Biometrics 70, 872–880
December 2014

DOI: 10.1111/biom.12213

Meta-Analysis Based Variable Selection for Gene Expression Data

Quefeng Li,[1] Sijian Wang,[1,2] Chiang-Ching Huang,[3] Menggang Yu,[2] and Jun Shao[4,*]

[1]Department of Statistics, University of Wisconsin, Madison, Wisconsin, U.S.A.
[2]Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, Wisconsin, U.S.A.
[3]Feinberg School of Medicine, Northwestern University, Illinois, U.S.A.
[4]School of Finance and Statistics, East China Normal University, Shanghai, China
[*]*email:* shao@stat.wisc.edu

Summary. Recent advance in biotechnology and its wide applications have led to the generation of many high-dimensional gene expression data sets that can be used to address similar biological questions. Meta-analysis plays an important role in summarizing and synthesizing scientific evidence from multiple studies. When the dimensions of datasets are high, it is desirable to incorporate variable selection into meta-analysis to improve model interpretation and prediction. According to our knowledge, all existing methods conduct variable selection with meta-analyzed data in an "all-in-or-all-out" fashion, that is, a gene is either selected in all of studies or not selected in any study. However, due to data heterogeneity commonly exist in meta-analyzed data, including choices of biospecimens, study population, and measurement sensitivity, it is possible that a gene is important in some studies while unimportant in others. In this article, we propose a novel method called meta-lasso for variable selection with high-dimensional meta-analyzed data. Through a hierarchical decomposition on regression coefficients, our method not only borrows strength across multiple data sets to boost the power to identify important genes, but also keeps the selection flexibility among data sets to take into account data heterogeneity. We show that our method possesses the gene selection consistency, that is, when sample size of each data set is large, with high probability, our method can identify all important genes and remove all unimportant genes. Simulation studies demonstrate a good performance of our method. We applied our meta-lasso method to a meta-analysis of five cardiovascular studies. The analysis results are clinically meaningful.

Key words: Gene selection; High dimension; Meta-analysis; Weak oracle property.

1. Introduction

High-dimensional gene expression data analysis is a useful tool to discover complicated biological mechanisms. Many reported findings, however, are not reproducible (Ntzani et al., 2003), sensitive to mild data perturbations (Ein-Dor et al., 2005; Michiels, Koscielny, and Hill, 2005), or lack of generalizability (Ferguson, 2004), due to small sample sizes relative to large number of genes and low signal-to-noise ratios in most gene expression data sets. As a potential solution to these problems, meta-analysis is a relatively inexpensive option, since many genomic databases are nowadays publicly available. There exist gene expression meta-analysis methods based on combining univariate summary statistics (they will be referred to as CSS in what follows), for example, p-values (Rhodes et al., 2002), effect sizes (Choi et al., 2003; Grützmann et al., 2005; Bhattacharjee et al., 2012; Han and Eskin, 2012), or ranks (DeConde et al., 2006; Zintzaras and Ioannidis, 2008). Li and Tseng (2011) extends the well-known Fisher's method to an adaptively weighted (AW) sum of logarithm of p-values with the optimal weights chosen by an exhaustive search over all subsets of studies. Similar strategies could be applied to combine effect sizes of Fixed Effects Model (FEM) or Random Effects Model (REM) from individual studies. A comprehensive review of these methods is given in Tseng, Ghosh, and Feingold (2012). The CSS tends to gain more power to identify differentially expressed (DE) gene. But it neglects correlations among genes. This motivates us to develop a novel variable selection method to handle high-dimensional gene expression data.

For a single data set, there exist many variable selection methods, for example, lasso (Tibshirani, 1996), elastic-net (Zou and Hastie, 2005), and others described in a review by Fan and Lv (2010). An ad hoc approach for meta analysis is to apply one of these methods to each individual data set. To fully make use of different data sets, we may stack all data sets into one large data set and then apply variable selection. Heterogeneity among data sets may be addressed (e.g., Ma and Jian, 2009; Liu, Dunson, and Zou, 2011; Ma, Huang, and Song, 2011). However, most existing methods conduct variable selection in an "all-in-or-all-out" fashion; that is, they identify a gene to be either important in all studies or unimportant in all studies. Data heterogeneity can arise from not only inconsistent experiment conditions, sample preparation methods and sample qualities, but also different choices of biospecimens, distinct study populations, and varying measurement sensitivity or precision. Thus, it is quite likely that an important gene in some studies may have null effect in other studies, and it is important to allow such flexibility due to the exploratory and discovery nature of gene expression studies.

In this article, we propose a lasso method that borrows strength across multiple data sets, but meanwhile keeps the selection flexibility in each individual data set. We develop an efficient algorithm and study the theoretical property of our proposed method. We prove a weak "oracle" property (Fan and Lv, 2011) that with large probability, our method can remove unimportant genes and consistently estimate the effect of important genes in each individual data set. In a simulation study, we examine empirical properties of the proposed lasso and compare it with an ad hoc lasso, a stack lasso, a group lasso, and four CSS methods in Li and Tseng (2011). A real data example is also considered.

Since our method is likelihood-based, it has the advantage to generate good prediction accuracy and perform variable selection simultaneously, but is restricted to some model assumptions. The simple CSS methods are model-free, easier to compute compared with the proposed method, and may have high power in DE gene detection. However, they do not consider correlations among genes and the relationship of genes with responses.

2. Motivational Example

Atherosclerosis accounts for the vast majority of fatal and non-fatal cardiovascular disease (CVD) events. Additionally, about half of all strokes are caused by atherosclerosis, the process that leads to narrowing and hardening of the arteries. Although traditional risk factors such as hyperlipidemia, hypertension, diabetes, smoking are considered as the causes of atherosclerosis, substantial numbers of individuals with low burden of CVD risk factors or in middle age (40–55 years old) still develop atherosclerosis (Rietzschel et al., 2006; Kelley et al., 2011). Therefore, identification of molecular mechanisms mediating atherogenesis is imperative to enhance current knowledge of preventing CVD and stroke.

A wealth of recent investigations has revealed that inflammation plays a central role in atherosclerosis (Libby, 2002; Hansson, 2005). However, the essential genes modulating the inflammatory process in atherosclerosis have yet to be identified to develop effective treatment and prevention strategies. The immune system is the main defense machinery to regulate inflammatory responses. It is therefore justifiable to look for genes in the immune signaling pathways that are responsible for initiating and perpetuating atherogenesis. Recent studies on diseases of similar phenotype have shown that defects in genes, alone or in combination, in the same pathway can cause overlapping clinical manifestations (Wang, Li, and Bucan, 2007; Torkamani, Topol, and Schork, 2008; Askland, Read, and Moore, 2009; Farber, 2013). This suggests that analysis of gene expression profiling of immune cells from patients with atherosclerosis or its clinical events such as myocardial infarction or stroke may reveal common genes responsible for the disease progression. As a first attempt, we perform a meta-analysis in this article to investigate the expression changes for 88 genes in the toll-like receptor (TLR) signaling pathway using five microarray data sets. These 88 genes are identified from the Reactome database (www.reactome.org) and are found across all five data sets. TLRs are the most studied PRRs and are the major signaling pathway in the innate immunity to coordinate the adaptive immune response and trigger inflammation (Iwasaki and Medzhitov, 2004). Follow-up work involving multiple pathways related to the immune system is currently under investigation. The studies that we include in our analysis are case-control studies with the datasets publicly available on Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/). Table 1 gives a detailed summary of the five data sets we used for our analysis.

3. Methodology

Consider M independent studies, each of which contains n_m subjects. Denote y_{mi} as a binary phenotype and $\boldsymbol{x}_{mi} = (x_{mi,1}, \ldots, x_{mi,p})^{\mathrm{T}}$ as a vector of expression profiles of p genes of the ith subject in the mth data set, where the superscript T denotes the standard vector transpose. The p genes are assumed common in all data sets. We assume the conditional probability that y_{mi} takes value 1 given the vector of gene expression follows the logistic regression model

$$\log \frac{\Pr(y_{mi}=1|\boldsymbol{x}_{mi})}{\Pr(y_{mi}=0|\boldsymbol{x}_{mi})} = \beta_{m0} + \boldsymbol{x}_{mi}^{\mathrm{T}}\boldsymbol{\beta}_m, \quad i=1,\ldots,n_m, \; m=1,\ldots,M, \tag{1}$$

where β_{m0} is an intercept and $\boldsymbol{\beta}_m = (\beta_{m1}, \ldots, \beta_{mp})^{\mathrm{T}}$ is a vector of regression coefficients for the mth data. Because of the data heterogeneity, we allow β_{m0} and $\boldsymbol{\beta}_m$ in (1) to vary with m. Variable (gene) selection amounts to finding zero components of $\boldsymbol{\beta}_m$, $m = 1, \ldots, M$.

The ad hoc separate-fit approach mentioned in the introduction section maximizes M separate penalized likelihoods, each of which is

$$\ell_m(\beta_{m0}, \boldsymbol{\beta}_m) - \lambda_m J(\boldsymbol{\beta}_m), \tag{2}$$

where $\ell_m(\beta_{m0}, \boldsymbol{\beta}_m)$ is the log-likelihood for the mth data set, λ_m is a non-negative tuning parameter and J is a certain sparsity-induced penalty. In this article, we consider the lasso (L_1-norm) penalty $J(\boldsymbol{\beta}_m) = \sum_{j=1}^{p} |\beta_{mj}|$. We call this method as the "separate lasso."

A "stack lasso" assumes that $\beta_{m0} = \beta_0$ and $\boldsymbol{\beta}_m = \boldsymbol{\beta}$ for all m in model (1), and conducts variable selection by maximizing the penalized likelihood of the stacked data:

$$\sum_{m=1}^{M} \ell_m(\beta_0, \boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where λ is a non-negative tuning parameter.

Ma et al. (2011) proposed to treat the effect vector $(\beta_{1j}, \ldots, \beta_{Mj})^{\mathrm{T}}$ for each j as a group and use a group-effect penalty. Thus, a "group lasso" conducts variable selection by maximizing

$$\sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{\beta}_m) - \lambda \sum_{j=1}^{p} \left(\sum_{m=1}^{M} \beta_{mj}^2 \right)^{1/2}.$$

Both stack lasso and group lasso select variables in an "all-in-or-all-out" fashion. We propose a *joint* fitting approach that borrows strength across different data sets as well as incorporates data heterogeneity. We consider the following

Table 1
Summaries of datasets in the five cardiovascular studies

	GSE12288	GSE16561	GSE20129	GSE22255	GSE28829
Total sample size	222	63	119	40	29
Case group size	110	39	48	20	16
Control group size	112	24	71	20	13
No. of genes in TLR	95	95	99	98	98
Platform	Affymetrix U133A	Illumina HumanRef8 V3.0	Illumina HumanRef8 V2.0	Affymetrix U133plus2	Affymetrix U133plus2
Bioassay	Whole blood	Whole blood	Whole blood and peripheral blood leukocytes	Peripheral blood mononuclear cells	Atherosclerotic carotid artery segments

GSE12288 studied Coronary Artery Disease (CAD) with cases defined as those with at least a stenosis greater than 50% and controls without evidence of coronary stenosis (Sinnaeve et al., 2009). GSE16561 studied ischemic stroke with cases defined as patients with acute ischemic cerebrovascular syndrome and controls without stroke (Barr et al., 2010). GSE20129 studied atherosclerosis that included women aged 50 year or above in the Multi-Ethnic Study of Atherosclerosis cohort (Huang et al., 2011). The cases are women who had coronary artery calcium (CAC) score >100 and carotid intima-media thickness (IMT) > 1 mm and controls had CAC <10 and IMT <0.65 mm. GSE22255 studied ischemic stroke (Krug et al., 2012). GSE28829 is an atherosclerotic progression study that used cases with advanced lesions and controls with early lesions (Daissormont et al., 2011). For Affymetrix data (GSE12288, 22255, 28829), the RMA algorithm was used for data normalization. For Illumina data (GSE16561, 20129), quantile normalization procedure was applied.

hierarchical reparameterization:

$$\beta_{mj} = g_j \zeta_{mj}, \quad m = 1, \ldots, M; \; j = 1, \ldots, p. \tag{4}$$

The parameter g_j is an effect of the jth gene at the first level of the hierarchy and ζ_{mj}'s with different m's reflect effect differences for the jth gene among M data sets at the second level of the hierarchy. If there is no heterogeneity, then $g_j = \beta_j$ defined in (3) and $\zeta_{mj} = 1$ for all j and m. In reparameterization (4), exact values of g_j and ζ_{mj} are not identifiable but we can identify whether they are equal to 0.

With reparameterization (4), we propose a variable selection method by solving

$$\max_{\boldsymbol{\beta}_0, \boldsymbol{g}, \boldsymbol{\zeta}} \left\{ \sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m) - \lambda_g \sum_{j=1}^{p} |g_j| - \lambda_\zeta \sum_{j=1}^{p} \sum_{m=1}^{M} |\zeta_{mj}| \right\} \tag{5}$$

where

$$\ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m) = \sum_{i=1}^{n_m} y_{mi}\{\beta_{m0} + \boldsymbol{x}_{mi}^{\mathrm{T}}(\boldsymbol{g} \cdot \boldsymbol{\zeta}_m)\} - \log[1 + \exp\{\beta_{m0} + \boldsymbol{x}_{mi}^{\mathrm{T}}(\boldsymbol{g} \cdot \boldsymbol{\zeta}_m)\}],$$

$\boldsymbol{g} = (g_1, \ldots, g_p)^{\mathrm{T}}$, $\boldsymbol{\zeta}_m = (\zeta_{m1}, \ldots, \zeta_{mp})^{\mathrm{T}}$, $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{M0})^{\mathrm{T}}$, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^{\mathrm{T}}, ..., \boldsymbol{\zeta}_M^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{g} \cdot \boldsymbol{\zeta}_m$ means the element-wise product.

The tuning parameter λ_g controls variable selection at the entire-gene level and can effectively remove genes that are unimportant for all M data sets. The tuning parameter λ_ζ controls variable selection at the individual data set level: if g_j is not equal to zero, some ζ_{mj} and hence the corresponding β_{mj} can still be shrunken to zero. Since the estimation of g_j depends on all M data sets, the estimation of regression coefficients $\beta_{mj} = g_j \zeta_{mj}$ depends on all M data sets. For this reason, we call this method as the "meta lasso."

Two modifications may be made in the optimization problem (5). First, a pre-specified weight w_m may be multiplied to $\ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m)$ in (5) to reflect the importance of mth data set. Second, since all M data sets have the same genes, it is intuitively appealing to require that the non-zero estimated effects of an identified gene has the same sign across all M data sets. Therefore, we may add a sign constraint

$$\zeta_{mj} \geq 0, \quad m = 1, \ldots, M; \; j = 1, \ldots, p, \tag{6}$$

in (5). In fact, in the analysis of real data and simulation studies in Section 5, constraint (6) is applied. In the rest of this section and the theoretical study (Section 4), we ignore weighting and constraint (6). But the results can be easily extended to the case with the sign constraint or weights.

Although there are two tuning parameters in (5), it turns out that they can be simplified into one. Specifically, in Lemma 1 we can show that (5) is equivalent to the following optimization problem with one tuning parameter $\lambda = \lambda_g \lambda_\zeta$:

$$\max_{\boldsymbol{\beta}_0, \boldsymbol{g}, \boldsymbol{\zeta}} \left\{ \sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m) - \sum_{j=1}^{p} |g_j| - \lambda \sum_{j=1}^{p} \sum_{m=1}^{M} |\zeta_{mj}| \right\}. \tag{7}$$

Lemma 1. *If $(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\zeta}})$ is a solution of (5), then there exists a solution $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{g}}, \hat{\boldsymbol{\zeta}})$ of (7) such that $\tilde{g}_j \tilde{\zeta}_{mj} = \hat{g}_j \hat{\zeta}_{mj}$ and $\tilde{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0$. Similarly, if $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{g}}, \hat{\boldsymbol{\zeta}})$ is a solution of (7), then there exists a solution $(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{\zeta}})$ of (5) such that $\hat{g}_j \hat{\zeta}_{mj} = \tilde{g}_j \tilde{\zeta}_{mj}$ and $\hat{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}}_0$.*

Next, we show that (7) has an equivalent form in terms of β_{mj}'s, which is useful for the derivation of theoretical results in Section 4.

Lemma 2. *If $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{g}}, \hat{\boldsymbol{\zeta}})$ is a solution of (7), then $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{10}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{Mp})^{\mathrm{T}}$ with $\hat{\beta}_{mj} = \hat{g}_j \hat{\zeta}_{mj}$, for $j = 1, \ldots, p$, is a* *solution of*

$$\max_{\boldsymbol{\beta}} \left\{ \sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{\beta}_m) - 2\sqrt{\lambda} \sum_{j=1}^{p} \left(\sum_{m=1}^{M} |\beta_{mj}| \right)^{1/2} \right\}, \qquad (8)$$

where $\boldsymbol{\beta} = (\beta_{10}, \beta_{11}, \ldots, \beta_{Mp})^{\mathrm{T}}$. *On the other hand, if* $\hat{\boldsymbol{\beta}}$ *is a solution of (8), then* $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{g}}, \hat{\boldsymbol{\zeta}})$ *is a solution of (7), where* $\hat{\boldsymbol{\beta}}_0 = (\hat{\beta}_{10}, \ldots, \hat{\beta}_{M0})^{\mathrm{T}}$, $\|\hat{\boldsymbol{\beta}}_{(j)}\|_1 = \sum_{m=1}^{M} |\hat{\beta}_{mj}|$,

$$(\hat{\boldsymbol{g}}, \hat{\boldsymbol{\zeta}}) = \begin{cases} \hat{g}_j = 0, \hat{\boldsymbol{\zeta}}_{(j)} = \mathbf{0} & \textit{if } \hat{\boldsymbol{\beta}}_{(j)} = \mathbf{0}, \\ \hat{g}_j = (\lambda \|\hat{\boldsymbol{\beta}}_{(j)}\|_1)^{1/2}, \ \hat{\boldsymbol{\zeta}}_{(j)} = \hat{\boldsymbol{\beta}}_{(j)}/\hat{g}_j & \textit{if } \hat{\boldsymbol{\beta}}_{(j)} \neq \mathbf{0}, \end{cases}$$

where $\hat{\boldsymbol{\beta}}_{(j)} = (\hat{\beta}_{1j}, \ldots, \hat{\beta}_{Mj})^{\mathrm{T}}$, $\hat{\boldsymbol{\zeta}}_{(j)} = (\hat{\zeta}_{1j}, \ldots, \hat{\zeta}_{Mj})^{\mathrm{T}}$, *and* $\hat{\boldsymbol{\zeta}} = (\hat{\zeta}_{11}, \hat{\zeta}_{12}, \ldots, \hat{\zeta}_{Mp})^{\mathrm{T}}$.

If we regard one gene's effects among all studies as a "group," then (8) imposes a square root penalty on each group and an L_1 penalty on individual elements within a group. It turns out that this penalization scheme is the same as the group bridge penalty proposed in Huang et al. (2009). However, it is hard to see why solving (8) is a reasonable way to handle variable selection in meta-analysis, because it is not natural to form the effects of the jth gene in M different data sets as a group. Our reparameterization (4) is intuitively appealing and it naturally leads to the maximization problem (5) and its equivalent form (7), which happens to have a connection with (8) revealed by our Lemma 2. Moreover, (8) is a non-concave problem that is hard to solve, although Huang et al. (2009) provided an algorithm. With the aid of reparameterization (4), (7) can be decomposed into two concave problems, each of which views $\boldsymbol{g}$ or $\boldsymbol{\zeta}$ as fixed. Thus, we in fact obtain a more effective computation method to solve the group bridge problem (8). Finally, our theoretical results in Section 4 are established not only for variable selection in meta-analysis, but also for variable selection by the group bridge penalty, which extends the results in Huang et al. (2009) since they did not consider the case of large p.

We propose to solve $\boldsymbol{\beta}_0$, $\boldsymbol{g}$, and $\boldsymbol{\zeta}$ in (7) iteratively. We first fix $\boldsymbol{\beta}_0$ and $\boldsymbol{\zeta}$ in (7) to maximize over $\boldsymbol{g}$. Next, we maximize over $\boldsymbol{\zeta}$ by fixing $\boldsymbol{\beta}_0$ and $\boldsymbol{g}$. Finally, we maximize over $\boldsymbol{\beta}_0$ by fixing $\boldsymbol{g}$ and $\boldsymbol{\zeta}$. We iterate between these steps until the algorithm converges. Specifically, the algorithm is described as follows:

Algorithm:

1. For each dataset, standardize columns to have zero mean and unit variance. Initialize $\hat{\zeta}_{mj}^{(0)} = 1$ for $1 \leq m \leq M; 1 \leq j \leq p$ and $\hat{\beta}_{m0}^{(0)} = 0$ for $1 \leq m \leq M$.

2. In the kth iteration, let $\tilde{x}_{mi,j} = x_{mi,j}\hat{\zeta}_{mj}^{(k-1)}$, where $\hat{\zeta}_{mj}^{(k-1)}$ is the value of $\hat{\zeta}_{mj}$ at the $(k-1)$th step and

$$\ell(\boldsymbol{g}) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left[y_{mi} \left(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p} \tilde{x}_{mi,j} g_j \right) - \log \left\{ 1 + \exp \left(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p} \tilde{x}_{mi,j} g_j \right) \right\} \right].$$

Estimate g_j by $\hat{g}_j^{(k)} = \operatorname{argmax}_{g_j} [\ell(\boldsymbol{g}) - \sum_{j=1}^{p} |g_j|]$, $j = 1, \ldots, p$.

3. Let $\check{x}_{mi,j} = x_{mi,j}\hat{g}_j^{(k)}$ and

$$\ell(\boldsymbol{\zeta}) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left[y_{mi} \left(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p} \check{x}_{mi,j} \zeta_{mj} \right) - \log \left\{ 1 + \exp \left(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p} \check{x}_{mi,j} \zeta_{mj} \right) \right\} \right].$$

Estimate ζ_{mj} by $\hat{\zeta}_{mj}^{(k)} = \operatorname{argmax}_{\zeta_{mj}} [l(\boldsymbol{\zeta}) - \lambda \sum_{j=1}^{p} \sum_{m=1}^{M} |\zeta_{mj}|]$. Add constraint $\zeta_{mj} \geq 0$ for $m = 1, \ldots, M; j = 1, \ldots, p$, if necessary.

4. Let $\hat{\beta}_{mj}^{(k)} = \hat{g}_j^{(k)} \hat{\zeta}_{mj}^{(k)}$ for $m = 1, \ldots, M;\ j = 1, \ldots, p$ and

$$\ell(\boldsymbol{\beta}_0) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left[y_{mi} \left(\beta_{m0} + \sum_{j=1}^{p} x_{mi,j} \hat{\beta}_{mj}^{(k)} \right) - \log \left\{ 1 + \exp \left(\beta_{m0} + \sum_{j=1}^{p} x_{mi,j} \hat{\beta}_{mj}^{(k)} \right) \right\} \right].$$

Estimate β_{m0} by $\hat{\beta}_{m0}^{(k)} = \operatorname{argmax}_{\beta_{m0}} \ \ell(\boldsymbol{\beta}_0)$.

5. If $\max_{1 \leq m \leq M, 0 \leq j \leq p} |\beta_{mj}^{(k)} - \beta_{mj}^{(k-1)}|$ is less than some predefined threshold (we use 10^{-5}), terminate the algorithm; otherwise go back to step 2 and iterate.

Since the objective function in (7) increases in each iteration and the objective function itself is concave, the convergence of this algorithm is guaranteed. Step 2 is a lasso-type problem and Step 3 is a nonnegative garrote-type problem. Both can be effectively solved by the coordinate descent algorithm in (Friedman, Hastie, and Tibshirani, 2010).

4. Theoretical Properties

For the ease of presentation, we assume that the sample sizes are equal to n in all M studies. In addition, we assume the intercept term $\boldsymbol{\beta}_0 = \mathbf{0}$ and drop $\boldsymbol{\beta}_0$ from (5), (7), and (8). After multiplying a constant, we rewrite the optimization problem (8) in a matrix form as

$$\max_{\boldsymbol{\beta} \in \mathcal{R}^{Mp}} [\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}} \boldsymbol{b}(\boldsymbol{\theta}) - n\lambda_n \rho(\boldsymbol{\beta})] \qquad (9)$$

where $\boldsymbol{X} \in \mathcal{R}^{Mn \times Mp}$ is the block-diagonal design matrix whose mth block contains data from mth study, $\boldsymbol{Y} = (y_{11}, y_{12}, \ldots, y_{Mn})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{Mp})^{\mathrm{T}}$, $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$, $\boldsymbol{b}(\boldsymbol{\theta})$ is a $\mathcal{R}^{Mn} \to \mathcal{R}^{Mn}$ function that $b(\theta_i) = \log\{1 + \exp(\theta_i)\}$, the penalty term $\rho(\boldsymbol{\beta}) = \sum_{j=1}^{p} (\sum_{m=1}^{M} |\beta_{mj}|)^{1/2}$, and $\lambda_n > 0$ is a tuning parameter. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{11}, \hat{\beta}_{12}, \ldots, \hat{\beta}_{Mp})^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}}_{(j)} = (\hat{\beta}_{1j}, \ldots, \hat{\beta}_{Mj})^{\mathrm{T}}$ be a subvector of $\hat{\boldsymbol{\beta}}$ corresponding to the effects of jth gene. In addition, we denote $\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \ldots, b'(\theta_{Mn}))^{\mathrm{T}}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \operatorname{diag}(b''(\theta_1), \ldots, b''(\theta_{Mn}))$, where $b'(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}$ and $b''(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}^2$.

In spirit of Theorem 1 in Fan and Lv (2011), we give the necessary and sufficient condition for a vector $\hat{\boldsymbol{\beta}}$ to be the solution to (9).

Theorem 1. *Any $\hat{\boldsymbol{\beta}} \in \mathcal{R}^{Mp}$ is a strict local maximizer of the objective function in (9) if*

$$\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = n\lambda_n \nabla\rho(\hat{\boldsymbol{\beta}}_I), \tag{10}$$

$$\boldsymbol{X}_{II}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}_{II}^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = n\lambda_n \partial\rho(\hat{\boldsymbol{\beta}}_{II}), \tag{11}$$

$$\lambda_{\min}(\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\boldsymbol{X}_I) > n\lambda_n \kappa(\rho, \hat{\boldsymbol{\beta}}_I), \tag{12}$$

where $\hat{\boldsymbol{\beta}}_I = \{\hat{\beta}_{mj} | \hat{\beta}_{mj} \neq 0, \hat{\boldsymbol{\beta}}_{(j)} \neq \mathbf{0}\}$, $\hat{\boldsymbol{\beta}}_{II} = \{\hat{\beta}_{mj} | \hat{\beta}_{mj} = 0, \hat{\boldsymbol{\beta}}_{(j)} \neq \mathbf{0}\}$, $\boldsymbol{X}_I$ is the submatrix of $\boldsymbol{X}$ formed by columns in $I = \{(m, j) | \hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_I\}$, $\boldsymbol{X}_{II}$ is the submatrix of $\boldsymbol{X}$ formed by columns in $II = \{(m, j) | \hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{II}\}$, $\nabla\rho(\cdot)$ and $\partial\rho(\cdot)$ are the gradient and one subgradient of $\rho(\cdot)$ satisfying that

$$\nabla\rho(\hat{\beta}_{mj}) = \frac{1}{2}\mathrm{sgn}(\hat{\beta}_{mj})\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-1/2} \quad \textit{for } \hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_I,$$

$$\partial\rho(\hat{\beta}_{mj}) \in (-\frac{1}{2}\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-1/2}, \frac{1}{2}\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-1/2}) \quad \textit{for } \hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{II},$$

$\mathrm{sgn}(\hat{\beta}_{mj}) = 1$ *if* $\hat{\beta}_{mj} > 0$, $\mathrm{sgn}(\hat{\beta}_{mj}) = 0$ *if* $\hat{\beta}_{mj} = 0$, $\mathrm{sgn}(\hat{\beta}_{mj}) = -1$ *if* $\hat{\beta}_{mj} < 0$, *and* $\kappa(\rho, \hat{\boldsymbol{\beta}}_I) = \max_{\{j|\hat{\boldsymbol{\beta}}_j \neq \mathbf{0}\}} \frac{1}{4}\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-3/2}$ *denotes the local concavity of the penalty ρ. On the other hand, if $\hat{\boldsymbol{\beta}}$ is a solution to (9), then it must satisfy (10)–(12) with $>$ replaced by $\geq$ and $\partial\rho(\hat{\beta}_{mj}) \in [-\frac{1}{2}\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-1/2}, \frac{1}{2}\|\hat{\boldsymbol{\beta}}_{(j)}\|_1^{-1/2}]$ for $\hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{II}$.*

There is another KKT condition that $\boldsymbol{X}_{III}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}_{III}^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = n\lambda_n\partial\rho(\hat{\boldsymbol{\beta}}_{III})$, where $\hat{\boldsymbol{\beta}}_{III} = \{\hat{\beta}_{mj} | \hat{\beta}_{mj} = 0, \hat{\boldsymbol{\beta}}_{(j)} = \mathbf{0}\}$, $III = \{(m, j) | \hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{III}\}$ and $\partial\rho(\hat{\boldsymbol{\beta}}_{III})$ is one subgradient of $\rho(\cdot)$ at $\hat{\boldsymbol{\beta}}_{III}$. Since $\rho(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\sum_{m=1}^{M} |\beta_{mj}|)^{1/2}$, $\partial\rho(\hat{\beta}_{mj}) \in (-\infty, +\infty)$ for $\hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{III}$, that is, $\partial\rho(\hat{\beta}_{mj})$ could be any real number for $\hat{\beta}_{mj} \in \hat{\boldsymbol{\beta}}_{III}$. Thus, this KKT condition always holds.

We now consider the nonasymptotic weak oracle property of our proposed $\hat{\boldsymbol{\beta}}$ in (9), that is, with large probability, $\hat{\boldsymbol{\beta}}$ identifies the sparsity structure of the true parameter vector, and non-zero elements of $\hat{\boldsymbol{\beta}}$ are consistent in a rate slower than $\sqrt{n}$.

We need some notation. Let $\boldsymbol{\beta}^*$ be the true value of parameters and let $\boldsymbol{\theta}^* = \boldsymbol{X}\boldsymbol{\beta}^*$. With a slight abuse of notation, we denote $I = \{(m, j) | \beta^*_{mj} \neq 0, \boldsymbol{\beta}^*_{(j)} \neq \mathbf{0}\}$, $II = \{(m, j) | \beta^*_{mj} = 0, \boldsymbol{\beta}^*_{(j)} \neq \mathbf{0}\}$, and $III = \{(m, j) | \beta^*_{mj} = 0, \boldsymbol{\beta}^*_{(j)} = \mathbf{0}\}$, where $\boldsymbol{\beta}^*_{(j)} = (\beta^*_{1j}, \beta^*_{2j}, \ldots, \beta^*_{Mj})^{\mathrm{T}}$. Let $s_p = |I|$, the cardinality of set I, and $d_p = 2^{-1}\min\{|\beta^*_{mj}| : \beta^*_{mj} \in I\}$ be half of the minimum signal. In addition, we let $l_p = \min_{\{j:\boldsymbol{\beta}^*_{(j)} \neq \mathbf{0}\}}\|\boldsymbol{\beta}^*_{(j)}\|_1^{1/2}$ and $L_p = \max_{\{j:\boldsymbol{\beta}^*_{(j)} \neq \mathbf{0}\}}\|\boldsymbol{\beta}^*_{(j)}\|_1^{1/2}$.

We consider a fixed design matrix $\boldsymbol{X}$. Let $\boldsymbol{X}_I$, $\boldsymbol{X}_{II}$, and $\boldsymbol{X}_{III}$ denote columns of $\boldsymbol{X}$ with indices in I, II and III, respectively. We assume each block of $\boldsymbol{X}$ has been standardized to zero mean and unit variance. We fix M and let n, p diverge with $p \gg n$. For simplicity concern, we write several quantities in terms of an order of n. We let $\log p \asymp n^{1-2\alpha_p}$, $d_p \asymp n^{-\alpha_d}$ and $s_p \asymp n^{\alpha_s}$, where $\alpha_s > 0$, $\alpha_d > 0$ and $0 < \alpha_p < 1/2$ are constants. For any two sequence a_n, b_n, $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$.

To present our main result, we require the following conditions:

(C1) $0 < \alpha_s < \gamma < \alpha_p < 1/2$, where γ is shown as in (b) of Theorem 2.

(C2) $0 < \alpha_d < \min\{2(\alpha_p - \gamma), 2(\gamma - \alpha_s), \gamma\}$;

(C3) $\|[\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)\boldsymbol{X}_I]^{-1}\|_\infty = O(b_s n^{-1})$, where $b_s = o(\min\{n^{1/2-\gamma}/\sqrt{\log n}, n^{\gamma-\alpha_s}\})$;

(C4) $\|\boldsymbol{X}_{II}^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)\boldsymbol{X}_I[\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)\boldsymbol{X}_I]^{-1}\|_\infty \leq l_p/(2L_p)$;

(C5) $\max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{m,j}\lambda_{\max}\left[\boldsymbol{X}_I^{\mathrm{T}}\mathrm{diag}\{|\boldsymbol{X}_{mj}| \circ |\boldsymbol{\mu}''(\boldsymbol{X}_I\boldsymbol{\delta})|\}\boldsymbol{X}_I\right] = O(n)$,

where $\boldsymbol{X}_{mj}$ denotes the column of $\boldsymbol{X}$ corresponding to the jth variable in the mth dataset, the L_∞ norm of a matrix is the maximum of the L_1 norm of each row. $\mathcal{N}_0 = \{\boldsymbol{\delta} \in \mathcal{R}^{s_p} : \|\boldsymbol{\delta} - \boldsymbol{\beta}^*_I\|_\infty \leq d_p\}$, the derivative of $\boldsymbol{\mu}(\cdot)$ is taken componentwise, $\boldsymbol{\beta}^*_I$ is the subvector of $\boldsymbol{\beta}^*$ with indices in I and $\circ$ denotes the Hadamard (componentwise) product.

Condition (C1) is a requirement of the diverging rate of s_p and p. Condition (C2) is a requirement of the minimal signal of $\boldsymbol{\beta}^*$, which cannot be too small. Condition (C3) essentially requires that $\frac{1}{n}\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)\boldsymbol{X}_I$ should not be singular and we need a lower bound for its sup-norm. (C4) is a condition similar to irrepresentable condition of the lasso. Zhao and Yu (2006) showed that the irrepresentable condition is sufficient and almost necessary for the lasso to achieve model selection consistency. Since our penalty in (9) is an L_1 penalty within each gene, we need (C4). (C5) is a technical condition needed in the proof.

Theorem 2. *Under conditions (C1)–(C5), if we choose the penalty $\lambda_n \asymp n^{-\alpha_\lambda}$ satisfying*

$$2^{-1}\alpha_d + \gamma < \alpha_\lambda < \min\{2\gamma - \alpha_s, \alpha_p\}, \tag{13}$$

$\lambda_n b_s = o(n^{-\alpha_d/2-\gamma})$ *and* $\lambda_n\kappa_0 = o(\tau_0)$, *where* $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\kappa(\rho, \boldsymbol{\delta}) = \max_{\boldsymbol{\delta}\in\mathcal{N}_0}\max_{\{j|\boldsymbol{\delta}_{(j)}\neq\mathbf{0}\}} 4^{-1}\|\boldsymbol{\delta}_{(j)}\|_1^{-3/2}$ *for* $\boldsymbol{\delta}_{(j)} = (\delta_{1j}, \ldots, \delta_{Mj})^{\mathrm{T}}$ *and* $\tau_0 = \min_{\boldsymbol{\delta}\in\mathcal{N}_0}\lambda_{\min}(n^{-1}\boldsymbol{X}_I^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{X}_I\boldsymbol{\delta})\boldsymbol{X}_I)$, *then for sufficiently large n, with probability greater than* $1 - 2\{s_p n^{-1} + (Mp - s_p)\,\mathrm{e}^{-n^{1-2\alpha_p}\log n}\}$, *there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of (9), such that*

(a) (sparsity) $\hat{\boldsymbol{\beta}}_{II\cup III} = \mathbf{0}$;
(b) (L_∞ consistency) $\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}^*_I\|_\infty \leq n^{-\gamma}$.

The convergence rate in (b) of Theorem 2 is slower than $\sqrt{n}$-rate, because the L_1 penalty in (9) cannot achieve $\sqrt{n}$-rate. The rate γ depends on α_s, α_p and α_d satisfying (C1)–(C2). In general, the smaller p and s_p are and the larger d_p is, the larger γ we can have and the faster $\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}^*_I\|_\infty$ vanishes. In the best case, the convergence rate could be very close to $\sqrt{n}$. If we employ the sign constraint as in (6), then by adding another natural condition that the non-zero elements of $\boldsymbol{\beta}^*$ have the same sign, our main result (Theorem 2) still holds.

5. Numerical Results

5.1. *Simulations*

Simulation studies are performed to compare the finite sample performance of our proposed meta lasso with the separate lasso, stack lasso, group lasso, adaptively weighting (AW)

Table 2

Sensitivity and specificity of eight methods: presented values are the mean (standard error) over 100 simulations, sample size $n_m = 50$ for all 10 studies

		$\pi_0 = 0.9$	$\pi_0 = 0.5$	$\pi_0 = 0.2$
Meta lasso	Sensitivity	0.942 (0.071)	0.921 (0.075)	0.913 (0.097)
	Specificity	0.995 (0.001)	0.993 (0.002)	0.993 (0.004)
Separate lasso	Sensitivity	0.028 (0.019)	0.146 (0.064)	0.537 (0.170)
	Specificity	1.000 (0.000)	1.000 (0.000)	0.999 (0.000)
Stack lasso	Sensitivity	0.998 (0.011)	0.729 (0.166)	0.218 (0.232)
	Specificity	0.996 (0.002)	0.996 (0.002)	0.999 (0.001)
Group lasso	Sensitivity	0.820 (0.089)	0.801 (0.089)	0.774 (0.098)
	Specificity	0.999 (0.000)	0.996 (0.001)	0.995 (0.001)
AW	Sensitivity	0.787 (0.151)	0.780 (0.069)	0.800 (0.073)
	Specificity	0.997 (0.001)	0.995 (0.001)	0.992 (0.002)
Fisher	Sensitivity	0.806 (0.163)	0.703 (0.077)	0.578 (0.124)
	Specificity	0.999 (0.001)	0.997 (0.001)	0.996 (0.001)
FEM	Sensitivity	0.957 (0.050)	0.677 (0.032)	0.067 (0.103)
	Specificity	0.998 (0.001)	0.998 (0.001)	1.000 (0.000)
REM	Sensitivity	0.731 (0.130)	0.333 (0.215)	0.000 (0.000)
	Specificity	0.999 (0.001)	0.999 (0.001)	1.000 (0.000)

method by Li and Tseng (2011), Fisher's method, combination of effect sizes from Fixed Effects Model (FEM), and Random Effects Model (REM). In the simulation, the number of studies $M = 10$, sample size $n_m = 50$ for $m = 1, \ldots, 10$, and number of genes $p = 1{,}000$. The gene expression values $\boldsymbol{x}_{mi}$ are i.i.d. from standard normal distribution. The responses y_{mi}'s are generated from a logistic model by $\Pr(y_{mi} = 1|\boldsymbol{x}_{mi}) = \exp(\boldsymbol{x}_{mi}^{\mathrm{T}}\boldsymbol{\beta}_m^*)/[1 + \exp(\boldsymbol{x}_{mi}^{\mathrm{T}}\boldsymbol{\beta}_m^*)]$, where $\boldsymbol{\beta}_m^* = (\beta_{m1}^*, \beta_{m2}^*, \ldots, \beta_{mp}^*)^{\mathrm{T}}$ and the intercept term $\beta_{m0}^* = 0$. To allow possible data heterogeneity, we let $\beta_{mj}^* = z_{mj}b_{mj}$ for $m = 1, \ldots, M$, $j = 1, \ldots, 10$, $\beta_{mj}^* = 0$ otherwise, where z_{mj} are i.i.d. from $N(3, 0.5^2)$ and b_{mj} are i.i.d. from Bernoulli(π_0). This means that for each of the first 10 genes, in each dataset, it is important with probability π_0 and unimportant with probability $1 - \pi_0$. If the gene is important in a dataset, its effect is generated from $N(3, 0.5^2)$. We consider three values for π_0: 0.2, 0.5, and 0.9 to investigate different levels of heterogeneity among datasets. For each case we run 100 replicates.

The variable selection performance of the eight methods are evaluated using selection sensitivity and specificity. For lasso-type methods, sensitivity is the proportion of non-zero β_{mj}^*'s that are correctly estimated as non-zero and the specificity is the proportion of zero β_{mj}^*'s that are correctly estimated as zero. In particular, for the stack lasso, the estimated effects of genes are the same among all studies, that is, $\hat{\beta}_{1j} = \cdots = \hat{\beta}_{Mj}$, for $j = 1, \ldots, p$. For the CSS methods (AW, Fisher, FEM, and REM), the list of genes being identified is regarded to be significant among all studies. Sensitivity and specificity are defined accordingly.

For the meta lasso, the tuning parameters are selected by minimizing the BIC:

$$\mathrm{BIC}(\lambda) = \sum_{m=1}^{M} \left\{-2\ell_m(\hat{\boldsymbol{\beta}}_{m,\lambda}) + s_m \log(n_m)\right\}, \qquad (14)$$

where $\hat{\boldsymbol{\beta}}_{m,\lambda}$ is the estimated coefficients in the mth dataset based on tuning parameter λ, s_m is the number of non-zero elements of $\{\hat{\boldsymbol{\beta}}_{m,\lambda}\}$, n_m is the sample size of the mth study and $\ell_m(\hat{\boldsymbol{\beta}}_{m,\lambda})$ is the log-likelihood with $\boldsymbol{\beta}_m$ being replaced by its estimate $\hat{\boldsymbol{\beta}}_{m,\lambda}$. Tuning parameters for separate lasso are chosen dataset by dataset by minimizing $\mathrm{BIC}(\lambda_m) = -2\ell_m(\hat{\boldsymbol{\beta}}_{m,\lambda_m}) + s_m \log(n_m)$, where $\hat{\boldsymbol{\beta}}_{m,\lambda_m}$ is the estimated coefficients based on mth dataset with tuning parameter λ_m and other notations have the same meaning as in (14). Tuning parameters for the stack lasso and group lasso are obtained by minimizing the BIC in (14) with $s_m \log(n_m)$ replaced by $s \log(\sum_{m=1}^{M} n_m)$, and $\hat{\boldsymbol{\beta}}_{m,\lambda}$ replaced by $\hat{\boldsymbol{\beta}}_{\lambda}$, the estimated coefficients in the stacked data based on tuning parameter λ, where s is the number of non-zero elements of $\hat{\boldsymbol{\beta}}_{\lambda}$.

The four CSS methods are implemented by Li and Tseng's R package `MetaDE`. The simulation results are summarized in Table 2. Among all eight methods, our proposed meta lasso has the superior performance. The separate lasso in general does not have enough power to identify important genes. Its ostensibly large sensitivity for $\pi_0 = 0.2$ is a mere fact that the number of true non-zero coefficients is very small in this case. The performance of the stack lasso changes dramatically with π_0. When $\pi_0 = 0.9$, the stack lasso performs best as expected, due to high level of homogeneity among the data sets. On the other hand, when $\pi_0 = 0.2$, the stack lasso method performs badly, due to high level of heterogeneity among the data sets. Due to its "all-in-all-out" property, the group lasso doesn't perform as well as our method, especially when data heterogeneity is strong (π_0 is small). In general, the four CSS approaches perform worse than our method. AW performs the best among the four CSS methods. Like the stack lasso, the FEM and REM suffer from the same heterogeneity issue.

For our proposed meta lasso, it takes around 10 minutes to run each simulation, which includes the time to find optimal tuning parameters by the BIC.

5.2. *Real-Data Analysis*

We present our analysis results for the motivational example in Section 2. Table 1 gives some summaries of the five data

Table 3

Gene selections of eight methods in five cardiovascular studies

Selections by meta lasso and separate lasso in each dataset

Datasets	Meta lasso	Separate lasso
GSE12288	IFNA4 STAT1	None
GSE16561	STAT1 TLR8	CD14 CD86 CHUK MAPK11 MAPK14 PIK3CG PIK3R1 RAC1 STAT1 TLR2 TLR7 TLR8 TNF TRAF3
GSE20129	TLR8	None
GSE22255	IFNA4	None
GSE28829	IFNA4 STAT1 TLR8	CD14 IFNAR2 IRF5 MAPK9

Selections by other methods in all datasets

Method	Gene list
Stack lasso	None
Group lasso	CD86 FOS IFNAR2 IKBKE MAPK14 PIK3CA STAT1 TLR2 TLR8
AW	AKT1 AKT3 CASP8 CCL5 CD14 CD40 CD80 CD86 CHUK FADD FOS IFNAR1 IFNAR2 IKBKE IL1B IL8 IRAK1 IRF5 IRF7 JUN LBP LY96 MAP2K3 MAP2K4 MAP2K7 MAP3K7 MAP3K8 MAPK1 MAPK11 MAPK13 MAPK14 MAPK9 MYD88 PIK3CA PIK3CD PIK3CG PIK3R1 PIK3R5 RAC1 SPP1 STAT1 TBK1 TLR1 TLR2 TLR4 TLR5 TLR6 TLR7 TLR8 TNF TRAF3 TRAF6
Fisher	AKT1 AKT3 CASP8 CCL5 CD14 CD40 CD80 CD86 CHUK FADD FOS IFNAR1 IFNAR2 IKBKE IL1B IL8 IRAK1 IRF5 IRF7 JUN LBP LY96 MAP2K3 MAP2K4 MAP2K7 MAP3K7 MAP3K8 MAPK1 MAPK11 MAPK13 MAPK14 MAPK9 MYD88 PIK3CA PIK3CB PIK3CD PIK3CG PIK3R1 PIK3R5 RAC1 SPP1 STAT1 TBK1 TLR1 TLR2 TLR4 TLR5 TLR6 TLR7 TLR8 TNF TRAF3
FEM	CD86 IFNA4 STAT1 TLR2 TLR4 TLR5 TLR7 TLR8
REM	None

sets. For microarray data (GSE12288, 22255, 28829) generated from the Affymetrix platform, the RMA algorithm was used for data normalization. For Illumina data (GSE16561, 20129), quantile normalization procedure was applied. There are 88 common genes in all five studies and our analysis is based on those 88 genes. We coded control group as 0 and case group as 1. We applied aforementioned eight methods to select important genes, with the optimal tuning parameters chosen by the BIC as discussed above. Table 3 gives the names of genes selected in each data set. Our method identified Interferon Alpha-4 (IFNA4), Signal Transducer and Activator of Transcription 1 (STAT1), and Toll-Like Receptor 8 (TLR8) as important genes that may affect the atherogenesis. The separate lasso resulted in disparate recommendation of genes by over-emphasizing data heterogeneity: four genes were recommended from GSE28829; fourteen genes selected from GSE16561 with only one gene (CD14) that was also selected in GSE28829; and none selected from the other three data sets. The stack lasso failed to find any genes by over-emphasizing data homogeneity.

Several TLR family members, especially TLR2 and TLR4, have been well studied for their role in atherogenesis and autoimmune diseases, whereas the role of TLR8 in vascular diseases is less known due to the lack of a nonfunctional TLR8 in mice, making in vivo studies difficult (Diebold, 2008). However, the biological and clinical significance of TLR8 in immune and inflammatory response is emerging. For example, TLR8 expressions are associated with poor outcome and greater inflammatory response in patients experiencing acute ischemic stroke (Brea et al., 2011). The fact that our method identifies TLR8 as the sole TLR associated with various vascular diseases suggests its underappreciated role in modulating inflammatory signals. Once activated, pattern recognition receptors can lead to the production of type I interferons (IFNs), which consequently induces STAT1 phosphorylation (Dempoya et al., 2012), indicating the importance of STAT1 in maintaining very tight regulation of the innate immune system. Interferon (IFN)-induced Janus kinase (Jak)/Signal Transducer and Activator of Transcription (STAT) pathway has been known for its importance in controlling immune responses. Therefore, our finding of the two additional genes, INFA4 and STAT1, may provide a clue regarding major molecules in the TLR-IFN-STAT signaling cascade that mediates atherosclerotic and inflammatory process.

The three genes selected by our method are also being identified by other methods. In fact, STAT1 and TLR8 are identified by group lasso, AW, Fisher's method, FEM, and the separate lasso in dataset GSE16561. In addition, IFNA4 is also identified by FEM.

Gene CD14 is selected by separate lasso in studies GSE16561 and GSE28829, but not by meta lasso. An explanation is that CD14 is correlated with some of STAT1, TLR8, and IFNA4 to a certain degree (see Appendix B in the Supplementary Materials).

Study GSE20129 has demographics different from other studies. Hence it is important to address data heterogeneity. To show the importance of having this study with the second largest sample size, Appendix B shows a further analysis with this study left out. Since TLR8 is a very informative gene and it is the only gene selected by meta lasso under GSE20129, leaving GSE20129 out results in a reduced signal of TLR8 and, thus, more genes are selected by meta lasso to compensate. This shows the importance of borrowing strength across a reasonable number of studies with data heterogeneity properly addressed.

6. Discussion

In this article, we proposed a meta lasso method for variable selection in meta-analysis with high-dimensional gene expression data. Through a hierarchical decomposition on regression coefficients, our method not only borrows strength across multiple data sets to boost the power to identify important genes, but also takes into account data heterogeneity to relax the "all-in-or-all-out" rule. Under certain regularity conditions, we prove the gene selection consistency of our method. Simulation studies demonstrate that our method have good performances and are much better than other lasso methods. We applied the meta lasso to a cardiovascular study. The analysis results are clinically meaningful.

There could be other choices of penalty in the proposed maximization problem (5). When many genes are highly correlated to form clusters, the irrepresentable condition required for the validity of lasso is questionable and any lasso-type method may not work well. In such cases, we may replace the lasso penalty in (5) by the elastic-net penalty (Zou and Hastie, 2005) and select variables by maximizing

$$\sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m) - \lambda_g \alpha_1 \sum_{j=1}^{p} |g_j| - \lambda_g (1-\alpha_1) \sum_{j=1}^{p} g_j^2$$
$$-\lambda_\zeta \alpha_2 \sum_{j=1}^{p} \sum_{m=1}^{M} |\zeta_{mj}| - \lambda_\zeta (1-\alpha_2) \sum_{j=1}^{p} \sum_{m=1}^{M} \zeta_{mj}^2.$$

The weakness of the lasso may be alleviated and selection sensitivity may be improved when some genes are highly correlated. The two tuning parameters λ_g and λ_ζ can be reduced to one. Under the elastic-net penalty, it can be easily seen from Steps 2 to 4 of the algorithm that the problem is still concave at each iteration and can be efficiently solved. However, there are two more tuning parameters if the elastic-net penalty is applied.

Another solution to handle high correlations is to incorporate exogenous gene functional information. Our undergoing research extends meta lasso to pathway level, where a genome-wide study is possible and significant pathways could also be identified.

7. Supplementary Materials

Web Appendices A and B, referenced in Sections 3–5, are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank two referees, an associate editor, and an editor for their useful comments and suggestions. The research efforts were partly supported by the University of Wisconsin Carbone Cancer Center support grant, NIH/NIA P30 CA014520 (for Drs. Quefeng Li and Menggang Yu) and by a development fund from the University of Wisconsin Carbone Cancer Center (for Dr. Menggang Yu).

References

Askland, K., Read, C., and Moore, J. (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human Genetics* **125**, 63–79.

Barr, T. L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A., Matarin, M. (2010). Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology* **75**, 1009–1014.

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., Yeager, M., Chung, C. C., Chanock, S. J., and Chatterjee, N. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics* **90**, 821–835.

Brea, D., Sobrino, T., Rodriguez-Yanez, M., Ramos-Cabrer, P., Agulla, J., Rodriguez-Gonzalez, R., et al. (2011). Toll-like receptors 7 and 8 expression is associated with poor outcome and greater inflammatory response in acute ischemic stroke. *Clinical Immunology* **139**, 193–198.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**, 84–90.

Daissormont, I. T., Christ, A., Temmerman, L., Sampedro Millares, S., Seijkens, T., Manca, M., et al. (2011). Plasmacytoid dendritic cells protect against atherosclerosis by tuning T-cell proliferation and activity. *Circulation Research* **109**, 1387–1395.

DeConde, R., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., Etzioni, R., et al. (2006). Combining results of microarray experiments: A rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* **5**, 1204.

Dempoya, J., Matsumiya, T., Imaizumi, T., Hayakari, R., Xing, F., et al. (2012). Double-stranded RNA induces biphasic STAT1 phosphorylation by both type I interferon (IFN)-dependent and type I IFN-independent pathways. *Journal of Virology* **86**, 12760–12769.

Diebold, S. S. (2008). Recognition of viral single-stranded RNA by Toll-like receptors. *Advanced Drug Delivery Reviews* **60**, 813–823.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **21**, 171–178.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions* **57**, 5467–5484.

Farber, C. R. (2013). Systems-level analysis of genome-wide association data. *G3 (Bethesda)* **3**, 119–129.

Ferguson, L. (2004). External validity, generalizability, and knowledge utilization. *Journal of Nursing Scholarship* **36**, 16–22.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1.

Grützmann, R., Boriss, H., Ammerpohl, O., Lüttges, J., Kalthoff, H., Schackert, H., et al. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* **24**, 5079–5088.

Han, B. and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics* **8**, e1002555.

Hansson, G. K. (2005). Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine* **352**, 1685–1695.

Huang, C. C., Liu, K., Pope, R. M., Du, P., Lin, S., Rajamannan, N. M., et al. (2011). Activated TLR signaling in atherosclerosis among women with lower Framingham risk score: The multi-ethnic study of atherosclerosis. *PLoS ONE* **6**, e21067.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339–355.

Iwasaki, A. and Medzhitov, R. (2004). Toll-like receptor control of the adaptive immune responses. *Nature Immunology* **5**, 987–995.

Kelley, R., DasMahapatra, P., Wang, J., Chen, W., Srinivasan, S., Fernandez, C., et al. (2011). Prevalence of atherosclerotic plaque in young and middle-aged asymptomatic individuals: The bogalusa heart study. *Southern Medical Journal* **104**, 803–808.

Krug, T., Gabriel, J. P., Taipa, R., Fonseca, B. V., Domingues-Montanari, S., Fernandez-Cadenas, I., et al. (2012). TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. *Journal of Cerebral Blood Flow & Metabolism* **32**, 1061–1072.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994–1019.

Libby, P. (2002). Inflammation in atherosclerosis. *Nature* **420**, 868–874.

Liu, F., Dunson, D., and Zou, F. (2011). High-dimensional variable selection in meta-analysis for censored data. *Biometrics* **67**, 504–12.

Ma, S., Huang, J., and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* **12**, 763–775.

Ma, S. and Jian, H. (2009). Regularized gene selection in cancer microarray meta-analysis. *BMC Bioinformatics* **10**, 1–12.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy. *The Lancet* **365**, 488–492.

Ntzani, E., Ioannidis, J., et al. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. *The Lancet* **362**, 1439.

Rhodes, D., Barrette, T., Rubin, M., Ghosh, D., and Chinnaiyan, A. (2002). Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* **62**, 4427–4433.

Rietzschel, E., De Buyzere, M., De Bacquer, D., Segers, P., Bekaert, S., De Backer, G., et al. (2006). Prevalence of atherosclerosis in middle-age: Results from the asklepios study in 2524 subjects free from overt cardiovascular disease. *Circulation* **114**, 378–382.

Sinnaeve, P. R., Donahue, M. P., Grass, P., Seo, D., Vonderscher, J., Chibout, S. D., et al. (2009). Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS ONE* **4**, e7037.

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

Torkamani, A., Topol, E. J., and Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**, 265–272.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* **40**, 3785–3799.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics* **81**, 1278–1283.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

Zintzaras, E. and Ioannidis, J. (2008). Meta-analysis for ranked discovery datasets: Theoretical framework and empirical demonstration for microarrays. *Computational Biology and Chemistry* **32**, 39–47.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Received August 2013. *Revised May* 2014. *Accepted May* 2014.