Xuan Cao¹ / Lili Ding² / Tesfaye B. Mersha³

Joint variable selection and network modeling for detecting eQTLs

¹ Division of Statistics and Data Science, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221, USA, E-mail: xuan.cao@uc.edu. <https://orcid.org/0000-0002-6859-0030>.

² Division of Biostatistics and Epidemiology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH 45229, USA

³ Division of Asthma Research, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH 45229, USA

Abstract:

In this study, we conduct a comparison of three most recent statistical methods for joint variable selection and covariance estimation with application of detecting expression quantitative trait loci (eQTL) and gene network estimation, and introduce a new hierarchical Bayesian method to be included in the comparison. Unlike the traditional univariate regression approach in eQTL, all four methods correlate phenotypes and genotypes by multivariate regression models that incorporate the dependence information among phenotypes, and use Bayesian multiplicity adjustment to avoid multiple testing burdens raised by traditional multiple testing correction methods. We presented the performance of three methods (MSSL – Multivariate Spike and Slab Lasso, SSUR – Sparse Seemingly Unrelated Bayesian Regression, and OBFBF – Objective Bayes Fractional Bayes Factor), along with the proposed, JDAG (Joint estimation via a Gaussian Directed Acyclic Graph model) method through simulation experiments, and publicly available HapMap real data, taking asthma as an example. Compared with existing methods, JDAG identified networks with higher sensitivity and specificity under row-wise sparse settings. JDAG requires less execution in small-to-moderate dimensions, but is not currently applicable to high dimensional data. The eQTL analysis in asthma data showed a number of known gene regulations such as STARD3, IKZF3 and PGAP3, all reported in asthma studies. The code of the proposed method is freely available at GitHub (<https://github.com/xuan-cao/Joint-estimation-for-eQTL>).

Keywords: eQTL, join analysis, network modeling, variable selection

DOI: 10.1515/sagmb-2019-0032

1 Introduction

In the last few years, high-throughput technologies have made it possible for genetic studies to move from variant cataloging to the analysis of genome function, especially in the context of multi-omics data including transcriptomics. One of the most commonly explored links is genetic variation in the context of gene expression referred as expression quantitative trait loci (eQTL). eQTL assesses the effect of genetic variation on gene expression and improves our understanding on how genetic variation leads to phenotypic variation including its contribution to human disease. eQTL studies have offered promise not just for the characterization of functional sequence variation but also for the understanding of basic processes of gene regulation and interpretation of genome-wide association studies.

The most commonly used statistical framework to eQTL analysis is a regression model of phenotype, e.g. gene expression on genotype. In regression model, univariate analysis is performed for each phenotype-genotype pair, needing post-hoc adjustment for multiple comparisons and ignoring any correlations within genotypes and phenotypes. However, genes and their regulators are organized into correlated networks, thus do not function in isolation. Simple univariate analysis approach may miss important genetic information and makes associations harder to be detected partly due to multiple testing correction. Hence, univariate approaches are unlikely to be the best strategy, especially in the context of complex disease genetic studies where high levels of correlation happen as a rule rather than exception (Ackermann, Sikora-Wohlfeld & Beyer, 2013; GTEx Consortium, 2015).

Since gene expression is downstream of genetic variants in the causal pathway and given the high-dimension and high correlation among gene expression data, there are two important aspects of eQTL, (a)

Xuan Cao is the corresponding author.

© 2020 Walter de Gruyter GmbH, Berlin/Boston.

to find associations between a set of genetic variants and gene expression (variable selection), where both genotype and gene expression data could be high-dimensional; and (b) to understand the dependence among genes through expression levels while accounting for the genetic variation information (covariance/network estimation).

Variable selection and covariance estimation have individually received great attention in the statistical literature, and multiple methods were proposed (Banerjee & Ghosal, 2015; Bhadra & Mallick, 2013; Yang, Wainwright & Jordan, 2016). In high-dimensional multi-omics datasets where the number of variables is much higher than the number of samples, an important problem is to select a subset of predictor variables, usually sparse, that significantly contribute to given responses. Furthermore, the covariance matrix is the most fundamental object that quantifies relationships between the given multivariate responses (Cao, Khare & Ghosh 2019a; 2019b). In recent years, joint variable selection and covariance estimation in high-dimensional settings has generated a lot of interest (Banterle et al., 2018; Consonni, Rocca & Peluso, 2017; Deshpande, Ročková & George, 2019). Examples include the multi-omics data such as eQTL and metabolomics analysis, where gene expression and metabolomics data are simultaneously available and analyzed (see Lee et al., 2018; Nica & Dermitzakis, 2013; Rockman & Kruglyak, 2006).

There are two types of hierarchical Bayesian approaches for joint variable selection and covariance/network estimation. The first type of approaches assumes the coefficient matrix linking SNP to gene expression to be row-wise sparse, meaning that a certain SNP could either have impact or not have impact on all the targeted genes. Under this assumption, many methods have been proposed to address the joint selection problem including Bhadra and Mallick (2013) and Consonni, Rocca, and Peluso (2017). In Bhadra and Mallick (2013), the authors induce sparsity both in the rows of the coefficient matrix, and in the covariance matrix through an undirected decomposable graph. The authors then explore both the variable and graph space via a Markov chain Monte Carlo (MCMC) algorithm. In Consonni, Rocca, and Peluso (2017), similar approach on selecting the active rows in the coefficient matrix is taken. In terms of network estimation, the authors perform an objective Bayesian model selection and impose a fractional prior on the inverse covariance matrix, then give a conjugate analysis and provide closed-form expressions for the marginal likelihood of the graph. The authors adopt similar collapsed Metropolis-Hastings algorithm as in Bhadra and Mallick (2013), and show significant improvement compared with Bhadra and Mallick (2013) in joint selection performance. We refer the method in Consonni, Rocca, and Peluso (2017) as objective Bayes fractional Bayes factor (“OBFBF”) for easier reference in the sequel. Under the same assumption (row-wise sparse), we propose a hierarchical multivariate regression model with a spike and slab Cholesky prior introduced in Cao, Khare, and Ghosh (2019a) on the covariance matrix, and independent Bernoulli priors for each edge, and for each active row index of the coefficient matrix. We refer to our proposed method as “JDAG” (joint estimation via a Gaussian directed acyclic graph (DAG) model).

The novelty of our proposed method comes from the following aspects. In this work, we significantly extended previous work including Consonni, Rocca, and Peluso (2017) and Bhadra and Mallick (2013) by (a) applying the spike and slab Cholesky prior and performing conjugacy analysis in the covariate-adjusted graphical model setting; (b) ensuring the JDAG selection is consistent under the covariate-adjusted graphical model setting, which was not theoretically guaranteed in Consonni, Rocca, and Peluso (2017) and Bhadra and Mallick (2013); (c) applying the methodology to the asthma related real data analysis of eQTL and gene network.

The second type of approaches assumes the coefficient matrix linking gene expression to SNP information to be entry-wise sparse, meaning that each individual SNP could either have impact or not have impact on individual genes. Under this assumption, two most recent methods have been proposed. Deshpande, Ročková, and George (2019) tackle the problem of simultaneous variable and covariance selection with the multivariate spike and slab Lasso prior placed on both individual genotype-phenotype associations and the entries of inverse covariance matrix, and then utilize an Expectation/Conditional Maximization (ECM) algorithm for fast dynamic conditional posterior exploration. For the rest of paper, we refer to this method as “MSSL” (Multivariate Spike and Slab Lasso). Banterle et al. (2018) place a spike and slab prior on each individual genotype-phenotype association and a standard inverse Wishart prior on the covariance matrix, and use the Evolutionary Stochastic Search (Bottolo & Richardson, 2010) to explore the sample space. Since this method is based on a sparse seemingly unrelated Bayesian regression model, we refer to this method as “SSUR” (Sparse Seemingly Unrelated Bayesian Regression). The computation cost for SSUR is extremely high, while MSSL requires relatively short time for execution. Compared with JDAG which is designed for small to moderate datasets, SSUR and MSSL are shown to be working under larger dimensions.

The manuscript is organized as follows. First, we elucidate the different methods in Section 2. Second, we show simulation experimental settings. Third, we conduct the experiment in real data using asthma as an example. Fourth, we present our results in the simulated and real data settings and evaluated the performance of MSSL, SSUR, and OBFBF against JDAG, with respect to several different measures of accuracy including precision, sensitivity and specificity, and the Receiver Operating Characteristic (ROC) curve.

2 Materials and methods

2.1 Notation and general joint model setup

In this section, we provide an overview of current Bayesian model formulation in analyzing eQTL associations. We start by reviewing the matrix normal distribution and clarifying some required notations. Consider a setting where we have n independent observations on q variables, which can be gene expression data on q genes. In particular, let $Y_1, Y_2, \dots, Y_n$ be observed data arranged in $n \times q$ matrix form:

$$Y = \begin{pmatrix} Y_1^T \\ \vdots \\ Y_n^T \end{pmatrix} = (y_1, \dots, y_q), \quad (1)$$

in which $Y_i = (y_{i1}, \dots, y_{iq})^T$ represents the observed values on the q variables for the i -th subject and $y_j = (y_{1j}, \dots, y_{nj})^T$ is the observed values on the j -th variable for n subjects. Let X denote an $n \times p$ design matrix of covariates containing the row vectors x_i ($i = 1, 2, \dots, n$), where each x_i contains the genotype information on p SNP locations for the i -th subject.

We further assume each observation Y_i to be normally and independently distributed with mean $B^T x_i$ and covariance matrix Σ , where B is a $p \times q$ coefficient matrix, in which each entry B_{ij} can be viewed as the coefficient that links the i -th SNP to the j -th gene, and $\Sigma = (LD^{-1}L^T)^{-1}$ is a $q \times q$ positive definite matrix, where (L, D) is the modified Cholesky decomposition of $\Omega = \Sigma^{-1}$. Equivalently, we model the distribution of Y to be matrix normal, conditional on B and (L, D) , given by $Y \sim MN_{n \times q}(XB, I_n, \Sigma)$, with mean matrix XB , row covariance matrix I_n and column covariance matrix $\Sigma = (LD^{-1}L^T)^{-1}$. Recall that our goal is to jointly select the nonzero values in B , and to estimate the sparsity pattern in Σ^{-1} , which can be uniquely encoded in appropriate graphs. We demonstrate the model setup via the following workflow (Figure 1).

Figure 1: A schematic of how models are formulated for joint eQTL analysis and network estimation. Based on the genetic variation matrix X and gene expression matrix Y to jointly select the nonzero values in B (containing the information for SNP-gene association), and to estimate the sparsity pattern in Σ^{-1} , which can be uniquely encoded in appropriate graphs as edges (gene network).

2.2 Markov chain Monte Carlo (MCMC) based method

In this section, we present our proposed Bayesian hierarchical model for the joint selection (JDAG), and review the OBFBB approach in Consonni, Rocca, and Peluso (2017) and the SSUR approach in Banterle et al. (2018). In particular, all three methods are based on MCMC algorithms, and our method along with OBFBB focus more on selecting SNPs impacting expression of all genes, while the SSUR method could be used to select significant individual SNP-gene associations.

2.2.1 Joint estimation via a Gaussian DAG model (JDAG)

2.2.1.1 Model specification

In this setting with Y obeying a matrix normal distribution, recall that B is a $p \times q$ matrix with p predictors. Suppose we are merely interested in the active rows of B , i.e. significant SNPs that are associated with all gene expression data. In particular, we assume only a sparse subset of size p_z is active in the model. Let $z = (z_1, z_2, \dots, z_p)$ be a vector of indicators, where $z_j = 1$ ($j = 1, 2, \dots, p$) if and only if the j -th predictor is present in the model. Thus, $p_z = \sum_{j=1}^p z_j$. Denote X_z as the submatrix formed from the columns of X corresponding to the active components in the indicator vector z and B_z as the submatrix formed from the rows of B corresponding to the active parts of z . Our goal in this subsection is to jointly perform SNP (variable) and gene network (DAG) selection, whose sparsity pattern could be uniquely encoded in Cholesky factor of the inverse covariance matrix Σ^{-1} , i.e. to correctly identify all the active variables and all the non-zero elements in the modified Cholesky parameter of Σ^{-1} (Cao, Khare & Ghosh 2019a; 2019b). To facilitate this purpose, we consider the following hierarchical model with the spike and slab Cholesky prior on the Cholesky factors introduced in Cao, Khare, and Ghosh (2019a), for $1 \leq j < k \leq q$,

$$Y|B_z, (L, D), \mathcal{D} \sim MN_{n \times q}(X_z B_z, I_n, (LD^{-1}L^T)^{-1}), \quad (2)$$

$$B_z|z, (L, D), \mathcal{D} \sim MN_{p_z \times q}(0, \tau I_{p_z}, (LD^{-1}L^T)^{-1}), \quad (3)$$

$$L_{kj}|d_j, \gamma_{kj} \sim \gamma_{kj}N(0, \tau^2 d_j) + (1 - \gamma_{kj})\delta_0(L_{kj}), \quad (4)$$

$$d_j \sim \text{Inverse-Gamma}(\lambda_1, \lambda_2), \quad (5)$$

$$\pi(z) = \prod_{i=1}^p \eta_1^{z_i} (1 - \eta_1)^{1-z_i}, \quad (6)$$

$$\pi(\mathcal{D}) = \prod_{j=1}^{q-1} \eta_2^{\sum_{k=j+1}^q \gamma_{kj}} (1 - \eta_2)^{q-j - \sum_{k=j+1}^q \gamma_{kj}}, \quad (7)$$

where $\gamma_{kj} = \mathbb{I}\{(k, j) \in E(\mathcal{D})\}$, $1 \leq j < k \leq q$ represents the edge indicator, and $E(\mathcal{D})$ is the edge set for DAG $\mathcal{D}$. In particular, the prior density for DAGs above in (7) corresponds to an Erdos-Renyi type of distribution on the space of DAGs, where each directed edge is present with probability η_2 independently of the other edges. Here, similar to Banerjee and Ghosal (2015) and Cao, Khare, and Ghosh (2019b), we are essentially letting γ_{kj} , $1 \leq j < k \leq q$ be independently and identically distributed Bernoulli (η_2) random variables.

In (3), by imposing a spike and slab type of prior on the coefficient matrix B , we are also performing a variable selection on all the active rows (SNP locations) in B as in Bhadra and Mallick (2013) and Consonni, Rocca, and Peluso (2017). The prior (6) for the variable indicator z on the space of 2^p variable indicators is equivalent to letting each z_i obey an independent Bernoulli distribution with probability η_1 . The proposed model differs from Bhadra and Mallick (2013) in two aspects. First, Bhadra and Mallick (2013) impose the hyper-inverse Wishart (HIW) distribution on the covariance matrix. Note that HIW priors are only defined for decomposable graphs, while the spike and slab Cholesky priors are defined on the entire Cholesky space and could be used for sparsity selection in these Cholesky factors. Second, under the identical assumptions in Cao, Khare, and Ghosh (2019a), the hierarchical model (2)–(7) yields the DAG selection consistency given the data and the true variable indicator Z_0 (Cao, Khare & Ghosh, 2019a). Whether the consistency will be theoretically guaranteed in Consonni, Rocca, and Peluso (2017) and Bhadra and Mallick (2013) is yet to be investigated.

2.2.1.2 Computational strategy

Now by the expression and conjugacy property of the family of spike and slab Cholesky priors, the hierarchical model (2)–(7) and Bayes' rule, the conditional distribution of Y given the indicator z and DAG D is given by

$$\pi(Y|z, \mathcal{D}) = \frac{1}{(2\pi)^{nq/2}} \times \frac{\gamma(U + Y^T(I_n + \tau(X_z X_z^T))^{-1}Y, n + \alpha)}{|(I_n + \tau(X_z X_z^T))|^{\frac{q}{2}} \gamma(U, \alpha)}, \quad (8)$$

where $\gamma(\cdot, \cdot)$ is the normalized constant for the spike and slab Cholesky distribution, and we refer readers to equation (16) in Cao, Khare, and Ghosh (2019a) for more details. Following from (8), after integrating out the modified Cholesky parameter (L, D) , the conditional density $\pi(Y|z, \mathcal{D})$ is available in closed form. Our simulation procedure is a collapsed gibbs sampler and the joint search for the active set of q predictors and the sparse DAG can be accomplished by simply generating the chain between z and $\mathcal{D}$. Therefore, we can evaluate the parameter space with substantial computational gains. Similar procedures are also adopted in Bhadra and Mallick (2013) and Consonni, Rocca, and Peluso (2017).

The sampler iteratively generates sequences of z and $\mathcal{D}$ from their conditional posteriors via Metropolis–Hastings algorithm using random addition/deletion of variables (or edges in DAG, respectively) for z and $\mathcal{D}$. In our simulation studies, we let $\eta_1 = 1/p$ and $\eta_2 = 1/q^2$. The MCMC procedure for z conditional on $\mathcal{D}$ and Y is given by,

1. Given the current state z_{curr} , sample the next state $z_{\text{cand}} \sim q(\cdot, z_{\text{curr}})$;
2. Compute the acceptance probability

$$p_a = \min \left\{ 1, \frac{\pi(Y|z_{\text{cand}}, \mathcal{D}) \pi(z_{\text{cand}}) q(z_{\text{cand}}|z_{\text{curr}})}{\pi(Y|z_{\text{curr}}, \mathcal{D}) \pi(z_{\text{curr}}) q(z_{\text{curr}}|z_{\text{cand}})} \right\},$$

and change the current state from z_{curr} to z_{cand} with probability p_a .

We chose the kernel $q(\cdot|z_{\text{curr}})$ by randomly changing a nonzero component in z_{curr} to 0 with probability 0.5 or by randomly setting a zero component to 1 with probability 0.5.

The MCMC procedure for $\mathcal{D}$ conditional on z and Y is similar to that of z and given by,

1. Given the current DAG $\mathcal{D}_{\text{curr}}$, sample the next DAG $\mathcal{D}_{\text{cand}} \sim q(\cdot, \mathcal{D}_{\text{curr}})$;
2. Compute the acceptance probability

$$p_b = \min \left\{ 1, \frac{\pi(Y|\mathcal{D}_{\text{cand}}, z) \pi(\mathcal{D}_{\text{cand}}) q(\mathcal{D}_{\text{cand}}|\mathcal{D}_{\text{curr}})}{\pi(Y|\mathcal{D}_{\text{curr}}, z) \pi(\mathcal{D}_{\text{curr}}) q(\mathcal{D}_{\text{curr}}|\mathcal{D}_{\text{cand}})} \right\},$$

and change the current state from $\mathcal{D}_{\text{curr}}$ to $\mathcal{D}_{\text{cand}}$ with probability p_b .

Similarly, we chose the kernel $q(\cdot|\mathcal{D}_{\text{curr}})$ by randomly changing a nonzero entry in the adjacency matrix of $\mathcal{D}_{\text{curr}}$ to 0 with probability 0.5 or by randomly setting a zero entry to 1 with probability 0.5.

We ran the Metropolis–Hastings algorithm to conduct the posterior inference of joint estimation of z and $\mathcal{D}$. Each MCMC chain ran for an iteration of 15,000 times with a burnin period of 7000, which gives us 8000 posterior samples. We constructed the final estimates by collecting the SNPs and edges with inclusion probabilities greater than 0.5. In the remaining of the paper, we refer to this method as “JDAG”.

2.2.2 Objective Bayes Fractional Bayes Factor based estimation (OBFBF)

In Consonni, Rocca, and Peluso (2017), similar approaches on selecting the active rows of B are taken. In the aspect of graph selection, the authors perform an objective Bayesian model selection and impose the following fractional prior on (B, Ω) with hyperparameters a_D, n_0 ,

$$\pi(B_z, \Omega) \propto |\Omega|^{\frac{a_D + n_0}{2}} \exp \left\{ -\frac{n_0}{2} \text{tr}(\Omega \{(B_z - \hat{B}_z)^T \tilde{C}_z (B_z - \hat{B}_z) + \tilde{R}_z\}) \right\},$$

where $\hat{B}_z = (X_z^T X_z)^{-1} X_z^T Y$, $\tilde{C}_z = n^{-1} X_z^T X_z$ and $\tilde{R}_z = n^{-1} (Y - X_z \hat{B}_z)^T (Y - X_z \hat{B}_z)$. The remaining prior specifications for both the variable and graph indicators are exactly the same as in (6)–(7). The authors then carry out a conjugate analysis and provide closed-form expressions for the marginal likelihood of any DAG, not just for perfect DAG only (an equivalent class to decomposable graphs), which expands the model scope for using the HIW prior in Bhadra and Mallick (2013). The authors also adopt similar collapsed Metropolis–Hastings algorithm as in Bhadra and Mallick (2013) for exploring the model space.

2.2.3 Sparse seemingly unrelated Bayesian regression (SSUR)

The closed form in (8) is convenient in the sense that we can easily derive the collapsed gibbs sampler that facilitates the purpose of estimating significant SNPs and the inferred gene-network. However, the individual associations between each significant SNP and related gene expression are not simultaneously available. Banterle et al. (2018) further address this issue by placing a spike and slab prior not on each row of B , but on each individual entry of B . Banterle et al. (2018) further show that by reparameterizing the hyper-inverse Wishart prior induced over decomposable graphs, the hierarchical model now yields full posterior conditional distribution available in closed form. We refer the readers to Banterle et al. (2018) for more details.

For ease of computation, Banterle et al. (2018) use the Evolutionary Stochastic Search (ESS) proposed in Bottolo and Richardson (2010) that uses a particular form of Evolutionary Monte Carlo (EMC) introduced in Liang and Wong (2000). As indicated in Banterle et al. (2018), the usage of ESS helps the chain mix in the high dimensional binary space of both covariates selection and graph structure learning, where each chain has a single associated temperature that only tempers the likelihood. The binary variable selection indicators and the regression coefficients are updated jointly via Metropolis-Hastings. The sampler is implemented fully in C++ and freely available at https://github.com/mbant/Bayesian_SSUR. The proposed method is referred to as “SSUR”, following the nomenclature in Banterle et al. (2018).

2.3 Multivariate spike and slab Lasso (MSSL)

The ESS procedure developed in Banterle et al. (2018) is shown to have huge computational advantage compared to the Naive Gibbs sampler with simply the full conditional distributions. However, it can still take up to 2 hours when both p, q are around 100. To address the issues of computational burden in higher dimensions, the authors in Deshpande, Ročková, and George (2019) tackle the problem of simultaneous variable and covariance selection with the multivariate spike and slab Lasso type of priors placed on individual entries of B . The proposed method is referred to as “MSSL”, following the nomenclature in Deshpande, Ročková, and George (2019). The main difference between MSSL and other Bayesian methods is that rather than directly sampling from intractable posterior distribution with MCMC, MSSL maximizes the posterior density, seeking the posterior mode via an Expectation/Conditional Maximization (ECM) algorithm for fast dynamic conditional posterior exploration. Through simulation studies in Deshpande, Ročková, and George (2019), MSSL is shown to outperform regularization competitors with significantly less runtime.

3 Simulation studies

In the simulation studies, we aim to compare the performance of current state-of-the-art methods. We utilize the following five metrics with the simulation data: (1) Receiver Operating Characteristic (ROC) curve, (2) Precision, (3) Negative Predictive Value (NPV), (4) Sensitivity, and (5) Specificity. We consider the following three different sparsity settings of matrix B and Σ^{-1} : (1) B is row-wise sparse, meaning each row of B either entirely equaling or un-equaling to 0; (2) B is entry-wise sparse, meaning the sparsity of B could be impose on each individual entry of B , rather on the entire row; (3) Σ^{-1} has a banded structure following the simulation setting in Deshpande, Ročková, and George (2019), where Σ^{-1} is tri-diagonal.

3.1 B is row-wise sparse

For our first and second simulation studies regarding B , we set $(n, p, q) = (100, 300, 100)$, and then generate the simulated SNP data X using function `simulate SNPs` in the R package “`scrim`”, where a specified proportion of cases and controls is explained by specified set of SNP interactions. Next for the row-wise sparse setting, we random select 2% of the SNP locations and set the corresponding rows of B non-zero entries independently drawn uniformly from the interval $[-2, 2]$. Other remaining rows of B are set to be zero. We randomly choose 3% non-zero entries for the true Cholesky parameter L_0 , we construct a 100×100 lower triangular matrix with all the entries being zero and diagonal entries being 1. Then, we randomly choose 3% of the lower triangular entries and set them to be 1. We refer to this matrix as our true Cholesky factor L_0 . The matrix L_0 also gives us the true DAG $\mathcal{D}_0$. D_0 is taken to be the identity matrix for computational convenience. Finally, we generate Y from $MN_{n \times q}(XB, I_n, (L_0 D_0^{-1} L_0^T)^{-1})$.

We then compare the performance of the joint variable and DAG selection for JDAG, OBFBBF, SSUR and MSSL. For JDAG, OBFBBF and SSUR, the final model is constructed by collecting the predictors and edges with

inclusion probabilities greater than 0.5. For MSSL, the final model is constructed by collecting the non-zero entries in the coefficient matrix and the inverse covariance matrix as in Deshpande, Ročková, and George (2019). The model selection performance is then compared using several different measures such as precision, sensitivity and specificity (average over 20 independent repetitions). Precision represents the proportion of true edges/entries among all the edges/entries detected by the given procedure, and sensitivity measures the proportion of true edges/entries detected by the given procedure among all the true edges/entries from the true model. Specificity represents the proportion of non-true edges/entries successfully excluded by the given procedure among all the non-true edges/entries from the true model. Precision, negative predictive value (NPV), sensitivity and specificity are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \text{NPV} = \frac{TN}{TN + FN}, \text{Sensitivity} = \frac{TP}{TP + FN}, \text{Specificity} = \frac{TN}{TN + FP},$$

where TP, TN, FP and FN correspond to true positive, true negative, false positive and false negative, respectively. Note that one would also like the precision, sensitivity and specificity values to be as close to 1 as possible. To better illustrate the robustness of these three methods with respect to the cutoff (thresholding) values, we also plot the receiver operating characteristic (ROC) curves for evaluating the DAG selection performance.

The results for the row-wise sparse case are presented in Table 1, with the ROC curves provided in Figure 2. Based on the simulation results for the row-wise sparse case, JDAG method overall outperforms OBFBF, SSUR and MSSL. In particular, out of all the true entries in B , based on sensitivity, the SSUR method fails to capture more true active entries in B compared with other two methods MSSL and JDAG. The MSSL method captures more nonzero entries with a higher sensitivity, but in the same time falsely identifies more entries than SSUR and JDAG based on precision. In terms of DAG selection, JDAG could correctly identify more edges, while excluding more false edges compared with OBFBF, SSUR and MSSL. For both variable and DAG selection, OBFBF performs overall worse compared with the other three methods. As for the computational cost, SSUR and OBFBF require longer runtime compared with JDAG and MSSL. JDAG gave overall best joint selection performance without intensive computational cost.

Table 1: Joint selection performance under row-wise sparse B .

	Variable selection				Graph selection			
	JDAG	SSUR	MSSL	OBFBF	JDAG	SSUR	MSSL	OBFBF
Precision	0.79	0.37	0.14	0.33	0.99	0.75	0.71	0.03
NPV	0.99	0.77	1	0.86	0.99	0.93	0.97	0.99
Sensitivity	0.52	0.1	1	0.05	0.62	0.17	0.26	0.05
Specificity	1	0.98	0.98	0.98	1	1	1	0.99
Runtime (min)	7.21	37.48	3.21	48.42	7.21	37.48	3.21	48.42

Figure 2: ROC curves for DAG selection under row-wise sparse B for JDAG, MSSL, SSUR, and OBFBF. X-axis: true positive rate (sensitivity); Y-axis: false positive rate (1-specificity).

3.2 B is entry-wise sparse

For the simulation study when the individual level of sparsity is imposed on B , we first reconstruct B with 2% randomly placed non-zero entries independently drawn uniformly from the interval $[-2, 2]$, then generate X and Y by following the similar procedure as in the previous subsection. The results for the entry-wise sparse case are shown in Table 2, with the ROC curves provided in Figure 3. From the entry-wise sparse case simulation results, we observe that based on sensitivity, out of all the active entries, the JDAG method fails to capture true active entries in B compared with other two methods (SSUR and MSSL). The MSSL method captures more non-zero entries based on sensitivity, and in the same time excludes more non-true entries than MSSL and JDAG based on NPV and specificity. In terms of DAG selection, based on precision, out of all the edges detected, JDAG could correctly identify more edges compared with SSUR and MSSL. However, based on sensitivity, out of all the true edges, fewer edges are correctly detected by JDAG compared with SSUR and MSSL. For both variable and DAG selection, the performance of OBFBF is worse compared with all the other three methods. Overall, our results illustrate the need to choose appropriate method that matches the sparsity pattern that the true underlying B possesses to avoid model misspecification.

Table 2: Joint selection performance under entry-wise sparse B .

	Variable selection				Graph selection			
	JDAG	SSUR	MSSL	OBFBF	JDAG	SSUR	MSSL	OBFBF
Precision	0.35	0.51	0.37	0.12	0.62	0.47	0.54	0.04
NPV	0.71	0.76	1	0.89	0.94	0.98	0.97	0.99
Sensitivity	0.04	0.12	0.62	0.04	0.19	0.27	0.23	0.05
Specificity	0.98	0.99	0.99	0.98	0.99	0.99	0.99	0.99
Runtime (min)	7.56	43.29	3.15	52.65	7.56	43.29	3.15	52.65

Figure 3: ROC curves for DAG selection under entry-wise sparse B for JDAG, MSSL, SSUR, and OBFBF.

3.3 Σ^{-1} is banded

For the third simulation study when Σ^{-1} has a banded structure under higher dimensions $(n, p, q) = (100, 500, 200)$, we follow the simulation setting in Deshpande, Ročková, and George (2019) and Consonni, Rocca, and Peluso (2017). In particular, we first generate $\Sigma = (0.5^{|i-j|})_{1 \leq i, j \leq q}$. The resulting Ω will then be tri-diagonal. Of all 500 predictors, we randomly take 2% to be active. Next, we generate $B \sim MN(0, \tau I, \Sigma)$, $X_{ij} \sim N(10, 1)$, and $Y \sim MN(XB, I, \Sigma)$.

The results for the cases when the inverse covariance matrix has a banded structure are shown in Table 3, with the ROC curves provided in Figure 4. Based on sensitivity, out of all the true active entries, the MSSL method captures more true active entries in B compared with other two methods (SSUR and JDAG). However, MSSL falsely includes a higher proportion of non-true entries out of all detected entries compared with JDAG and OBFBF. In terms of DAG selection, based on precision, out of all the edges detected, SSUR could correctly

identify more edges compared with JDAG and MSSL, but based on sensitivity, out of all the true edges, fewer edges are correctly detected by SSUR compared with JDAG and MSSL. For both variable and DAG selection, the performance of OBFBF is considerably worse compared with all the other three methods. However, we would like to point out the performance differences between different methods might be due to the implementation. Since each method was originally coded in a different environment, when implementing the methods, rather than making much adjustment on the tuning parameters, we naturally adopt the default parameters for each of the competing methods.

Table 3: Joint selection performance under banded Ω .

	Variable selection				Graph selection			
	JDAG	SSUR	MSSL	OBFBF	JDAG	SSUR	MSSL	OBFBF
Precision	0.33	0.03	0.07	0.17	0.85	0.89	0.38	0.02
NPV	0.99	1	1	0.86	1	0.97	1	0.99
Sensitivity	0.22	0.46	1	0.01	0.72	0.16	0.98	0.02
Specificity	0.99	0.99	0.99	0.99	1	1	1	0.99
Runtime (min)	14.12	64.28	4.91	71.27	14.12	64.28	4.91	71.27

Figure 4: ROC curves for DAG selection under banded Ω for JDAG, MSSL, SSUR, and OBFBF.

4 Real data analysis

Since in simulation study, we observe OBFBF performs rather poorly compared with the other three methods, in real data analysis, we evaluate only the performance of JDAG, MSSL, and SSUR models. We focus on replication of established eQTLs of 189 SNPs located on Chromosome 17, and eQTL of asthma related Chromosome 17q12-21 locus while inferring the underlying gene network.

4.1 Compared with established eQTLs of 189 SNPs on Chromosome 17

We consider 82 unrelated individuals of Northern and Western European ancestry from Utah (CEU), whose genotypes are available from the International Hapmap project and publicly available at http://ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3/hapmap_format/draft_1/forward/.

The genotype is coded as 0, 1, and 2 for homozygous rare, heterozygous and homozygous common alleles. We focus on the 189 SNPs located on Chromosome 17 that have established eQTL associations with 31 probes as recorded at <http://www.exsnp.org/eQTL>. In addition to these 31 probes, we add 100 additional noise probes to evaluate the model performance. The gene expression data containing information for these 131 probes of 82 CEU subjects were analyzed as in Stranger et al. (2007). There were four replicates for each individual. The raw data were background corrected and then quantile normalized across four replicates of one single

subject and then median normalized across all subjects. The gene expression data are again publicly available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6536>. Therefore, in this setting, $n = 82$, $p = 189$, and $q = 131$, and the sparsity level in B is approximately 2%.

In real data analysis focusing on the 189 SNPs located on Chromosome 17, the JDAG yields five global significant SNPs: rs2074222, rs11651097, rs9900396, rs1470697, rs3744216. The results for detecting significant SNP-gene associations pertaining the SSUR and MSSL are shown in Table 4, where the associations are defined to be significant when either the corresponding values in the association matrix B are non-zero for MSSL or the posterior probabilities for the corresponding values in B are greater than the thresholding value for the other methods. MSSL identified more SNP associations compared with SSUR. The inferred graphs for the three methods are shown in Figure 5. In particular, the MSSL method estimates the DAG unrealistically with only three edges compared with SSUR and JDAG methods. For both methods, almost all of the detected loci have been previously established, which shows the utility of SSUR and MSSL for estimating the individual SNP-gene associations.

Table 4: Significant SNP-probe/gene associations on chromosome 17 identified by MSSL and SSUR.

	SNP	Probe ID	Gene symbol
SSUR	rs4130140, rs6565724	GI_42661283-S	C17orf97
	rs2255166	GI_27500527-S	SLFN5
	rs1122634, rs989128	GI_31542722-S	SPATA20
	rs6565724	GI_37544593-S	
	rs2779212	Hs.379903-S	
	rs883416	hmm5445-S	
	rs11868024	GI_29739147-S	ARL16
	rs883416	GI_42661257-S	SLFN5
	rs12450199	GI_11496988-S	GAA
	rs8081611	GI_38327653-S	CHRNE
	rs4790530	GI_4826681-S	CTNS
	rs2715555	GI_34147551-S	EIF5A
	rs238224	GI_14149701-S	RNF167
	rs402514	GI_44662825-I	ELP5
	rs8066241	GI_38142463-S	HEATR6
	rs2715555	GI_7705938-S	RAPGEFL1
	rs228293	GI_20336255-A	
	rs2725391	GI_27500527-S	
MSSL	rs9912300, rs989128	GI_31542722-S	SPATA20
	rs7502521	GI_29739147-S	ARL16
	rs2779212	Hs.379903-S	
	rs7214776, rs2521985, rs222837, rs402514	GI_44662829-A	ELP5
	rs12450199	GI_11496988-S	GAA
	rs228293, rs11651097, rs7225914	GI_21314623-S	PGS1
	rs3817182	GI_7705938-S	RAPGEFL1
	rs161370	GI_4826681-S	CTNS
	rs7217496	GI_23111027-I	SNX11
	rs7214776	Hs.446193-S	
	rs11651097	GI_14196461-A	PCDHGA2
	rs7225932	GI_22547185-I	SHMT1
	rs222852	GI_21389608-S	LSM14B
	rs228293	GI_20336255-A

Figure 5: Inferred gene/probe network for the CEU data. A: SSUR; B: MSSL; C: JDAG.

4.2 GWAS-based eQTL study on the Chromosome 17q12-21 locus

In this setting, we perform a GWAS-based eQTL study on the Chromosome 17q12-21 locus, which includes several genes that have been linked to asthma susceptibility in previous study (see Marinho et al., 2012; Cai et al., 2014; Lavoie-Charland et al., 2016). Out of the 47,293 total available probes corresponding to different Illumina TargetID, we select the 100 most variable probes, each of which corresponding to a different transcript. We focus on the 104 SNPs located on the Chromosome 17q12-21 locus. The SNP data and the gene expression data are both publicly available as in previous setting. Therefore, in this setting, $n = 82$, $p = 104$, and $q = 100$.

For the second real data analysis focusing on Chromosome 17q12-21 locus, the JDAG method yields five significant SNPs: rs1476278 (PGAP3), rs2313171 (PGAP3), rs881844 (STARD3), rs9972882 (STARD3), rs9635726 (IKZF3). All the three genes and SNPs have been previously reported to be associated with asthma development and susceptibility (see Marinho et al., 2012; Cai et al., 2014; Lavoie-Charland et al., 2016).

The significant SNP-gene association results based on SSUR and MSSL are shown in Table 5. The inferred graphs for three methods are given in Figure 6. The SSUR method yields a total of twelve significant SNPs: rs17405722, rs8070945, rs7220372, rs2277618, rs12952407, rs7223784, rs6963, rs630539, rs7218686, rs9898532, rs4796793, rs665268. Out of these twelve SNPs, two SNPs, rs17405722, rs4796793 are on gene STAT3. STAT3 has been previously reported to be associated with asthma susceptibility (see Lautner-Csorba et al., 2012; Zhang et al., 2015; Singh et al., 2017). For MSSL method, out of three SNPs detected, only one SNP, rs2290400 on gene GSDMB has been previously reported to be associated with asthma susceptibility (see Marinho et al., 2012; Stein et al., 2018).

Table 5: Significant SNP-probe/gene associations on 17q12-21.

	SNP	Probe ID	Gene Symbol
SSUR	rs7218686	GI_26665892-S	HLA-DRB5
	rs9898532	GI_42476191-S	TUBB2B

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

MSSL	rs7220372	GI_11993934-S	CHI3L2
	rs2277618	GI_11993934-S	CHI3L2
	rs12952407	GI_42660576-S	
	rs7223784	GI_34222299-S	EPS8
	rs6963	GI_8923472-S	NLRP2
	rs17405722	GI_41197088-S	
	rs8070945	GI_41190507-S	
	rs630539	GI_26665892-S	HLA-DRB5
	rs4796793	GI_22035637-A	MGST1
	rs665268	GI_22035637-A	MGST1
	rs2934971	GI_37546619-S	
	rs2290400	GI_14211892-S	GSDMB
	rs4252612	GI_37546619-S

Figure 6: GWAS Inferred gene network for the CEU data. A: SSUR; B: MSSL; C: JDAG.

5 Discussion

In this study, we present our Bayesian hierarchical model for joint variable selection and network estimation, and evaluate its performance along with three other existing methods (MSSL, SSUR and OBFBB) in simulation as well as real data analysis to detect eQTLs and estimate gene networks. For real data analysis, we use publicly available HapMap data, and acquire the SNP/gene associations and gene networks simultaneously. Compared with univariate regression approach, these approaches incorporate the dependence information among genotypes, and conveniently avoid the issues raised by multiple testing correction. In simulation study, our proposed method outperforms the other three methods in the row-wise sparse setting. In real data analysis, our method is able to detect significant SNPs that have been previously linked to asthma. In particular, all the other three methods lack the precision and accuracy in variable selection and network estimation under the row-wise sparse setting. Our analysis and interpretation requires replication with additional analysis

and simulation work. In particular, our work, rather than making much adjustment pertaining to the tuning parameters, adopt the default parameters for each of the competing methods that are coded under different computational environment. Future work will include optimization of each methods under similar computing environment.

We conducted downstream analyses of the genes identified both within the regulatory networks and in association with asthma via Ingenuity Pathway Analyses (IPA; www.ingenuity.com). The pathways which are related to the identified genes are listed in Table 6, and the corresponding pathway network is given in Figure 7. Immune system pathway is the most enriched pathways based on differential gene regulators. Immune system pathways have been implicated in the development and progression of asthma. Differentially regulated genes including IKZF3, STAT3, GSDMB, and PGAP3. For example, Ikaros family zinc finger protein 3 (IKZF3) is known to play important roles in tissue growth and remodeling and linked to several disease conditions including asthma (Torgerson et al., 2011). IKZF3 plays a critical role in regulating B and T cell development, and several alternative coding as well as noncoding transcripts variants (Duhamel et al., 2008). In genome-wide study, the most statistically significant SNP associated with asthma resided in a coding region of the IKZF3 gene (Galanter et al., 2014). Several studies have identified several IKZF3 SNPs to be associated with a childhood wheezing phenotype, childhood eczema and asthma (Marenholz et al., 2015; Marinho et al., 2012), early-life environmental tobacco smoke exposure, and lung function (Blekic et al., 2013). The gasdermin B (GSDMB) locus was one of the strongest associations for asthma in the EVE Consortium (Torgerson et al., 2011). In summary, although we focused on eQTL analysis and confirmed previously discovered asthma genes, the JDAG approach can be easily extended for multi-omics analysis including metabolomics or proteomics datasets.

Table 6: Enriched pathways from differential gene regulatory network between asthma vs. controls.

Categories	Diseases or functions annotation	<i>p</i> -value	Molecules
Immunological disease	Systemic autoimmune syndrome	0.0027	GSDMB, IKZF3, STAT3
Inflammatory response, organismal injury and abnormalities	Inflammation of organ	0.0033	IKZF3, PGAP3, STAT3
Inflammatory disease, respiratory disease	Asthma	0.0010	GSDMB, IKZF3
Inflammatory response	Inflammation of absolute anatomical region	0.0289	PGAP3, STAT3
Immunological disease, organismal injury and abnormalities	Lymphadenopathy of cervical lymph node	0.0002	STAT3
Inflammatory disease, organismal injury and abnormalities, respiratory disease	Fibrosis of peribronchial tissue	0.0005	GSDMB

Figure 7: Four genes (STAT3, GSDMB, PGAP3, IKZF3) pathway network.

Acknowledgments

This research was supported in part by the National Institutes of Health (NIH) Funder Id: <http://dx.doi.org/10.13039/100000009>, grant R01HL132344 and the Simons Foundation's collaboration grant (Funder Id: <http://dx.doi.org/10.13039/100000893>, No. 635213).

Conflict of interest statement: The authors declare that they have no competing interests.

References

- Ackermann, M., W. Sikora-Wohlfeld and A. Beyer (2013): Impact of natural genetic variation on gene expression dynamics. *PLoS Genet.*, 9, e1003514.
- Banerjee, S. and S. Ghosal (2015): Bayesian structure learning in graphical models. *J. Multivar. Anal.*, 136, 147–162.
- Banterle, M., Bottolo, L. Richardson, S. Ala-Korpela, M.-R. Järvelin, A. Lewin (2018): Sparse variable and covariance selection for high-dimensional seemingly unrelated bayesian regression. Preprint from bioRxiv. DOI: <https://doi.org/10.1101/467019>.
- Bhadra, A. and B. K. Mallick (2013): Joint high-dimensional Bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics*, 69, 447–457.
- Blekic, M., B. Kljaic Bukvic, N. Aberle, S. Marinho, J. Hankinson, A. Custovic and A. Simpson (2013): 17q12-21 and asthma: Interactions with early-life environmental exposures. *Ann. Allergy Asthma Immunol.*, 110, 347–353.
- Bottolo, L. and S. Richardson (2010): Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.*, 5, 583–618.
- Cai, X., Y. Qiao, C. Diao, X. Xu, Y. Chen, S. Du, X. Liu, N. Liu, S. Yu, D. Chen and Y. Jiang (2014): Association between polymorphisms of the *ikzf3* gene and systemic lupus erythematosus in a Chinese han population. *PLoS One*, 9, 1–6.
- Cao, X., K. Khare and M. Ghosh (2019a): Consistent bayesian sparsity selection for high-dimensional gaussian DAG models with multiplicative and beta-mixture priors. Preprint from arXiv. <https://arxiv.org/abs/1903.03531>.
- Cao, X., K. Khare and M. Ghosh (2019b): Posterior graph selection and estimation consistency for high-dimensional Bayesian dag models. *Ann. Statist.*, 47, 319–348.
- Consonni, G., L. L. Rocca and S. Peluso (2017): Objective bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.*, 44, 741–764.

- Deshpande, S. K., V. Rocková and E. I. George (2019): Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *J. Comput. Graph. Stat.*, 28, 921–931.
- Duhamel, M., I. Arrouss, H. Merle-Beral and A. Rebollo (2008): The Aiolos transcription factor is up-regulated in chronic lymphocytic leukemia. *Blood*, 111, 3225–3228.
- Galanter J. M., C. R. Gignoux, D. G. Torgerson, L. A. Roth, C. Eng, S. S. Oh, E. A. Nguyen, K. A. Drake, S. Huntsman, D. Hu, S. Sen, A. Davis, H. J. Farber, P. C. Avila, E. Brigino-Buenaventura, M. A. LeNoir, K. Meade, D. Serebrisky, L. N. Borrell, W. Rodríguez-Cintrón, A. M. Estrada, K. S. Mendoza, C. A. Winkler, W. Klitz, I. Romieu, S. J. London, F. Gilliland, F. Martinez, C. Bustamante, L. K. Williams, R. Kumar, J. R. Rodríguez-Santana and E. G. Burchard (2014): Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the genes-environments & Admixture in Latino Americans study. *J. Allergy Clin. Immunol.*, 134, 295–305.
- GTEX Consortium (2015): The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660.
- Lautner-Csorba, O., A. Gézsi, A. F. Semsei, P. Antal, D. J. Erdélyi, G. Schermann, N. Kutszegi, K. Csordás, M. Hegyi, G. Kovács, A. Falus and C. Szalai (2012): Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. *BMC Med. Genomic*, 5, 42.
- Lavoie-Charland, E., J. Brub, L. Boulet and Y. Boss (2016): Asthma susceptibility variants are more strongly associated with clinically similar subgroups. *J. Asthma*, 53, 907–913.
- Lee, J., S. Kim, J. Jhong and J. Koo (2018): Variable selection and joint estimation of mean and covariance models with an application to eQTL data. *Comput. Math. Method. Med.*, 2018, 1–13.
- Liang, F. and W. H. Wong (2000): Evolutionary monte carlo: applications to cp model sampling and change point problem. *Stat. Sin.*, 10, 317–342.
- Marenholz I, J. Esparza-Gordillo, F. Ruschendorf, A. Bauerfeind, D. P. Strachan, B. D. Spycher, H. Baurecht, P. Margaritte-Jeannin, A. Sääf, M. Kerkhof, M. Ege, S. Baltic, M. C. Matheson, J. Li, S. Michel, W. Q. Ang, W. McArdle, A. Arnold, G. Homuth, F. Demenais, E. Bouzigon, C. Söderhäll, G. Pershagen, de J. C. Jongste, D. S. Postma, C. Braun-Fahrlander, E. Horak, L. M. Ogorodova, V. P. Puzyrev, E. Y. Bragina, T. J. Hudson, C. Morin, D. L. Duffy, G. B. Marks, C. F. Robertson, G. W. Montgomery, B. Musk, P. J. Thompson, N. G. Martin, A. James, P. Sleiman, E. Toskala, E. Rodriguez, Fölster-R. Holst, A. Franke, W. Lieb, C. Gieger, A. Heinzmann, E. Rietschel, T. Keil, S. Cichon, M. M. Nöthen, C. E. Pennell, P. D. Sly, C. O. Schmidt, A. Matanovic, V. Schneider, M. Heinig, N. Hübner, P. G. Holt, S. Lau, M. Kabesch, S. Weidinger, H. Hakonarson, M. A. R. Ferreira, C. Laprise, M. B. Freidin, J. Genuneit, G. H. Koppelman, E. Melén, M. H. Dizier, A. J. Henderson and Y. A. Lee. (2015): Meta-analysis identifies seven susceptibility loci involved in the atopic march. *Nat. Commun.* 6, 8804.
- Marinho, S., A. Custovic, P. Marsden, J. Smith and A. Simpson (2012): 17q12-21 variants are associated with asthma and interact with active smoking in an adult population from the United Kingdom. *Ann. Allergy Asthma Immunol.*, 108, 402–411.e9.
- Nica, A. C. and E. T. Dermitzakis (2013): Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368, 20120362.
- Rockman, M. V. and L. Kruglyak (2006): Genetics of global gene expression. *Nat. Rev. Genet.*, 7, 862–872.
- Singh, D., P. Bagam, M. Sahoo and S. Batra (2017): Immune-related gene polymorphisms in pulmonary diseases. *Toxicology*, 383, 24–39.
- Stein, MM, E. E. Thompson, N. Schoettler, B. A. Helling, K. M. Magnaye, C. Stanhope, C. Igartua, A. Morin, C. Washington 3rd, D. Nicolae, K. Bønnelykke and C. Ober (2018): A decade of research on the 17q12-21 asthma locus: Piecing together the puzzle. *J. Allergy Clin. Immunol.*, 142, 749–764.e3.
- Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles and E. T. Dermitzakis (2007): Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315, 848–853.
- Torgerson, DG, E. J. Ampleford, G. Y. Chiu, W. J. Gauderman, C. R. Gignoux, P. E. Graves, B. E. Himes, A. M. Levin, R. A. Mathias, D. B. Hancock, J. W. Baurley, C. Eng, D. A. Stern, J. C. Cledón, N. Rafaels, D. Capurso, D. V. Conti, L. A. Roth, M. Soto-Quiros, A. Toghias, X. Li, R. A. Myers, I. Romieu, D. J. Van Den Berg, D. Hu, N. N. Hansel, R. D. Hernandez, E. Israel, M. T. Salam, J. Galanter, P. C. Avila, L. Avila, J. R. Rodriguez-Santana, R. Chapela, W. Rodriguez-Cintrón, G. B. Diette, N. F. Adkinson, R. A. Abel, K. D. Ross, M. Shi, M. U. Faruque, G. M. Dunston, H. R. Watson, V. J. Mantese, S. C. Ezurum, L. Liang, I. Ruczinski, J. G. Ford, S. Huntsman, K. F. Chung, H. Vora, X. Li, W. J. Calhoun, M. Castro, J. J. Sienna-Monge, B. del Rio-Navarro, K. A. Deichmann, A. Heinzmann, S. E. Wenzel, W. W. Busse, J. E. Gern, R. F. Lemanske Jr, T. H. Beaty, E. R. Bleecker, B. A. Raby, D. A. Meyers, S. J. London, Mexico City Childhood Asthma Study (MCAAS), F. D. Gilliland, Children's Health Study (CHS) and HARBORS study, E. G. Burchard, Genetics of Asthma in Latino Americans (GALA) Study, Study of Genes-Environment and Admixture in Latino Americans (GALA2) and Study of African Americans, Asthma, Genes & Environments (SAGE), F. D. Martinez, Childhood Asthma Research and Education (CARE) Network, S. T. Weiss, Childhood Asthma Management Program (CAMP), L. K. Williams, Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE), K. C. Barnes, Genetic Research on Asthma in African Diaspora (GRAAD) Study, C. Ober and D. L. Nicolae (2011): Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.*, 43, 887–892.
- Yang, Y., M. Wainwright and M. Jordan (2016): On the computational complexity of high-dimensional bayesian variable selection. *Ann. Stat.*, 44, 2497–2532.
- Zhang, Y., J. Li, C. Wang and L. Zhang (2015): Association between the interaction of key genes involved in effector T-cell pathways and susceptibility to develop allergic rhinitis: a population-based case-control association study. *PLoS One* 10, e0131248.