotype is normally distributed with homoscedastic variance σ^2_e, note that integrating $\boldsymbol{\theta}$ out of the model results in

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{ZGu}, \sigma^2_e \boldsymbol{I} + \sigma^2_\theta \boldsymbol{R'R}), \tag{7}$$

which is essentially the mixed model of Kang et al. (2010) with genetic variance σ^2_θ and low-dimensional representation of the kinship matrix $\boldsymbol{R'R}$. Note that for samples that also have familial relatedness, we could include another effect to account for known pedigree information as in the model of Yu et al. (2006).

The coefficients $\boldsymbol{\theta}$ and $\boldsymbol{u}$ are confounded in equation (6) because of overlap in the column spaces of $\boldsymbol{R}$ and $\boldsymbol{ZG}$. Variation in the phenotype can be explained by both SNP effects as well as the population structure of individuals. To alleviate confounding, we adopted a restricted regression approach (Reich et al., 2006) and reparameterized the model as

$$\boldsymbol{\mu} = \boldsymbol{R\theta} + \boldsymbol{X\beta} + \boldsymbol{Ku}, \tag{8}$$

where $\boldsymbol{K} = (\boldsymbol{I} - \boldsymbol{P_R})\boldsymbol{ZG}$ and $\boldsymbol{P_R} = \boldsymbol{R}(\boldsymbol{R'R})^{-1}\boldsymbol{R'}$ is the projection matrix onto the column space of $\boldsymbol{R}$. The restricted model, equation (8), gives priority to $\boldsymbol{\theta}$ to explain

all the contested sources of variation in the response (Reich et al., 2006). Conceptually, the restricted model is similar to the step-wise procedures in that the SNP effects are estimated from the residual variation in $\boldsymbol{y}$ remaining from regressions that correct for population structure. Because SNPs are measured for an effect on the phenotype that is not explained by population structure, both procedures lack power to detect the relevant variants that are confounded with the ancestries or relatedness of individuals. The benefit of the restricted model is that all parameters are estimated simultaneously such that their uncertainties are accounted for in the model. For observational data, the environmental and genetic effects may also be confounded, and restricted regression could be used for $\boldsymbol{\beta}$ (i.e., $\boldsymbol{K} = (\boldsymbol{I} - \boldsymbol{P_R})(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{ZG}$).

2.3 Spatially Structured Inclusion Probabilities

Traditional approaches to BVSR have treated the inclusion probabilities as a fixed hyperparameter, $\pi_j = \pi$ for $j = 1, \ldots, n_u$ (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). While fixing π_j may be appropriate for some analyses, in many cases, the sparsity may not be known even to the correct order of magnitude. Guan and Stephens (2011) specified the log uniform prior

$$\log(\pi) \sim \mathcal{U}(a, b), \tag{9}$$

with $a = \log(1/n_u)$ and $b = \log(M/n_u)$, so that the lower and upper limits on π correspond to an expectation of 1 and M variables included in the model, respectively, where M is a hyperparameter. The log uniform prior puts approximately equal probability on different magnitudes of sparsity (e.g., $10^{-3}, 10^{-4}$, and 10^{-5}) whereas a uniform prior would favor larger magnitudes. In GWAS, prior information on the sparsity of relevant SNPs is rarely available, making the log uniform prior an appealing choice.

Guan and Stephens (2011) showed the log uniform prior provides accurate posterior inference for a wide range of sparsities, but assuming a common sparsity across all chromosomes and within each chromosome may not always be appropriate. A growing synthesis of evidence suggests mutations contributing to phenotypic variation (i.e., the evolutionarily relevant SNPs) are not distributed uniformly across all genetic regions (Stern and Orgogozo, 2009). The vast majority of eukaryotic deoxyribonucleic acid is non-exonic and likely has no influence on phenotype (Van Straalen and Roelofs, 2011). Furthermore, for many traits, the SNPs associated with phenotype can be restricted to relatively few clustered loci in protein coding or cis-regulatory regions (Wang et al., 2010). In humans, for example, about 95% of disease-causing mutations occur in exonic regions that only make up $1-2\%$ of the entire genome (Posey, 2019). If a trait is monogenic, all relevant SNPs could fall within a gene of less than 20 kilobases (kb). Even for polygenic traits, evolutionary relevant mutations tend to accumulate near a reduced number of input-output gene nexuses in regulatory networks. In *Drosophila*, hundreds of genes regulate trichome (hair-like structures used for locomotion) development on the larval cuticle, but all evolutionarily relevant mutations are located in the cis-regulatory region of the input-output gene *shavenbaby* (McGregor et al., 2007). The relevant mutations are restricted to the cis-regulatory region of *shavenbaby* because mutations in the

upstream (input) genes or protein coding region of *shavenbaby* itself would interfere with organ development and would not persist in the population. Furthermore, coordinated expression among multiple downstream (output) genes is required for trichome development (Chanut-Delalande et al., 2006), such that mutations in a single downstream gene would not influence phenotype.

Parallel evolution provides additional evidence for genetic hotspots (Stern and Orgogozo, 2009). Many species, as well as reproductively isolated populations, have developed mutations targeting the same gene and regulatory networks. For example, 11 insect species have developed DDT resistance through mutations on one of two amino acids for the voltage-gated sodium channel gene *para* (ffrench Constant et al., 1998). In *Arabidopsis*, over 20 populations have independently evolved mutations for silencing the *Frigida* gene which induces early flowering (Shindo et al., 2005). The accumulation of evolutionarily relevant mutations in restricted genetic regions could be related to pleiotropy and epistasis. Mutations occurring in pleiotropic genes effect multiple traits and are unlikely to increase fitness thereby occluding their selection (Cooper et al., 2007). Likewise, the effect of mutations on epistatic genes is dependent on the genetic background and will experience slower selection compared to mutations that improve fitness for all genotypes. To summarize, the vast majority of mutations likely have no influence on phenotype. Mutations that reduce fitness are quickly removed from the population, and those which only improve fitness for certain genetic background may be selected in natural populations more slowly. This leaves a small number of genetic hotspots that harbor the majority of observable adaptive variation in natural populations. Motivated by the clustering of evolutionarily relevant mutations, we propose modeling SNP inclusion indicator jointly to aggregate spatial signals and increase power (Benjamini and Heller, 2007).

Spatial patterns in association statistics are ubiquitous in GWAS. Nearby SNPs tend to have similar marginal associations as a result of strong mutual correlations resulting from limited recombination during meiosis, a pattern known as linkage disequilibrium (Jorde, 2000). Figure 1 shows a Manhattan plot of marginal associations with flowering time in *Arabidopsis thaliana* for the full set of $\approx$ 7 SNPs and a reduced set of $\approx$ 550k SNPs that have been filtered for linkage disequilibrium (see Section 2.5). Flowering time in *Arabidopsis thaliana* is known to have a complex polygenic architecture (Zan and Carlborg, 2018), and this is reflected by many moderate signals distributed throughout the genome. We also see spatial structure among signals with the largest marginal associations often occurring in clusters. Clusters in the Manhattan plot are more pronounced for the full set of SNPs, but spatial signals are still prevalent in the linkage disequilibrium filtered set. The remaining spatial signal for the filtered set may be related to functional genomics because nearby SNPs are more likely to be part of the same gene networks and regulatory pathways.

Motivated by the clustering of evolutionarily relevant mutations in GWAS, we investigated accounting for spatial dependence in the SNP inclusion indicators using Markov random fields. We considered the Markov random field represented by a physically contiguous arrangement of SNPs in each chromosome. In reality, SNPs are separated by thousands of codons, but modeling them as contiguous has been argued as a viable

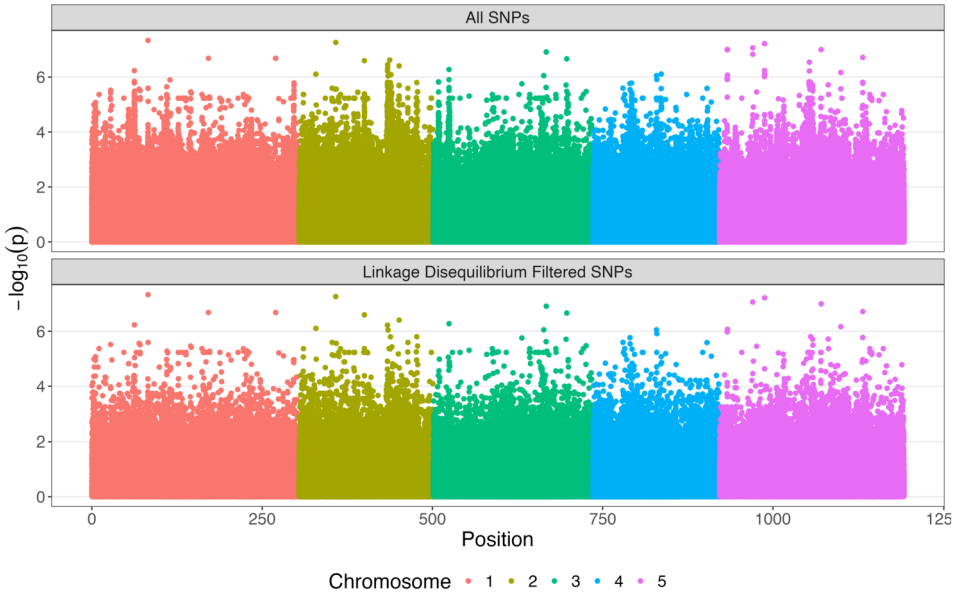

Figure 1: Manhattan plot of marginal associations with flowering time for ≈ 7 million single nucleotide polymorphisms (SNPs) in *Arabidopsis thaliana*. The vertical axis is the negative log (base 10) of the p-values from t-tests for univariate linear regressions of each SNP with flowering time. The bottom panel shows the same marginal associations but for a reduced set of ≈ 550k (SNPs) that have been filtered for linkage disequilibrium.

approach (Sesia et al., 2019). Li and Zhang (2010) explored modeling the inclusion indicators of variables in the Markov random field using an Ising model. Under the linear chain Ising model Li and Zhang (2010) proposed for GWAS, the inclusion probability of the jth SNP conditional on the other inclusion indicators is

$$P(\nu_j = 1 | \boldsymbol{\nu}_{-j}, c, d) = \text{logit}^{-1}(c(\nu_{j-1} + \nu_{j+1}) + d), \tag{10}$$

where c and d are parameters for controlling the clustering and sparsity of relevant SNPs, respectively. Li and Zhang (2010) showed the Ising model improved power to detect relevant SNPs for simulated spatial signals. A computational hurdle for fully Bayesian inference of c and d under the linear chain Ising model is the normalizing factor in the joint likelihood,

$$[\boldsymbol{\nu} | c, d] \equiv \frac{1}{C(c,d)} \exp\left(c \sum_{j=1}^{n_u - 1} \nu_j \nu_{j+1} + d \sum_{j=1}^{n_u} \nu_j \right). \tag{11}$$

The normalizing factor $C(c, d)$ is the sum over all 2^{n_u} configurations of $\boldsymbol{\nu}$ and is infeasible to evaluate for large n_u. Li and Zhang (2010) fixed c and d but this approach is unappealing for GWAS because the sparsity of the relevant SNPs is rarely known (Guan and Stephens, 2011).

We developed a more flexible model for inducing spatial dependence in inclusion indicators that allows for fully Bayesian inference using a reduced rank Gaussian process. Within each chromosome, we modeled the correlation in the SNP inclusion probabilities with the conditional autoregressive process,

$$\text{logit}(\boldsymbol{\pi}) \sim \mathcal{N}(\mu_\pi \mathbf{1}, \tau \boldsymbol{L}(\rho)), \tag{12}$$

where $\boldsymbol{L}(\rho) = (\text{diag}(\boldsymbol{A}\mathbf{1}) - \rho\boldsymbol{A})^{-1}$, $\boldsymbol{A}$ is a proximity matrix, and $\mathbf{1}$ is a $n_u \times 1$ column of ones such that $\boldsymbol{A}\mathbf{1}$ is the row sums of $\boldsymbol{A}$ (Ver Hoef et al., 2018; Hooten and Hefley, 2019). We specified the neighborhood structure,

$$a_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j,\ d_{ij} \leq d_{\text{thresh}} \\ 0, & i \neq j,\ d_{ij} \geq d_{\text{thresh}} \end{cases}, \tag{13}$$

where d_{ij} is the number of base pairs between SNP i and j, and d_{thresh} is selected *a priori*. We assume neighboring SNPs should have nearly identical inclusion indicators and set $\rho \to 1$ to induce an approximate intrinsic conditional autoregressive (ICAR) process (Ver Hoef et al., 2018).

Inverting $\boldsymbol{L}(\rho)$ is prohibitive for the large n_u typically encountered in GWAS. To reduce computational burden, we used a basis function approach for incorporating spatial dependence (Hefley et al., 2017), and let

$$\text{logit}(\boldsymbol{\pi}) = \mu_\pi \mathbf{1} + \boldsymbol{B}\boldsymbol{\alpha}, \tag{14}$$

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \tau \boldsymbol{B}'\boldsymbol{L}(\rho)\boldsymbol{B}), \tag{15}$$

where $\boldsymbol{B}$ is a basis expansion of $\boldsymbol{L}(\rho)$. Specifically, we let $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}$, where $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ is the spectral decomposition of $\frac{n_u}{\mathbf{1}'\boldsymbol{A}\mathbf{1}}\boldsymbol{A}$. The basis functions are sinusoidal with equal amplitude but increasing frequency allowing for more fine-scale spatial patterns (Reich and Hodges, 2008). We reduced computational burden by using the first n_α vectors from the basis expansion. The accuracy of the approximation decreases with fewer vectors included but selecting $n_\alpha << n_u$ generally has negligible effects on posterior inference (Hughes and Haran, 2013). We show in the Web Supplement (Van Ee et al., 2025) that the basis functions are insensitive to the choice of d_{thresh}.

2.4 Controlling False Discovery Rate via Knockoffs

Accounting for population structure with principal components as in equation (6) reduces spurious associations but does not provide information on the expected number of false discoveries. The prevailing practice in BVSR is to select all variables with posterior inclusion probabilities greater than 0.5 (George and McCulloch, 1993; O'Hara and Sillanpää, 2009). Referred to as the median probability model, Barbieri and Berger (2004) showed that for linear regression using a 0.5 threshold minimizes predictive error, but no theoretical guarantees are implied for FDR control. We calibrated selection of relevant SNP to control FDR using a knockoff variable approach.

Barber and Candès (2015) introduced the knockoff filter for general variable selection and several extensions have been proposed in the context of GWAS (Candès et al., 2018; Sesia et al., 2019, 2020, 2021). Knockoffs are synthetic variables constructed to be exchangeable with the original predictors but independent of the response. The knockoff filter leverages the synthetic variables to calibrate the selection procedure such that the FDR is controlled at the desired level. Unlike traditional approaches that control FDR asymptotically or assume independence of the tested hypotheses (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001), FDR control with the knockoff filter is exact for finite sample sizes regardless of the design or covariates, the number of variables in the model, or the signal-to-noise ratio (Barber and Candès, 2015). The knockoff filter can be applied in a wide range of models, but for brevity, we describe the method in the context of our BVSR model.

Consider the augmented model,

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \phi], \tag{16}$$

$$f(\boldsymbol{\mu}) = \boldsymbol{R\theta} + \boldsymbol{X\beta} + \boldsymbol{Ku} + \tilde{\boldsymbol{K}}\tilde{\boldsymbol{u}}, \tag{17}$$

where $\tilde{\boldsymbol{K}} = (\boldsymbol{I} - \boldsymbol{P_R})\boldsymbol{Z}\tilde{\boldsymbol{G}}$ and $\tilde{\boldsymbol{G}}$ is a knockoff of $\boldsymbol{G}$ with n_g synthetic genotypes for n_u SNPs. The vector $\tilde{\boldsymbol{u}}$ controls the effect of the knockoff SNPs on $\boldsymbol{y}$. We describe the properties and construction of $\tilde{\boldsymbol{G}}$ in the following sections. We expressed a joint spike-and-slab model for the original and knockoff SNP effects, $(u_j, \tilde{u}_j)$, as

$$(u_j, \tilde{u}_j) \sim (\delta_j, \tilde{\delta}_j)\mathcal{N}(0, \sigma_a^2), \tag{18}$$

$$(\delta_j, \tilde{\delta}_j) \sim \begin{cases} (0,1), \text{ w.p. } 1/2 & \text{for } \nu_j = 1 \\ (1,0), \text{ w.p. } 1/2 & \text{for } \nu_j = 1 \\ (0,0), & \text{for } \nu_j = 0 \end{cases}, \tag{19}$$

$$\nu_j \sim \text{Bernoulli}(\pi_j), \tag{20}$$

where w.p. is an abbreviation for "with probability." We define the quantity $w_j = \delta_j - \tilde{\delta}_j$ and denote its posterior mean $\bar{w}_j = \mathbb{E}(w_j|\boldsymbol{y})$. True discoveries are indicated by $\bar{w}_j > 0$ whereas false discoveries (i.e., selected SNPs with no association to phenotype conditional on the inclusion of all relevant SNPs) correspond to $\bar{w}_j \leq 0$. We also redefine the variance of the original and knockoff SNP effects

$$\sigma_a^2(h, \boldsymbol{\delta}, \tilde{\boldsymbol{\delta}}, \boldsymbol{s}^2, \tilde{\boldsymbol{s}}^2) = \frac{h}{1-h}\frac{1}{\sum_{j:\delta_j=1} s_j^2 + \sum_{j:\tilde{\delta}_j=1} \tilde{s}_j^2}, \tag{21}$$

so that σ_a^2 shrinks as either type of variable is added to the model.

Suppose we select all SNPs having $\bar{w}_j > t^\star$ for some $t^\star \in (0,1)$. We let $\hat{S} \subset \{1, \ldots, n_u\}$ be the subset of SNPs selected. The FDR of this procedure is

$$\text{FDR} = \mathbb{E}\left(\frac{\#\{j : u_j = 0 \text{ and } j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} \vee 1}\right), \tag{22}$$

where we use the notation $a \vee b$ to denote $\max\{a, b\}$. The goal is to chose $t^\star$ as small as possible subject to the constraint that FDR is controlled below some prespecified threshold q. Barber and Candès (2015) proved the optimal $t^\star$ is given by

$$t^\star = \min\left\{t \in (0,1) : \frac{1 + \#\{j : w_j \leq -t\}}{\#\{j : w_j \geq t\} \vee 1} \leq q\right\}. \tag{23}$$

The threshold, $t^\star$, in equation (23) controls the expected number of false discoveries, but note that for any one analysis the observed proportion of false discoveries may exceed q.

'Model-X' Knockoffs

Knockoff variables as proposed in Barber and Candès (2015) are constructed geometrically and only valid if $n_u < 2n_g$. Candès et al. (2018) introduced probabilistically constructed 'Model-X' knockoffs for high-dimensional variable selection. Given a family of random variables $\boldsymbol{g} = (g_1, \ldots, g_{n_u})'$, a 'Model-X' knockoff, $\tilde{\boldsymbol{g}} = (\tilde{g}_1, \ldots, \tilde{g}_{n_u})'$, satisfies two properties:

1. for any subset $S \subset \{1, \ldots, n_u\}$, $(\boldsymbol{g}, \tilde{\boldsymbol{g}})_{\text{swap}(S)} \overset{d}{=} (\boldsymbol{g}, \tilde{\boldsymbol{g}})$,
2. $\tilde{\boldsymbol{g}} \perp\!\!\!\perp \boldsymbol{y} | \boldsymbol{g}$,

where $\overset{d}{=}$ denotes equality in distribution and $(\boldsymbol{g}, \tilde{\boldsymbol{g}})_{\text{swap}(S)}$ is obtained by swapping the variables g_j and $\tilde{g}_j$ for all $j \in S$. Following Candès et al. (2018), henceforth, we refer to criteria 1 and 2 as the exchangeability and nullity of knockoffs, respectively.

A trivial knockoff satisfying exchangeability and nullity is $\tilde{\boldsymbol{g}} = \boldsymbol{g}$. This knockoff would be of little practical use because $\bar{w}_j = 0$ for all j yielding no power. As a more relevant example, if the variables follow a Gaussian distribution, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, one possible knockoff construction is

$$(\boldsymbol{g}, \tilde{\boldsymbol{g}}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}), \text{ where } \boldsymbol{H} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) \\ \boldsymbol{\Sigma} - \text{diag}(\boldsymbol{s}) & \boldsymbol{\Sigma} \end{pmatrix}, \tag{24}$$

and $\text{diag}(\boldsymbol{s})$ is any diagonal matrix selected in such a way that the joint covariance matrix is positive definite. In general, knockoffs become more powerful as the absolute pairwise correlation between each variable and its knockoff decreases.

Hidden Markov Model Knockoffs

The knockoff construction presented in equation (24) is challenging to implement for high-dimensional data. As the number of variables, n_u, grows, the domain of $\boldsymbol{s}$ that ensures $\boldsymbol{H}$ is positive definite collapses, and knockoffs generated using equation (24) become indistinguishable from the original variables. Simulating $\tilde{\boldsymbol{g}}$ from equation (24) has computational complexity $\mathcal{O}(n_u^3)$ and memory requirement $\mathcal{O}(n_u^2)$, which may not

be feasible. The Gaussian assumption is particularly unappealing in the context of GWAS where $\boldsymbol{g}$ is discrete-valued, and simulations from Rosenblatt et al. (2019) suggest that discrete knockoffs generated under the Gaussian construction often violate exchangeability.

Sesia et al. (2019) developed a knockoff generator, `SNPknock`, for SNP data based on the hidden Markov model (HMM) described by Scheet and Stephens (2006) that addresses the aforementioned challenges. The HMM approach produces discrete knockoffs and accommodates linkage disequilibrium patterns in haplotype blocks that lead to high mutual correlation among neighboring SNPs (Jorde, 2000). Compared to the Gaussian knockoff construction, `SNPknock` substantially reduces computational burden and has computational complexity and memory requirements $\mathcal{O}(n_u^2)$ and $\mathcal{O}(n_u)$, respectively. Simulations using real genotype matrices demonstrated knockoffs generated using `SNPknock` control the false discovery rate while yielding high power (Sesia et al., 2019).

In principle, the knockoffs generated with `SNPknock` should also account for population structure, but this may not be the case if the SNPs have been heavily filtered prior to analysis. Sesia et al. (2021) extended their approach to accommodate genome-wide correlations related to population structure. Specifically, kinship coefficients and principal components calculated from knockoffs generated using the new algorithm, `knockoffGWAS`, and from a reduced number of SNPs are indistinguishable from the original genotypes. We describe our implementation of their HMM framework for the *Arabidopsis thaliana* in Section 2.5.

2.5 Implementation

Identifying the relevant SNPs in GWAS is particularly challenging because of the massive number of variables considered and high mutual correlations among SNPs arising from linkage disequilibrium (Weinwurm et al., 2013). Clustering and pruning SNPs prior to analysis can improve performance (Wang et al., 2010; Lu et al., 2015; Candès et al., 2018; Sesia et al., 2019), and several approaches have been suggested in the context of GWAS (Selinski and Ickstadt, 2008; Wang et al., 2015; Candès et al., 2018). While clustering and pruning variables based on correlation prior to selection has several disadvantages (Lippitt et al., 2024), correlational clustering may be warranted in GWAS (Candès et al., 2018; Sesia et al., 2019, 2020; He et al., 2024). Fitting models to all SNPs jointly is often computationally infeasible necessitating some degree of variable thinning. Furthermore, a common post hoc analysis in GWAS involves augmenting the group considered for genetic fine mapping by adding SNPs in linkage disequilibrium with the selected ones using `PLINK` (Purcell et al., 2007). The motivation for post hoc aggregation is that the unit of inference is not an individual SNP but genomic regions. However, if FDR control is specified at the SNP level, as is common practice, this standard pipeline creates a mismatch between the units for true and false discoveries (Benjamini and Heller, 2007; Brzyski et al., 2017).

The solution proposed by recent GWAS methods (Candès et al., 2018; Sesia et al., 2019, 2020; He et al., 2024) involves clustering SNPs prior to analysis and selecting

individuals within clusters as representatives for broader genomic associations that are identifiable from a variable selection perspective. Correlational clustering achieves both of these efforts by grouping SNPs using genetic distance, which coincides with correlation as a result of linkage disequilibrium, and removing SNPs that are too similar to discriminate. We hierarchically clustered the SNPs using their absolute correlation as a measure of similarity, and cut the dendrogram at the height such that collections of SNPs having mutual correlations of 0.5 or more were classified into clusters. We then used 20% of the observations of flowering time to calculate marginal t-tests for each SNP within clusters and choose the SNP with the lowest p-value as the cluster representative. A full description and justification of this procedure is provided in the Web Supplement of Candès et al. (2018).

The commonly used variable pruning procedure reduced the total number of SNPs considered to $n_u = 558{,}321$. Larger n_u increases the challenge of simulating knockoffs that satisfy exchangeability in Section 2.4 and are distinct from the original genotypes (Sesia et al., 2021). To improve power, we generated knockoffs for the pruned genotype using `SNPknock` (Sesia et al., 2019) rather than simulating knockoffs for all ≈ 7 million SNPs. *Arabidopsis thaliana* is a highly self-fertilizing species, and it is assumed that the wild collected lines are fully inbred. The genotype matrix available for *Arabidopsis thaliana* is a variant matrix where 1 and 0 encode homozygous for the reference and alternate alleles, respectively (Togninalli et al., 2018). Because the genotype matrix is binary, we used the haplotype implementation of `FastPhase` and `SNPknock` even though *Arabidopsis thaliana* is diploid. This implementation reflects the breeding structure of *Arabidopsis thaliana* in that both haplotypes are inherited from the same parent and therefore identical. Note that while `SNPknock` does not explicitly account for population structure, principal components calculated from the knockoffs were similar to those calculated from the original genotypes (see Web Supplement). Following Candès et al. (2018), we set the rows of the knockoff genotype matrix corresponding to the observations used for determining the cluster representatives to their original values to ensure they met the exchangeability lemma.

We obtained a posterior sample for all unobserved quantities in our model using Markov chain Monte Carlo (MCMC). Implementing BVSR for large sets of variables with MCMC is computationally demanding (Griffin et al., 2021). A pivotal component of making BVSR computationally feasible in GWAS is the Rao-Blackwellization (Casella and Robert, 1996) of the SNP marginal posterior inclusion probabilities (Guan and Stephens, 2011). With the massive number of SNPs considered, mixing of the inclusion indicators, ν_j, is poor, and Monte Carlo estimates of the posterior inclusion probabilities calculated as the proportion of MCMC samples for which $\nu_j = 1$ are prone to high sampling variance. Guan and Stephens (2011) suggested a Rao-Blackwellized estimate for SNP posterior inclusion probabilities that can dramatically improve mixing. In our approach, the knockoff statistics are the difference in posterior inclusion probabilities for the SNP and its knockoff and are amenable to Rao-Blackwellization regardless of whether inclusion probabilities have spatial dependence. The derivations of the Rao-Blackwellized knockoff statistics are provided in the Web Supplement as well as a description of several additional sampling strategies we used to improve convergence. For the Ising model, we obtained samples of c and d using a Gaussian random walk

Metropolis-Hastings algorithm. The normalizing function in the Ising model is computationally intractable, and we used the pseudolikelihood (Besag, 1975) approximation $[\boldsymbol{\nu}|c,d] \approx \prod_{j=1}^{n_u}[\nu_j|\boldsymbol{\nu}_{-j},c,d]$.

3 Results

We fit our spatial BVSR model with an ICAR process on SNP inclusion probabilities as in (12) to a variety of simulated datasets and flowering time observations for 1058 wild accessions of *Arabidopsis thaliana*. We also fit a spatial BVSR model with an Ising process and a non-spatial BVSR model with a log uniform prior on SNP inclusion probabilities (equations (11) and (9), respectively). To quantify the relative advantages of the BVSR approach over other common GWAS methods, we fit modified LASSO regression (Tibshirani, 1996) and linear mixed models (Kang et al., 2010) to all datasets. Specifically, we fit a LASSO regression model to the genotype matrix augmented with the same knockoffs used in the BVSR models. The model also included the same environmental covariates and population structure basis functions (for datasets simulated with population structure), but no penalization was applied to these effects. We calculated the knockoff statistics for each SNP as the difference in the magnitude of effects for the original and knockoff variables. This model and knockoff procedure follows the approach presented by Sesia et al. (2019). All linear mixed models were fit using efficient mixed-model association eXpedited (EMMAX) with the same kinship matrix from which the population structure basis functions were derived (Kang et al., 2010). We estimated the effects of the original and knockoff variables for each SNP and calculated knockoff statistics as the difference in magnitudes like before.

While Model-X knockoffs do not place any assumptions on the distribution of $\boldsymbol{y}|\boldsymbol{G}$, our BVSR model as well as its competitors (e.g., LASSO) implicitly assume sparse and linear associations between SNPs and phenotype. The assumption of sparsity is valid because the majority of genetic variants are known to have no impact on phenotype, but effects of the relevant SNPs are undoubtedly non-linear and interactive (Zuk et al., 2012). To capture the discrepancy between the biological processes that give rise to phenotype and the models used for detecting relevant SNPs, we simulated the effect of each relevant SNP and phenotype using a neural network framework with rectified linear unit (ReLU) link function to capture non-linear genotype-phenotype associations. Across simulations we varied the signal-to-noise ratio, degree of spatial structure, and the inclusion of both population structure and linkage disequilibrium. In addition to measuring the false discovery and true positive rate, we also assessed out-of-sample predictive performance for the BVSR and LASSO models by withholding a number of genotypes from model fitting for each simulated dataset. For the analysis of Arabidopsis flowering time, we randomly generated 10 folds of 100 genotypes and let the eleventh fold contain the remaining 58 genotypes. We provide full model statements, data simulation and preprocessing details, and model fitting statistics for the simulated and real datasets in Web Supplements B and C, respectively.

Power decreased in all models as the signal-to-noise ratio increased (Figure 2). Power was generally lowest for the single SNP approach (EMMAX) and LASSO. Performance

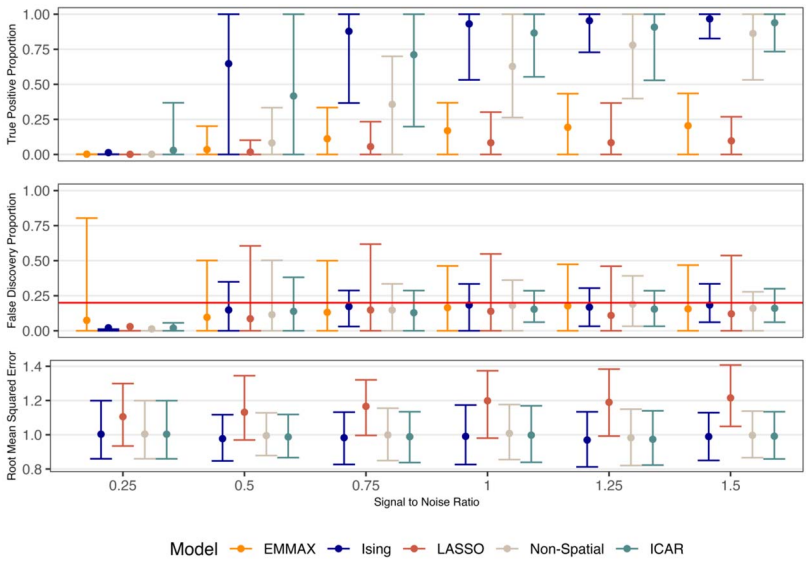

Figure 2: Means, 5th, and 95th quantiles of false discovery proportion, true positive proportion, and out-of-sample predictive performance. The true positive proportion is the number of relevant SNPs correctly identified divided by 30, the true number of simulated relevant SNPs. We varied the signal-to-noise ratio by setting $\sigma_e^2 = 1$ in equation (2) and letting $u_j = 0.25, 0.50, \ldots, 1.5$ in equation (3) for the 30 relevant SNPs. Results are pooled across six clustering regimes: all relevant SNPs in 1, 3, 5, 10, 15, and 30 clusters, respectively. We simulated 10 datasets for each clustering regime and signal-to-noise ratio, for a total of 360 datasets. All simulated datasets are for $n_u = 20,000$ uncorrelated SNPs (no linkage disequilibrium) observed across $n = 1,000$ unrelated individuals with no population structure. The red horizontal line depicts the targeted false discovery proportion of $q = 0.20$. The root mean squared error was calculated for 100 random genotypes that were withheld from model fitting and is divided by the standard deviation of $\boldsymbol{y}$.

was similar for the three BVSR approaches, but the spatial BVSR approaches had a higher upper bound. All three approaches gave similar power for datasets simulated without spatial structure, but when the relevant SNPs were clustered, the spatial BVSR models attained higher power (Figure 3). For datasets with moderate levels of spatial structure (i.e., clusters of 3-6 relevant SNPs), the Ising model had higher power than the ICAR model. As expected, the mean FDR did not depend on the signal-to-noise ratio or clustering regime and was close to the targeted rate of $q = 0.20$. There was little difference in predictive performance across the BVSR methods. Even for datasets for which the spatial BVSR models attained significantly higher power, there was no appreciable difference in predictive performance. Note that if a variable had a large effect on predictive performance, it would be identified well by each method. The spatial BVSR models leverage dependence among the relevant variables to detect predictors

that are only weakly associated with the response.

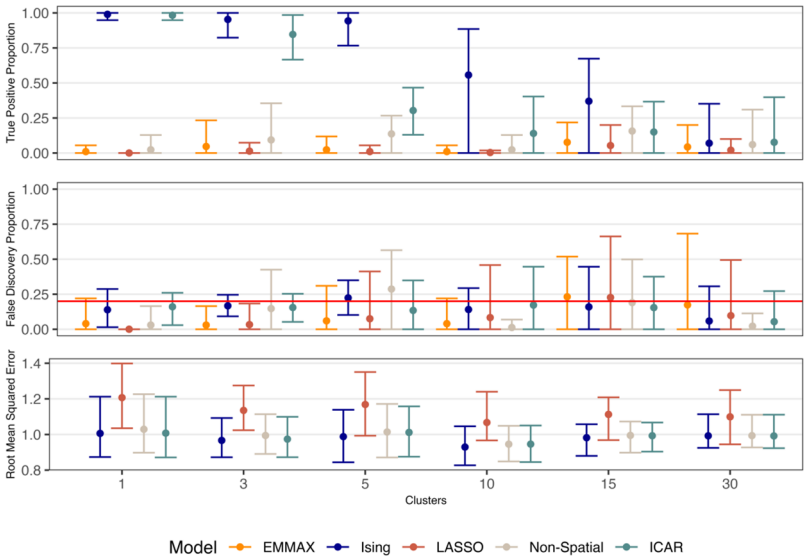

Figure 3: Means, 5th, and 95th quantiles of false discovery proportion, true positive proportion, and out-of-sample predictive performance. The true positive proportion is the number of relevant SNPs correctly identified divided by 30, the true number of relevant SNPs. The horizontal axis depicts the number of clusters of relevant SNPs and ranges from extreme (all relevant SNPs in one cluster) to no spatial structure (random uniform position for all relevant SNPs). Results shown are for a signal-to-noise ratio of 0.5. Results are summarized across 10 datasets for each clustering regime. All simulated datasets are for $n_u = 20,000$ uncorrelated SNPs (no linkage disequilibrium) observed across $n = 1,000$ unrelated individuals with no population structure. The red horizontal line depicts the targeted false discovery proportion of $q = 0.20$. The root mean squared error was calculated for 100 random genotypes that were withheld from model fitting and is divided by the standard deviation of $\boldsymbol{y}$.

Power decreased for all methods in the presence of linkage disequilibrium and population structure (Figure 4). When SNPs are correlated, associations can become masked lowering power to detect relevant SNPs. For the datasets simulated without population structure and linkage disequilibrium, the relevant SNPs tended to be among the top 100 SNPs with highest marginal association, but when we introduced population structure and linkage disequilibrium, it was not uncommon to observe relevant SNPs with marginal associations that ranked in the bottom half. This trend is exacerbated with increased clustering of the relevant SNPs because linkage disequilibrium decays with distance, and we observed that power slightly decreased for the LASSO and non-spatial BVSR model as we increased spatial dependence. The spatial BVSR models, on the other hand, increased in power as the relevant SNPs became more clustered. For extreme spatial clustering (i.e., all 30 relevant SNPs in a single cluster), the ICAR model

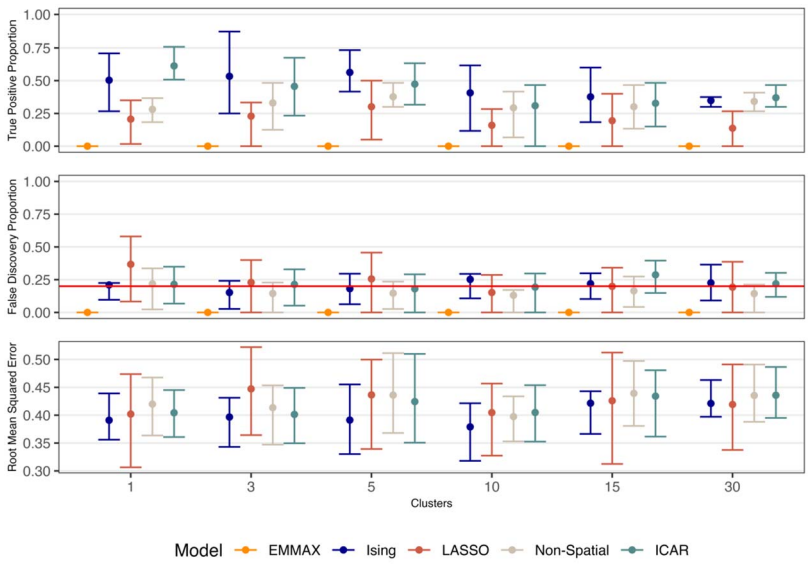

Figure 4: Means, 25th, and 75th quantiles of false discovery proportion, true positive proportion, and out-of-sample predictive performance. The true positive proportion is the number of relevant SNPs correctly identified divided by 30, the true number of relevant SNPs. The horizontal axis depicts the number of clusters of relevant SNPs and ranges from extreme (all relevant SNPs in one cluster) to no spatial structure (random uniform position for all relevant SNPs). Results shown are for a signal-to-noise ratio of 1 (i.e., $\sigma_e^2 = 1$ in equation (2) and $u_j = 1$ in equation (3)). Results are summarized across 27 datasets for each clustering regime. All simulated datasets are for $n_u = 20,000$ SNPs extracted from chromosomes 1-5 of *Arabidopsis thaliana* and $n = n_g = 1,058$ individuals with population structure mimicking *Arabidopsis thaliana*. Linkage disequilibrium among SNPs was partially attenuated using the variable pruning procedure described in Section 2.5. The red horizontal line depicts the targeted false discovery proportion of $q = 0.20$. The root mean squared error was calculated for 20 random genotypes that were withheld from model fitting and is divided by the standard deviation of $\boldsymbol{y}$.

had the greatest power, but, as before, the Ising model showed slightly greater power for datasets with more moderate levels of spatial dependence. The non-spatial BVSR model generally had greater power than the LASSO model. The EMMAX method failed to detect any relevant SNPs at the targeted false discovery proportion. As before, there was no appreciable difference in predictive performance across approaches.

Figure 5 presents the knockoff statistics for $n_u = 558{,}321$ SNP cluster representatives in *Arabidopsis thaliana*. At a false discovery proportion threshold of 0.20, we made 0, 25, 27, 40, and 65 discoveries with the EMMAX, Ising BVSR, LASSO, non-spatial BVSR, and ICAR BVSR models, respectively. We identified SNPs likely tagging flowering time genes by noting whether any of the SNPs in their cluster fell within 10 kb, the estimated

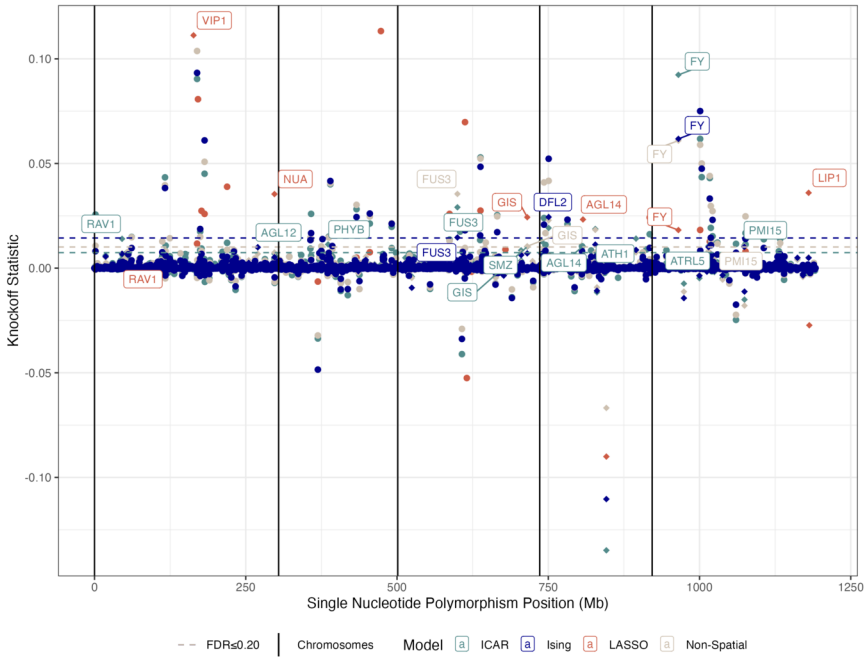

Figure 5: *Arabidopsis thaliana* flowering time knockoff statistics for $n_u = 558{,}321$ single nucleotide polymorphism (SNP) cluster representatives considered. Labels denote the 3, 7, 6, and 14 knockoff statistics for SNP clusters falling in buffers of flowering time genes selected in the Ising BVSR, LASSO, non-spatial BVSR, and ICAR BVSR models, respectively. Knockoff statistics have been scaled to yield the same variance across methods. The five vertical lines depict the starting points of each chromosome. The dashed horizontal lines depict the threshold needed to obtain the targeted false discovery proportion of $q = 0.20$. Shape indicates whether a SNP cluster fell within a 10 kilobase buffer of one of the 282 described flowering time genes in Brachi et al. (2010).

linkage disequilibrium rate of *Arabidopsis thaliana* (Kim et al., 2007), of a described gene. The number of cluster representatives in flowering time gene buffers was 0, 3, 7, 6 and 14, respectively. Approximately 15.8% of the 558,321 SNP clusters in our subset had a cluster member that fell within the 10 kb buffer of one of the 282 flowering time genes described in Brachi et al. (2010). Thus, in the non-spatial and Ising BVSR models, we selected roughly the proportion of SNP clusters tagging flowering time genes that we would have expected by chance, whereas in the LASSO and spatial BVSR models, we selected a higher proportion of SNP representatives tagging flowering time genes (26% and 22%, respectively).

All four models selected SNP clusters that had representatives within the 10 kb buffers of the flowering time genes AT5G13480 (*FY*) and three of four models had a representative in AT3G58070 (*GIS*). *FY* had one of the strongest estimated associations and encodes a protein with similarity to yeast Pfs2p, a messenger ribonucleic acid

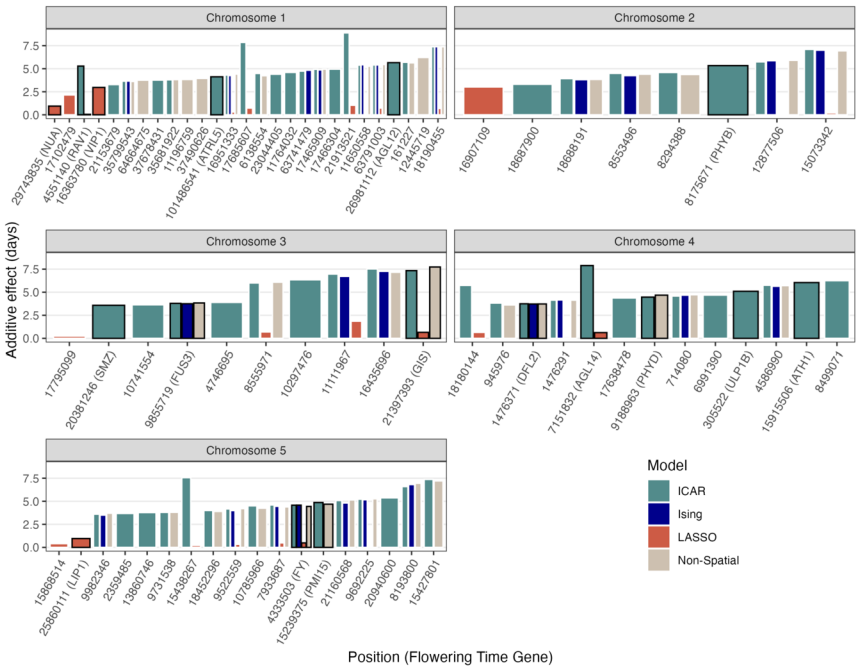

Figure 6: *Arabidopsis thaliana* change in flowering time effects for 25, 27, 40, and 65 single nucleotide polymorphism (SNP) cluster representatives selected in the Ising BVSR, LASSO, non-spatial BVSR, and ICAR BVSR models, respectively. Effects represent the change in days to flowering induced by swapping two copies of the early flowering allele for two copies of the late flowering allele at a particular locus. Selected SNP clusters falling within a 10 kilobase buffer of one of the 282 described flowering time genes in Brachi et al. (2010) are depicted by columns with a black outline and gene names are given in parenthesis. Bar widths vary based on the number of discoveries in each chromosome and how many methods identified a SNP.

processing factor that influences flowering time (Schmid et al., 2005). In the BVSR models, associations were also strong for AT3G26790 (*FUS3*), which regulates the hormones gibberellin and abscisic acid that can influence flowering time via their interactions with the transcription of the gene AT5G61850 (*LEAFY*) (Gazzarrini et al., 2004). Many of the selected SNP clusters, while not directly linked to genes in the list described by Brachi et al. (2010), are likely involved in the biochemical pathways regulating flowering time. For example, two additional discoveries made by the ICAR BVSR model in chromosome 4 had cluster members in the genes AT4G01650 and AT4G02120 (*CTPS3*), which are expressed during many development stages including flowering (Schmid et al., 2005).

Figure 6 gives the absolute value of the coefficients, u_j, for the subset of SNPs selected in each model. Because $\boldsymbol{G}$ is a variant matrix (i.e., binary), the absolute value

of the coefficients, u_j, represents the expected change in days to flowering associated with replacing both copies of the early flowering allele with the late flowering allele at locus j. For the BVSR models, effects varied from 2-9 days and were similar for cluster representatives selected in both models. Effects were shrunk near zero for the LASSO model. The mean, minimum, and maximum relative root mean squared error across the 11 folds of out-of-sample data were 0.97 $(0.90, 1.10)$, 1.04 $(0.95, 1.14)$, 1.01 $(0.90, 1.22)$, and 1.03 $(0.95, 1.14)$ for the LASSO, non-spatial BVSR, Ising BVSR, and ICAR BVSR models, respectively. Note that for the LASSO model, our estimates for out-of-sample predictive performance are optimistic because we used the same observations both for choosing the optimal penalty value and model fitting. The LASSO model also applied greater regularization to the SNP effects than did the BVSR models (Figure 6), which could also explain its improved predictive performance (Hooten and Hobbs, 2015).

4 Discussion

We developed a spatial BVSR model for improving power and controlling false discoveries. While we motivated and validated our spatial BVSR model in the context of GWAS, our method is more broadly applicable for high-dimensional variable selection and is tailored for contexts in which the relevant variables have latent spatial structure (e.g., gene expression (Stingo et al., 2011) or functional MRI studies (Li and Zhang, 2010)). Variable selection in GWAS is particularly challenging because SNP effects are often confounded with population structure. We demonstrated how to alleviate the confounding between SNP and population structure effects with restricted regression. Our restricted regression approach is appealing in GWAS because the SNP effects are estimated from the residuals of $\boldsymbol{y}$, or "corrected" trait, as proposed in previous methods (Guan and Stephens, 2011; Li et al., 2011; Lu et al., 2015; Sesia et al., 2021). However, rather than correcting for population structure and then estimating SNP effects sequentially, we estimate all quantities simultaneously in a fully Bayesian model that accounts for all sources of uncertainty.

In both the simulation study and analysis of *Arabidopsis thaliana* flowering time, we showed that jointly modeling SNP inclusion probabilities can improve power when the relevant variants have latent spatial structure. In the analysis of flowering time, the ICAR BVSR model had the most discoveries and a greater proportion of discoveries near described flowering time genes. In simulations, the Ising and ICAR BVSR models generally gave similar performance, but, for the flowering time analysis, the Ising approach was much less powerful. We suspect the lower power of the Ising approach is related to overly localized neighborhood effects. SNP inclusion indicators are only dependent on the two nearest neighbors, which is likely too fine-scale for the 550k SNPs considered. We could broaden dependency in the Ising model by increasing neighborhood size, but it is unclear how to choose a suitable neighborhood structure *a priori*. An analogous choice for the ICAR model relates to specifying d_{thresh} in equation (13), but the ICAR approach is insensitive to the choice of d_{thresh}, and we used the same adjacency matrix for both the simulation study and real data analysis. Thus, while it does increase computation burden, the ICAR approach offers more flexibility.

The case study of *Arabidopsis thaliana* flowering time also highlighted the advantages of the BVSR approach for identifying relevant variants as compared to standard single SNP approaches. In addition to being more powerful than the traditional single SNP approaches, the BVSR model also selected variables that varied considerably in their marginal associations. For example, among the 65 SNPs selected in the ICAR BVSR model, rankings based on marginal associations varied from 2 to 2654, with only 9 SNP clusters selected from the top 100 ranks. In the BVSR model, many of the low ranked SNPs are uncorrelated with the trait conditional on the hundreds of other SNPs included in the model, and hence go unselected. Congruent with our focus on the infinitesimal genetic model (Barton et al., 2017), we simulated phenotypes influenced by many SNPs with small effect sizes and validated our approach on flowering time, a trait with polygenic architecture. While the infinitesimal model is widely applicable, some traits may only be influenced by a few large effect loci for which single SNP testing approaches, like EMMAX, have been shown to outperform multiple SNP approaches (Kang et al., 2010; Buzdugan et al., 2016).

Zan and Carlborg (2018) provided a comprehensive description of the genetic architecture underpinning *Arabidopsis thaliana* flowering time and identified 33 SNPs that collectively described 55.1% of the total phenotypic variance in flowering time. Zan and Carlborg (2018) first reduced the number of SNPs considered by screening for SNPs associated with the 282 flowering time genes described in Brachi et al. (2010). The joint effect and associations of the remaining SNPs was then estimated with a backward elimination association analysis with an adaptive FDR based threshold of 0.15. The selected SNPs were congruent with functional genomics with 11 SNPs located within a flowering time gene.

While the adherence of the SNPs selected with our ICAR BVSR model to flowering time genes was less strong, the selected SNPs still tended to be closer to flowering time genes than we would expect by chance. In cases where prior information is available regarding the location of genes, Fridley (2009) and Stephens and Balding (2009) noted that performance could be improved by inflating the prior inclusion probabilities of these exonic SNPs and their cis-reglatory regions by specifying a non-zero mean for $\boldsymbol{\alpha}$ in equation (15). Similarly, if prior information is available on gene networks, we could extend our framework to include additional layers of dependence based on distance metrics related to functional genomics (i.e., dependence in inclusion indicators for SNPs belonging to the same gene network). For comparison with the non-spatial methods, we specified homogeneous prior inclusion probabilities for all SNPs in the *Arabidopsis thaliana* analysis and did not incorporate any prior information on gene networks. One caveat to correlating true discoveries with gene proximity is that in genetic fine mapping SNPs selected in GWAS can be linked to causal genes up to 2 Mbps away (Brodie et al., 2016). Furthermore, some causal SNPs show no associations to genes or their regulators entirely (Niu et al., 2019). Hence, even though a number of the SNPs selected from our BVSR model were far from described flowering time genes, their associations to flowering time is not necessarily spurious. Implementing knockoff variable approaches within LASSO for GWAS, Candès et al. (2018) and Sesia et al. (2019) also reported a number of discoveries that were not replicated in previous analyses.

Model-X knockoffs are appealing choice for FDR control in GWAS because they are compatible with a wide range of modeling frameworks that can accommodate additional sources of information (e.g., prior information or data on gene pathways and locations). While several more recently proposed variable selection methods may perform well in simple settings (Wang and Ramdas, 2022; Dai et al., 2023; Xing et al., 2023), it is not readily apparent how to incorporate biological information into these approaches. Given the biological complexity of how mutations contribute to phenotypic variation coupled with the variability in these mechanisms across organisms (Stern and Orgogozo, 2009), some have cautioned against restrictive assumptions on $\boldsymbol{y}|\boldsymbol{G}$ (Sesia et al., 2019, 2021). Under the Model-X knockoff approach, distributional assumptions are shifted to $\boldsymbol{G}$. In practice, the distribution of $\boldsymbol{G}$ is always unknown, but prior information is readily available from both well-described mechanisms of genetic inheritance and the plethora of observed genomes. Barber et al. (2020) showed that under the Model-X knockoff approach, inflation of the false discovery rate is proportional to the error in estimating the distribution of each feature conditional on all the rest (i.e., $\boldsymbol{g}_j|\boldsymbol{G}_{-j}$). The accuracy of several genotype imputation methods (Scheet and Stephens, 2006; Delaneau et al., 2012) suggest that this error will be small and justify the Model-X knockoffs approach for GWAS. Incorporating Model-X knockoffs into a fully parametric model, as we have described, necessarily implies other assumptions (i.e., linearity, normality, sparsity, etc.), but in the simulation study, we found that our BVSR model can detect relevant variants even when these assumptions are not met.

Because Model-X knockoffs are stochastic, different runs of an algorithm can produce discrepancies in the selected variables based on the generated knockoff variables (Ren et al., 2023). In our analysis of *Arabidopsis thaliana*, the lowest knockoff statistic for all three models occurred in chromosome 4 and is linked to a SNP cluster with representative in the 10 kb buffer of the flowering time gene AT4G20370 (*TSF*). AT4G20370 is the "twin sister" of AT1G65480 (*FT*), a previously identified flowering time gene for traditional GWAS approaches (Alonso-Blanco et al., 2016). Just by chance, the knockoff of this SNP was more associated with flowering time than the true SNP. Gu and Yin (2021) suggested treating the knockoffs as random variables and sampling them directly in the MCMC algorithm. An advantage of this approach is that it can stabilize the knockoff filter by attenuating issues related to only using one realization of the knockoff variables. In principle, we could embed a Bayesian implementation of the HMM proposed by Sesia et al. (2019) within our BVSR and sample the knockoff SNPs at each MCMC iteration, but this approach would be computationally infeasible for the large number of SNPs considered in most GWAS settings. We could take a derandomized knockoff approach and summarize the selected variables across multiple instances fit in parallel (Ren et al., 2023), but this would also be computationally infeasible for GWAS.

The consequences of clustering for variable selection and multiple testing have been discussed at length (Benjamini and Heller, 2007; Dai and Barber, 2016; Brzyski et al., 2017). In GWAS, treating SNPs as independent units may not be appropriate if the goal is to discover genomic regions associated with a phenotype (Wang et al., 2010; Lu et al., 2015; Brodie et al., 2016; Brzyski et al., 2017; Candès et al., 2018; Sesia et al., 2019). Variable selection within quantitative trait loci may not always be feasible because of high multicollinearity among clustered SNPs. We preprocessed the genotype

matrix by first clustering SNPs with high mutual correlations and then choosing cluster representatives (Candès et al., 2018; Sesia et al., 2019, 2020; He et al., 2024). Another option would be to include all SNPs but assign clusters to joint inclusion indicators as proposed by Lu et al. (2015) using group knockoffs (Dai and Barber, 2016). Our spatial BVSR model can be viewed as a more flexible version of this model that encourages nearby SNPs to have the same inclusion indicator as a result of the spatially structured inclusion probabilities. Modeling either the inclusion indicators or probabilities jointly has the potential to both improve power and reduce false discoveries because they can magnify true but negligible individual effects as well as dilute one-off spurious associations (Benjamini and Heller, 2007; Zhang et al., 2014; Lu et al., 2015; Brzyski et al., 2017; Lee et al., 2023).

Increased genomic and phenotypic data collection has highlighted the importance of methods for understanding the association between a response and an increasingly large number of predictors. Our proposed BVSR approach incorporates several recent advances in variable selection into one cohesive hierarchical model tailored for GWAS. Using restricted regression, we stabilized posterior computation for confounded factors that are generally estimated in a step-wise procedure. We achieved rigorous FDR control with knockoff variables customized to the study system. We corroborated previous research that showed incorporating biological mechanisms into the selection of variables can improve performance (Li and Zhang, 2010). Lastly, by combining reduced rank approximations and sampling strategies, we demonstrated the computational feasibility of Bayesian methods for high-dimensional variable selection.

Acknowledgments

We thank several members from the BromeCast research network for their contributions to this manuscript.

Funding

Justin J. Van Ee and Mevin B. Hooten were supported by NSF 2222525 and 1927177.

Diana Gamba and Jesse R. Lasky were supported by NSF 1927009.

Megan L. Vahsen was supported by NSF 1927282.

Supplementary Material

Web Supplement: Spatial knockoff Bayesian variable selection in genome-wide association studies (DOI: 10.1214/25-BA1556SUPP; .pdf). We provide the details of our Markov chain Monte Carlo algorithm, full model statements and data preprocessing for the simulation studies and flowering time analysis, and the derivation of Rao-Blackwellized estimates of the knockoff statistics.

References

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K. J., Schmitz, R. J., Ümit Seren, Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. (2016). "1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*." *Cell*, 166(2): 481–491. 4, 23

Barber, R. F. and Candès, E. J. (2015). "Controlling the false discovery rate via knockoffs." *The Annals of Statistics*, 43(5): 2055–2085. MR3375876. doi: https://doi.org/10.1214/15-AOS1337. 3, 10, 11, 12

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020). "Robust inference with knockoffs." *The Annals of Statistics*, 48(3): 1409–1431. MR4124328. doi: https://doi.org/10.1214/19-AOS1852. 23

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870–897. URL https://doi.org/10.1214/009053604000000238 MR2065192. doi: https://doi.org/10.1214/009053604000000238. 10

Barton, N. H., Etheridge, A. M., and Véber, A. (2017). "The infinitesimal model: Definition, derivation, and implications." *Theoretical Population Biology*, 118: 50–73. 2, 5, 22

Benjamini, Y. and Heller, R. (2007). "False discovery rates for spatial signals." *Journal of the American Statistical Association*, 102(480): 1272–1281. MR2412549. doi: https://doi.org/10.1198/016214507000000941. 8, 13, 23, 24

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300. MR1325392. 1, 2, 11

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency." *Annals of Statistics*, 1165–1188. MR1869245. doi: https://doi.org/10.1214/aos/1013699998. 11

Besag, J. (1975). "Statistical analysis of non-lattice data." *Journal of the Royal Statistical Society Series D: The Statistician*, 24(3): 179–195. 15

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., and Roux, F. (2010). "Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature." *PLoS Genetics*, 6(5): e1000940. 4, 19, 20, 22

Brodie, A., Azaria, J. R., and Ofran, Y. (2016). "How far from the SNP may the causative genes be?" *Nucleic Acids Research*, 44(13): 6046–6054. 2, 22, 23

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). "Controlling the rate of GWAS false discoveries." *Genetics*, 205(1): 61–75. 1, 2, 13, 23, 24

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., and Bühlmann, P. (2016). "Assessing statistical significance in multivariable genome wide association analysis." *Bioinformatics*, 32(13): 1990–2000. 2, 3, 22

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3): 551–577. MR3798878. doi: https://doi.org/10.1111/rssb.12265. 3, 11, 12, 13, 14, 22, 23, 24

Casella, G. and Robert, C. P. (1996). "Rao-Blackwellisation of sampling schemes." *Biometrika*, 83(1): 81–94. MR1399157. doi: https://doi.org/10.1093/biomet/83.1.81. 14

Chanut-Delalande, H., Fernandes, I., Roch, F., Payre, F., and Plaza, S. (2006). "Shavenbaby couples patterning to epidermal cell shape control." *PLoS Biology*, 4(9): e290. 8

Cooper, T. F., Ostrowski, E. A., and Travisano, M. (2007). "A negative relationship between mutation pleiotropy and fitness effect in yeast." *Evolution*, 61(6): 1495–1499. 8

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023). "False discovery rate control via data splitting." *Journal of the American Statistical Association*, 118(544): 2503–2520. MR4681600. doi: https://doi.org/10.1080/01621459.2022.2060113. 23

Dai, R. and Barber, R. (2016). "The knockoff filter for FDR control in group-sparse and multitask regression." In *International Conference on Machine Learning*, 1851–1859. PMLR. 23, 24

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). "A linear complexity phasing method for thousands of genomes." *Nature Methods*, 9(2): 179–181. 23

Donnelly, P. (2008). "Progress and challenges in genome-wide association studies in humans." *Nature*, 456(7223): 728–731. 2

ffrench Constant, R. H., Pittendrigh, B., Vaughan, A., and Anthony, N. (1998). "Why are there so few resistance–associated mutations in insecticide target genes?" *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1376): 1685–1693. 8

Fridley, B. L. (2009). "Bayesian variable and model selection methods for genetic association studies." *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1): 27–37. 3, 5, 22

Frommlet, F., Ruhaltinger, F., Twaróg, P., and Bogdan, M. (2012). "Modified versions of Bayesian Information Criterion for genome-wide association studies." *Computational Statistics & Data Analysis*, 56(5): 1038–1051. MR2897552. doi: https://doi.org/10.1016/j.csda.2011.05.005. 3

Gazzarrini, S., Tsuchiya, Y., Lumba, S., Okamoto, M., and McCourt, P. (2004). "The transcription factor FUSCA3 controls developmental timing in *Arabidopsis* through the hormones gibberellin and abscisic acid." *Developmental Cell*, 7(3): 373–385. 20

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85(410): 398–409. MR1141740. 4

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 5

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 7, 10

Griffin, J. E., Łatuszyński, K., and Steel, M. F. (2021). "In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p." *Biometrika*, 108(1): 53–69. MR4226189. doi: https://doi.org/10.1093/biomet/asaa055. 14

Gu, J. and Yin, G. (2021). "Bayesian Knockoff Filter." *arXiv preprint arXiv:2102.05223*. MR4294543. doi: https://doi.org/10.1111/rssb.12430. 3, 23

Guan, Y. and Stephens, M. (2011). "Bayesian variable selection regression for genome-wide association studies and other large-scale problems." *Annals of Applied Statistics*, 5(3): 1780–1815. MR2884922. doi: https://doi.org/10.1214/11-AOAS455. 2, 3, 4, 5, 6, 7, 9, 14, 21

He, Z., Chu, B., Yang, J., Gu, J., Chen, Z., Liu, L., Morrison, T., Belloy, M. E., Qi, X., Hejazi, N., Mathur, M., Le Guen, Y., Tang, H., Hastie, T., Ionita-laza, I., Sabatti, C., and Candès, E. (2024). "Beyond guilty by association at scale: Searching for causal variants on the basis of genome-wide summary statistics." *bioRxiv*. 2, 13, 24

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017). "The basis function approach for modeling autocorrelation in ecological data." *Ecology*, 98(3): 632–646. 10

Hooten, M. B. and Hefley, T. J. (2019). *Bringing Bayesian Models to Life*. CRC Press. 10

Hooten, M. B. and Hobbs, N. T. (2015). "A guide to Bayesian model selection for ecologists." *Ecological Monographs*, 85(1): 3–28. 21

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1): 139–159. MR3008275. doi: https://doi.org/10.1111/j.1467-9868.2012.01041.x. 10

Jorde, L. (2000). "Linkage disequilibrium and the search for complex disease genes." *Genome Research*, 10(10): 1435–1444. 8, 13

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). "Variance component model to account for sample

structure in genome-wide association studies." *Nature Genetics*, 42(4): 348–354. 2, 6, 15, 22

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. (2007). "Recombination and linkage disequilibrium in Arabidopsis thaliana." *Nature Genetics*, 39(9): 1151–1155. 19

Lee, E., Ibrahim, J. G., and Zhu, H. (2023). "Bayesian bi-level variable selection for genome-wide survival study." *Genomics & Informatics*, 21(3): e28. 2, 24

Li, F. and Zhang, N. R. (2010). "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics." *Journal of the American Statistical Association*, 105(491): 1202–1214. MR2752615. doi: https://doi.org/10.1198/jasa.2010.tm08177. 2, 3, 9, 21, 24

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). "The Bayesian lasso for genome-wide association studies." *Bioinformatics*, 27(4): 516–523. MR3172942. doi: https://doi.org/10.2174/1875036201307010027. 3, 5, 6, 21

Lippitt, W., Carlson, N. E., Arbet, J., Fingerlin, T. E., Maier, L. A., and Kechris, K. (2024). "Limitations of clustering with PCA and correlated noise." *Journal of Statistical Computation and Simulation*, 94(10): 2291–2319. MR4769269. doi: https://doi.org/10.1080/00949655.2024.2329976. 13

Lu, C., Tzovaras, B. G., and Gough, J. (2021). "A survey of direct-to-consumer genotype data, and quality control tool (*GenomePrep*) for research." *Computational and Structural Biotechnology Journal*, 19: 3747–3754. 2

Lu, Z.-H., Zhu, H., Knickmeyer, R. C., Sullivan, P. F., Williams, S. N., and Zou, F. (2015). "Multiple SNP set analysis for genome-wide association studies through Bayesian latent variable selection." *Genetic Epidemiology*, 39(8): 664–677. 3, 5, 6, 13, 21, 23, 24

McGregor, A. P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D. G., Payre, F., and Stern, D. L. (2007). "Morphological evolution through multiple cis-regulatory mutations at a single gene." *Nature*, 448(7153): 587–590. 7

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 5, 7

Niu, H.-M., Yang, P., Chen, H.-H., Hao, R.-H., Dong, S.-S., Yao, S., Chen, X.-F., Yan, H., Zhang, Y.-J., Chen, Y.-X., et al. (2019). "Comprehensive functional annotation of susceptibility SNPs prioritized 10 genes for schizophrenia." *Translational Psychiatry*, 9(1): 56. 22

O'Hara, R. B. and Sillanpää, M. J. (2009). "A review of Bayesian variable selection methods: what, how and which." *Bayesian Analysis*, 4(1): 85–117. MR2486240. doi: https://doi.org/10.1214/09-BA403. 10

Posey, J. E. (2019). "Genome sequencing and implications for rare disorders." *Orphanet Journal of Rare Diseases*, 14(1): 153. 7

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." *Nature Genetics*, 38(8): 904–909. 2, 3, 6

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). "New approaches to population stratification in genome-wide association studies." *Nature Reviews Genetics*, 11(7): 459–463. 2, 3, 6

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). "Association mapping in structured populations." *The American Journal of Human Genetics*, 67(1): 170–181. 3

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American Journal of Human Genetics*, 81(3): 559–575. 13

Reich, B. J. and Hodges, J. S. (2008). "Identification of the variance components in the general two-variance linear model." *Journal of Statistical Planning and Inference*, 138(6): 1592–1604. URL https://www.sciencedirect.com/science/article/pii/S037837580700314X MR2427291. doi: https://doi.org/10.1016/j.jspi.2007.05.046. 10

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). "Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models." *Biometrics*, 62(4): 1197–1206. MR2307445. doi: https://doi.org/10.1111/j.1541-0420.2006.00617.x. 6, 7

Ren, Z., Wei, Y., and Candès, E. (2023). "Derandomizing knockoffs." *Journal of the American Statistical Association*, 118(542): 948–958. MR4595468. doi: https://doi.org/10.1080/01621459.2021.1962720. 23

Risch, N. and Merikangas, K. (1996). "The future of genetic studies of complex human diseases." *Science*, 273(5281): 1516–1517. 2

Rosenblatt, J. D., Ritov, Y., and Goeman, J. J. (2019). "Discussion of 'Gene hunting with hidden Markov model knockoffs'." *Biometrika*, 106(1): 29–33. MR3912381. doi: https://doi.org/10.1093/biomet/asy062. 13

Saeys, Y., Inza, I., and Larranaga, P. (2007). "A review of feature selection techniques in bioinformatics." *Bioinformatics*, 23(19): 2507–2517. 1

Schaid, D. J., Chen, W., and Larson, N. B. (2018). "From genome-wide associations to candidate causal variants by statistical fine-mapping." *Nature Reviews Genetics*, 19(8): 491–504. 2

Scheet, P. and Stephens, M. (2006). "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." *The American Journal of Human Genetics*, 78(4): 629–644. 13, 23

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf,

B., Weigel, D., and Lohmann, J. U. (2005). "A gene expression map of *Arabidopsis thaliana* development." *Nature Genetics*, 37(5): 501–506. 20

Selinski, S. and Ickstadt, K. (2008). "Cluster analysis of genetic and epidemiological data in molecular epidemiology." *Journal of Toxicology and Environmental Health*, 71(11-12): 835–844. 13

Sesia, M., Bates, S., Candès, E., Marchini, J., and Sabatti, C. (2021). "False discovery rate control in genome-wide association studies with population structure." *Proceedings of the National Academy of Sciences*, 118(40): e2105841118. 1, 2, 3, 11, 13, 14, 21, 23

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2020). "Multi-resolution localization of causal variants across the genome." *Nature Communications*, 11(1): 1093. 3, 11, 13, 24

Sesia, M., Sabatti, C., and Candès, E. J. (2019). "Gene hunting with hidden Markov model knockoffs." *Biometrika*, 106(1): 1–18. MR3912377. doi: https://doi.org/10.1093/biomet/asy033. 2, 3, 6, 9, 11, 13, 14, 15, 22, 23, 24

Shindo, C., Aranzana, M. J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., and Dean, C. (2005). "Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*." *Plant Physiology*, 138(2): 1163–1173. 4, 8

Shirsekar, G., Devos, J., Latorre, S. M., Blaha, A., Queiroz Dias, M., González Hernando, A., Lundberg, D. S., Burbano, H. A., Fenster, C. B., and Weigel, D. (2021). "Multiple sources of introduction of North American *Arabidopsis thaliana* from across Eurasia." *Molecular Biology and Evolution*, 38(12): 5328–5344. 4

Spisák, S., Lawrenson, K., Fu, Y., Csabai, I., Cottman, R. T., Seo, J.-H., Haiman, C., Han, Y., Lenci, R., Li, Q., et al. (2015). "CAUSEL: An epigenome-and genome-editing pipeline for establishing function of noncoding GWAS variants." *Nature Medicine*, 21(11): 1357–1363. 2

Stephens, M. and Balding, D. J. (2009). "Bayesian statistical methods for genetic association studies." *Nature Reviews Genetics*, 10(10): 681–690. 3, 22

Stern, D. L. and Orgogozo, V. (2009). "Is genetic evolution predictable?" *Science*, 323(5915): 746–751. 3, 7, 8, 23

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes." *The Annals of Applied Statistics*, 5(3). MR2884929. doi: https://doi.org/10.1214/11-AOAS463. 3, 21

Storey, J. D. and Tibshirani, R. (2003). "Statistical significance for genomewide studies." *Proceedings of the National Academy of Sciences*, 100(16): 9440–9445. MR1994856. doi: https://doi.org/10.1073/pnas.1530509100. 2

Sul, J. H., Martin, L. S., and Eskin, E. (2018). "Population structure in genetic studies: Confounding factors and mixed models." *PLoS Genetics*, 14(12): e1007309. 2, 6

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288. MR1379242. 15

Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., and et. al. (2018). "The AraGWAS Catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog." *Nucleic Acids Research*, 46(D1): D1150–D1156. 14

Van Ee, J. J., Gamba, D., Lasky, J. R., Vahsen, M. L., and Hooten, M. B. (2025). "Supplement to "Spatial knockoff Bayesian variable selection in genome-wide association studies"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1556SUPP. 10

Van Straalen, N. M. and Roelofs, D. (2011). *An Introduction to Ecological Genomics*. Oxford University Press. 7

Vannucci, M., Stingo, F. C., and Berzuini, C. (2010). "Bayesian models for variable selection that incorporate biological information." *Bayesian Statistics*, 9: 1–20. MR3204022. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0022. 2, 3

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., and Fortin, M.-J. (2018). "Spatial autoregressive models for statistical inference from ecological data." *Ecological Monographs*, 88(1): 36–59. 10

Visscher, P. M., Goddard, M. E., Derks, E. M., and Wray, N. R. (2012). "Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses." *Molecular Psychiatry*, 17(5): 474–485. 2, 5

Wang, C., Kao, W.-H., and Hsiao, C. K. (2015). "Using Hamming distance as information for SNP-sets clustering and testing in disease association studies." *PloS One*, 10(8): e0135918. 13

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). "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. MR4176343. 2

Wang, K., Li, M., and Hakonarson, H. (2010). "Analysing biological pathways in genome-wide association studies." *Nature Reviews Genetics*, 11(12): 843–854. 2, 7, 13, 23

Wang, R. and Ramdas, A. (2022). "False discovery rate control with e-values." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3): 822–852. MR4460577. 2, 23

Weinwurm, S., Sölkner, J., and Waldmann, P. (2013). "The effect of linkage disequilibrium on Bayesian genome-wide association methods." *Journal of Biometrics & Biostatistics*, 4(5): 180. 13

Xing, X., Zhao, Z., and Liu, J. S. (2023). "Controlling false discovery rate using Gaussian mirrors." *Journal of the American Statistical Association*, 118(541): 222–241. MR4571118. doi: https://doi.org/10.1080/01621459.2021.1923510. 23

Yu, J., Pressoir, G., Briggs, H., Vroh, I., Yamasaki, M., Doebley, J., McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., and Buckler, E. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nature Genetics*, 38(2): 203–208. 2, 6

Zan, Y. and Carlborg, Ö. (2018). "A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected *Arabidopsis thaliana*." *Molecular Ecology Resources*, 18(4): 798–808. 4, 8, 22

Zhang, X., Xue, F., Liu, H., Zhu, D., Peng, B., Wiemels, J. L., and Yang, X. (2014). "Integrative Bayesian variable selection with gene-based informative priors for genome-wide association studies." *BMC Genetics*, 15: 1–11. 2, 3, 24

Zhou, X. and Stephens, M. (2012). "Genome-wide efficient mixed-model analysis for association studies." *Nature Genetics*, 44(7): 821–824. MR3172942. doi: https://doi.org/10.2174/1875036201307010027. 2, 6

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." *Proceedings of the National Academy of Sciences*, 109(4): 1193–1198. 4, 5, 15