The Annals of Applied Statistics
2012, Vol. 6, No. 2, 497–520
DOI: 10.1214/11-AOAS526

A BAYESIAN MODEL AVERAGING APPROACH FOR OBSERVATIONAL GENE EXPRESSION STUDIES

BY XI KATHY ZHOU[1], FEI LIU[2] AND ANDREW J. DANNENBERG[3]

Weill Medical College of Cornell University, IBM Watson Research Center and Weill Medical College of Cornell University

Identifying differentially expressed (DE) genes associated with a sample characteristic is the primary objective of many microarray studies. As more and more studies are carried out with observational rather than well controlled experimental samples, it becomes important to evaluate and properly control the impact of sample heterogeneity on DE gene finding. Typical methods for identifying DE genes require ranking all the genes according to a preselected statistic based on a single model for two or more group comparisons, with or without adjustment for other covariates. Such single model approaches unavoidably result in model misspecification, which can lead to increased error due to bias for some genes and reduced efficiency for the others. We evaluated the impact of model misspecification from such approaches on detecting DE genes and identified parameters that affect the magnitude of impact. To properly control for sample heterogeneity and to provide a flexible and coherent framework for identifying simultaneously DE genes associated with a single or multiple sample characteristics and/or their interactions, we proposed a Bayesian model averaging approach which corrects the model misspecification by averaging over model space formed by all relevant covariates. An empirical approach is suggested for specifying prior model probabilities. We demonstrated through simulated microarray data that this approach resulted in improved performance in DE gene identification compared to the single model approaches. The flexibility of this approach is demonstrated through our analysis of data from two observational microarray studies.

1. Introduction. In recent years, as the rapid advances in biotechnology have markedly driven down the cost of microarray experiments, more and more large scale studies are carried out with heterogeneous samples, conveniently collected from subjects of different phenotypic characteristics and exposure histories. Such microarray studies are considered observational rather than experimental in nature [Potter (2003)] because the effects of confounding or correlation in covariates

Received June 2011; revised October 2011.

[1]Supported in part by a pilot award from the Clinical and Translation Science Center at Weill Cornell Medical College through the National Institute of Health (UL1-RR024996) and National Cancer Institute 1R21CA133260-01A2.

[2]Supported in part by the University of Missouri Research Board Award.

[3]Supported in part by funding from the Flight Attendants Medical Research Institute.

Key words and phrases. Bayesian model averaging, differential gene expression, microarray, observational study.

need to be properly handled. The sample complexity of such studies presents both opportunities and challenges to the analysis. Considering the differential gene expression studies, with multifaceted sample characteristics, one may explore more complex questions that are not possible with a more homogeneous sample such as the identification of differentially expressed (DE) genes associated with not just one sample characteristic but multiple characteristics and/or their interactions. For example, Boyle et al. (2010) investigated DE genes associated with smoking as well as smoking $\times$ gender interaction. In another study involving smokers and never smokers [Carolan et al. (2008)], microarray data were obtained for an unbalanced lung airway epithelium sample involving different tissue sites from subjects of different gender, age and ethnicity. An interesting question is to identify DE genes associated with either a single or multiple sample characteristics. To address these questions, one needs to quantify the strength of association between the expression of each gene and a set of sample characteristics. This differs from the gene set enrichment analysis [Efron and Tibshirani (2007); Efron (2010)], where the interest is to quantify the strength of association between a set of genes and a single sample characteristic. Direct application of currently available approaches to these questions does not provide a coherent solution and has clear limitations.

Methods for identifying DE genes are typically based on the ranking of statistics for between group differences associated with one sample characteristic (also known as a factor or a covariate), such as the t-, F-statistics, their nonparametric counterparts, their modified forms, or the Bayesian versions [see Jeffery, Higgins and Culhane (2006) for an excellent review of the various approaches]. These methods are suited for well controlled experiments. Their lack of control for confounding factors attracts increasing concern when applied to observational microarray studies [Potter (2003); Webb et al. (2007); Troester, Millikan and Perou (2009)]. With observational samples, the results may be confounded by a variety of sample characteristics, such as age, sex, genetic profile, exposure and treatment history, etc., which can lead to an increased number of false discoveries. Recent studies by Scheid and Spang (2007) and Leek and Storey (2007) suggested that hidden traces of unknown confounders may exist in DE gene studies and that ranking statistics need to be adjusted accordingly. To account for the effects of possible confounders, several approaches have been adapted from traditional observational studies and applied to microarray data [Smyth (2004); Hummel, Meister and Mansmann (2008)], including adjustment via multiple regression on known confounders or on surrogate variables for unknown confounders [Leek and Storey (2007)], or via a matched study design [Heller, Manduchi and Small (2009)].

Regardless of covariate adjustment, the aforementioned approaches rank the genes based on the effect sizes estimated using the same model, that is, a model with the same structure and same set of covariates, for all genes. Such a single model approach can be problematic for high-dimensional microarray data because different genes may be involved in different biological processes and their expression may be affected by different sets of covariates. More specifically, as shown in

Section 2, such an approach leads to model misspecification for a certain proportion of the genes and does not offer the same level of accuracy and efficiency for the effect size estimation for genes under investigation.

To avoid model misspecification in microarray data analysis, an ideal solution could be to apply different models to different sets of genes whereby each model contains only the set of covariates relevant to the genes it is describing. Identifying appropriate models for different sets of genes can be challenging because model uncertainty makes it difficult to identify a single best model. The Bayesian model averaging (BMA) approach offers an attractive alternative solution to this problem. Hoeting et al. (1999) provides a review of this approach in more traditional settings. In recent years, BMA approaches have been developed to handle various problems involving high throughput genetic data. For example, they were used to improve the assessment of candidate gene effects in the genome-wide association studies [Wu et al. (2010); Xu, Craiu and Sun (2011)] and to improve sample classification using gene expression microarray data [Yeung, Bumgarner and Raftery (2005)]. They have also been shown to improve the DE gene detection in settings where the microarray data involved two different distributional assumptions [Sebastiani, Xie and Ramoni (2006)] or were from different sources [Conlon, Song and Liu (2006)]. All these approaches are computationally expensive, as MCMC simulation is used to obtain estimates of model parameters. In this study, we propose a BMA approach for observational microarray studies based on linear regression models. It does not require MCMC simulations for estimating model parameters and offers a flexible and coherent framework to identify simultaneously DE genes associated with a single factor, multiple factors and/or their interactions.

In the next section we discuss limitations of single model approaches. In particular, we evaluate the impact of model misspecification from such approaches on DE gene finding. We also identify parameters that affect the magnitude of impact. In Section 3 we propose to find DE genes with a BMA approach that properly controls for sample heterogeneity and model uncertainty. In Section 4 we compare the performances of ranking statistics based on a simple model, a complex model and the BMA approach in simulated microarray studies. Section 5 concludes with applications of BMA to two existing microarray data sets. Our analysis supports the utility of the BMA method as a useful tool for capturing and quantifying the complex relationship between gene expression patterns and sample characteristics in observational microarray studies.

2. Limitation of the single model approaches. In this section we consider a general framework to describe gene expression variations in microarrays. Under this framework, we argue that the single model approaches to DE gene detection are overly simplified and subjected to the impact of model misspecification, for example, the omission of relevant covariates when a simple model is used and the inclusion of irrelevant covariates when a complex model is used. The consequences

of such model misspecification have been discussed extensively in the linear regression setting [Rao (1971, 1973); Rosenberg and Levy (1972)]. The implication of these results, however, has not been fully investigated in DE gene studies. In this section we evaluate the consequences of model misspecification from the single model approaches on performance measures often used in DE gene studies, including the false discovery rate (FDR) and sensitivity. We conclude this section with a summary of the main results.

2.1. *Notation*. We consider an observational microarray study which aims to identify DE genes associated with different values of a factor X_1, for example, cigarette smoking exposure. Expression profiles of J genes are obtained for n subjects with different values of X_1. Without loss of generality, a typical model for identifying X_1 related DE genes can be written as

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1i} + \cdots + \beta_{kj}x_{ki} + \eta_{ij} \tag{2.1}$$

or

$$y_{ij} = \alpha_{0j} + \alpha_{1j}x_{1i} + \cdots + \alpha_{kj}x_{ki} + \alpha_{(k+1)j}x_{(k+1)i} + \varepsilon_{ij}, \tag{2.2}$$

where y_{ij} is the normalized and typically log-transformed expression level of gene j in subject i; x_{1i} is the factor level for X_1 in subject i; $x_{2i}, \ldots, x_{ki}$ are levels for other factors, denoted by $X_2, \ldots, X_k$, that affect the expression of all the genes, for example, experimental parameters involved in the microarray experiments; $x_{(k+1)i}$ is the level of a potential confounding factor X_{k+1}, for example, gender, age, race, alcohol exposure, etc.; η_{ij} and ε_{ij} denote normally distributed random errors.

To identify DE genes related to X_1, p-values based on t-statistic of estimate of either β_{1j} or α_{1j} can be used as the ranking statistics. If model (2.1) is used, the relevant t-statistic for gene j is $t_{M_1,1j} = \hat{\beta}_{1j}/sd(\hat{\beta}_{1j})$, where $\hat{\beta}_{1j}$ is the least square estimate of β_{1j}. If model (2.2) is used, the t-statistic for gene j is calculated as $t_{M_2,1j} = \hat{\alpha}_{1j}/sd(\hat{\alpha}_{1j})$. It can be shown that the two statistics are related as follows:

$$t_{M_1,1j} = \frac{S_{1\cdot 23\cdots k}}{S_{1\cdot 23\cdots k+1}} t_{M_2,1j} + \frac{S^{-2}_{k+1\cdot 1\cdots k} b_{k+1,1} e^T_{k+1} Y_j}{sd(\hat{\beta}_{1j})}, \tag{2.3}$$

where $S^2_{k+1\cdot 1\cdots k}$, b_{k+1} and e_{k+1} are the residual sum of squares, least square parameter estimates and residual, respectively, from the following auxiliary regression equation:

$$X_{k+1} = Xb_{k+1} + e_{k+1}, \tag{2.4}$$

where $X = (X_1, \ldots, X_k)$. $S^2_{1\cdot 23\cdots k.k+1}$ is the residual sum of squares for the auxiliary regression with X_1 as the outcome and $X_2, \ldots, X_{k+1}$ as the covariates.

For an observational microarray study, such single model approach with or without covariate adjustment has an intrinsic limitation, that is, neither model can be the true model for all the genes. For the aforementioned hypothetical microarray

study, model (2.1) is the true model only for genes not related to X_{k+1} (X_{k+1} null genes, or M_1 genes), and model (2.2) is the true model only for genes related to X_{k+1} (X_{k+1} DE genes, or M_2 genes). Based on these considerations, a multi-model approach that uses p-values of $t_{M_1,1\cdot}$ to rank the M_1 genes and p-values of $t_{M_2,1\cdot}$ to rank the M_2 genes is preferable.

The performance difference between the single model and the multi-model approaches can be compared by utilizing the relationship between the two t-statistics. Let $F_1(t)$ and $F_2(t)$ be the density distributions of the ranking statistics $t_{M_1,1\cdot}$ and $t_{M_2,1\cdot}$, respectively. Under the multi-model approach, the density distribution of the ranking statistics can be written as

$$F(t) = (1-f)F_1(t) + fF_2(t),$$

where f is the proportion of M_2 genes. $F_1(t)$ and $F_2(t)$ can further be written as

$$F_1(t) = (1-p_1)F_{10}(t) + p_1F_{11}(t),$$
$$F_2(t) = (1-p_2)F_{20}(t) + p_2F_{21}(t),$$

where p_1 and p_2 are the proportions of DE genes in M_1 and M_2 genes, $F_{\cdot 0}(t)$ and $F_{\cdot 1}(t)$ are distributions of the test statistic for the null and DE genes, respectively. For a given cutoff $c > 0$, the false discovery rate and sensitivity can be calculated as

$$FDR(c) = \frac{(1-f)(1-p_1)[1-F_{10}(c)]}{(1-f)[1-F_1(c)] + f[1-F_2(c)]} + \frac{f(1-p_2)[1-F_{20}(c)]}{(1-f)[1-F_1(c)] + f[1-F_2(c)]} \tag{2.5}$$

and

$$S(c) = 2(1-f)p_1[1-F_{11}(c)] + 2fp_2[1-F_{21}(c)].$$

We discuss the impact of the two single model approaches on the FDR and sensitivity separately.

2.2. *Single model without covariate adjustment.* When model (2.1) is used, the FDR can be written as

$$FDR^{M_1}(c) = \frac{(1-f)(1-p_1)[1-F_{10}(c)]}{(1-f)[1-F_1(c)] + f[1-F_2^{M_1}(c)]} + \frac{f(1-p_2)[1-F_{20}^{M_1}(c)]}{(1-f)[1-F_1(c)] + f[1-F_2^{M_1}(c)]}.$$

The sensitivity can be written as

$$S^{M_1}(c) = 2(1-f)p_1[1-F_{11}(c)] + 2fp_2[1-F_{21}^{M_1}(c)].$$

Superscript M_1 is used to denote that the distribution of t-statistic is derived from model (2.1), which is misspecified for the M_2 genes because of omitting relevant covariate X_{k+1}.

Omission of relevant covariate leads to bias in the model parameter estimates [Rao (1971)]. Specifically, the bias can be written as

$$Bias(\hat{\beta}_{1j}) = E(S^{-2}_{k+1\cdot 1\cdots k} b_{k+1,1} e^T_{k+1} Y_j) = \alpha_{k+1,j} \cdot b_{k+1,1}, \tag{2.6}$$

where $b_{k+1,1.23\cdots k}$ is the least square estimate of the parameter associated with X_1 in the auxiliary regression (2.4). Therefore, we have for the M_2 gene j

$$E(t_{M_1,1j}) \approx \frac{S_{1\cdot 23\cdots k}}{S_{1\cdot 23\cdots k+1}} \left[E(t_{M_2,1j}) + \frac{b_{k+1,1}\alpha_{k+1}}{\sigma_{2j}/S_{1\cdot 23\cdots k+1}} \right].$$

It is known that $S^2_{1\cdot 23\cdots k.k+1} \le S^2_{1\cdot 23\cdots k}$.

For the M_2 DE genes, because $t_{M_1,1j}$ can be greater or less than $t_{M2,1j}$ depending on the values of α_1 and $Bias(\hat{\beta}_1)$, $F^{M_1}_{21}(t)$ is unlikely to be systematically different from $F_{21}(t)$ and results in great changes in sensitivity.

However, for the M_2 null genes, the above results indicate $E|t_{M_1,1j}| \ge E|t_{M_2,1j}|$, that is, the distribution of $t_{M_2,1j}$ for the M_2 null genes moves away from zero. Hence, $1 - F^{M_1}_{20}(c) \ge 1 - F_{20}(c)$. Let a and b be the denominator and numerator of $FDR(c)$ as written in (2.5), respectively. Let δ be the difference between the numerators of $FDR^{M_1}(c)$ and $FDR(c)$, that is,

$$\delta = f(1-p_2)\{[1 - F^{M_1}_{20}(c)] - [1 - F_{20}(c)]\},$$

and δ' be the difference between the denominators of the two FDRs,

$$\begin{aligned}\delta' = {} & f(1-p_2)\{[1 - F^{M_1}_{20}(c)] - [1 - F_{20}(c)]\} \\ & + fp_2\{[1 - F^{M_1}_{21}(c)] - [1 - F_{21}(c)]\}.\end{aligned}$$

As discussed above, $[1 - F^{M_1}_{21}(c)]$ is comparable to $[1 - F_{21}(c)]$ because the bias is unlikely to lead to systematic difference between $F^{M_1}_{21}(t)$ and $F_{21}(t)$. Additionally, p_2 generally is much smaller than $1 - p_2$ in microarrays. Therefore, $\delta' \approx \delta$ and $FDR^{M_1}(c)$ can be approximated by $(b+\delta)/(a+\delta)$. Since $(b+\delta)/(a+\delta) \ge b/a$ for any $a > b > 0$ and $\delta \ge 0$, this indicates $FDR^{M_1}(c) \ge FDR(c)$, that is, increased FDR with this single model approach.

2.3. *Single model with covariate adjustment.* When model (2.2) is used, the FDR and sensitivity at a given cutoff can be written as

$$\begin{aligned}FDR^{M_2}(c) = {} & \frac{(1-f)(1-p_1)[1 - F^{M_2}_{10}(c)]}{(1-f)[1 - F^{M_2}_{1}(c)] + f[1 - F_2(c)]} \\ & + \frac{f(1-p_2)[1 - F_{20}(c)]}{(1-f)[1 - F^{M_2}_{1}(c)] + f[1 - F_2(c)]}\end{aligned}$$

and

$$S^{M_2}(c) = 2(1-f)p_1[1 - F_{11}^{M_2}(c)] + 2fp_2[1 - F_{21}(c)],$$

due to the potential change in the distributions of test statistics for the M_1 genes. The relationship of the two t-statistics can be written as

$$t_{M_2,1j} = \frac{S_{1\cdot 23\cdots k+1}}{S_{1\cdot 23\cdots k}} t_{M_1,1j} + \frac{S_{k+1\cdot 1\cdots k}^{-2} b_{k+1,1} e_{k+1}^T Y_j}{sd(\hat{\alpha}_{1j})}.$$

It is known that, with the inclusion of an irrelevant covariate, model (2.2) does not result in a biased parameter estimate for the M_1 genes. However, since $sd(\hat{\beta}_{1j}) \le sd(\hat{\alpha}_{1j})$ in general, $E(|t_{M_1,1}|) \ge E(|t_{M_2,1}|)$ for M_1 DE genes. Therefore, the distribution $F_{11}^{M_2}(t)$ moves toward 0 and results in $S^{M_2}(c) \le S(c)$, that is, reduced sensitivity in detecting DE genes in M_1 genes. As $|t_{M_1,1}|$ in general is likely to be greater than $|t_{M_2,1}|$, F_{10}^{M2} also shrinks toward 0. It is likely that $FDR^{M_2}(c)$ will be comparable to $FDR^{M_1}(c)$. Hence, reduced sensitivity in detecting DE genes in M_1 genes will be the main consequence resulted from applying the complex model for all the genes.

2.4. *Summary.* The above results suggested that the single model approaches with or without covariate adjustment can lead to inferior performance. It is expected that the impact on FDR and sensitivity could be greater if more X_{k+1}-like covariates exist in the sample. These results will be further demonstrated in the simulation study. The above discussion also suggested that the performance for DE gene detection can be improved by applying the correct model for the right sets of genes. Yet, such knowledge is commonly not available beforehand. In the following section, we propose a BMA approach as a practical substitute for the multi-model approach for DE gene detection that takes into account both sample heterogeneity and model uncertainty.

3. A Bayesian model averaging approach. In this section we discuss an efficient Bayesian model averaging approach to identifying DE genes associated with a covariate of interest. The methodology proposed in this paper is closely related to methods discussed in Liang et al. (2008) and we largely follow their notation. Consider a series of possible models for describing the expression pattern of each gene. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ be a binary vector of length K, with each element indicating the inclusion status of the kth covariate in the model, that is,

$$\gamma_k = \begin{cases} 0, & \text{if } \beta_k = 0, \\ 1, & \text{if } \beta_k \neq 0. \end{cases}$$

Each model in the model space can then be labeled by $\boldsymbol{\gamma}$, namely, $\mathcal{M}_{\boldsymbol{\gamma}}$. For gene j, $j = 1, \ldots, J$, the model can be written as

$$\mathcal{M}_{\boldsymbol{\gamma} j} : \mathbf{Y}_j = \alpha_{\boldsymbol{\gamma} j} \mathbf{1}_n + \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma} j} + \mathrm{N}(\mathbf{0}, \phi_{\boldsymbol{\gamma} j}^{-1} \mathbf{I}_n),$$

where $\alpha_{\boldsymbol{\gamma} j}$ is the intercept term; $\mathbf{X}_{\boldsymbol{\gamma}}$ is the submatrix of $\mathbf{X}$ consisting of columns associated with nonzero γ_k; $\boldsymbol{\beta}_{\boldsymbol{\gamma} j}$ and $\phi_{\boldsymbol{\gamma} j}$ are parameters under this model.

The marginal posterior inclusion probability for variable X_k and gene j, is then defined as

$$P_{kj} = P(\gamma_{kj} \neq 0 | \mathbf{Y}_j) = \sum_{\boldsymbol{\gamma}} \mathbf{1}_{\gamma_{kj}=1} \times P(\mathcal{M}_{\boldsymbol{\gamma} j} | \mathbf{Y}_j), \tag{3.1}$$

which is the sum of posterior probabilities of all models that include the covariate of interest. It quantifies the strength of association between covariate X_k and the expression level of the jth gene and can be used to rank the DE genes.

The posterior model probability for $\mathcal{M}_{\boldsymbol{\gamma} j}$ can be calculated based on Bayes factors of pairs of models, for example,

$$P(\mathcal{M}_{\boldsymbol{\gamma} j} | \mathbf{Y}_j) = \frac{p(\mathcal{M}_{\boldsymbol{\gamma} j}) BF(\mathcal{M}_{\boldsymbol{\gamma} j} : \mathcal{M}_{\mathbf{0} j})}{\sum_{\boldsymbol{\gamma}'} p(\mathcal{M}_{\boldsymbol{\gamma}' j}) BF(\mathcal{M}_{\boldsymbol{\gamma}' j} : \mathcal{M}_{\mathbf{0} j})}, \tag{3.2}$$

where $p(\mathcal{M}_{\boldsymbol{\gamma} j})$ is the prior model probability for genes measured in the microarray experiment and the Bayes factor $BF(\mathcal{M}_{\boldsymbol{\gamma} j} : \mathcal{M}_{\mathbf{0} j})$ is defined as

$$BF(\mathcal{M}_{\boldsymbol{\gamma} j} : \mathcal{M}_{\mathbf{0} j}) = \frac{f(\mathbf{Y}_j | \mathcal{M}_{\boldsymbol{\gamma} j})}{f(\mathbf{Y}_j | \mathcal{M}_{\mathbf{0} j})},$$

that is, the ratio of marginal likelihood under $\mathcal{M}_{\boldsymbol{\gamma} j}$ and the base model, $\mathcal{M}_{\mathbf{0} j}$. Here the null model (i.e., the model with only the intercept term) is used as the base model. For $\mathcal{M}_{\boldsymbol{\gamma} j}$, the marginal likelihood is obtained by integrating out the model parameters from the joint posterior probability

$$f(\mathbf{Y}_j | \mathcal{M}_{\boldsymbol{\gamma} j}) = \int f(\mathbf{Y}_j | \boldsymbol{\Theta}_{\boldsymbol{\gamma} j}) \pi(\boldsymbol{\Theta}_{\boldsymbol{\gamma} j} | \mathcal{M}_{\boldsymbol{\gamma} j}) \, d\boldsymbol{\Theta}_{\boldsymbol{\gamma} j},$$

where $\boldsymbol{\Theta}_{\boldsymbol{\gamma} j} = (\alpha_{\boldsymbol{\gamma} j}, \boldsymbol{\beta}_{\boldsymbol{\gamma} j}, \phi_{\boldsymbol{\gamma} j})$, and $\pi(\boldsymbol{\Theta}_{\boldsymbol{\gamma} j} | \mathcal{M}_{\boldsymbol{\gamma} j})$ is the prior of model parameters under $\mathcal{M}_{\boldsymbol{\gamma} j}$.

To determine the Bayes factor, proper priors, $\pi(\boldsymbol{\Theta}_{\boldsymbol{\gamma} j} | \mathcal{M}_{\boldsymbol{\gamma} j})$, are needed. We utilized the Zellner–Siow prior for model parameters [Zellner and Siow (1980)] in our study. Liang et al. (2008) have shown that this prior resolves several consistency issues associated with fixed g-priors while retaining several attractive properties such as adaptivity, good shrinkage properties, robustness to the misspecification of g and fast marginal likelihood calculation. When comparing two nested models as in our case, a flat prior is placed on common coefficients, $(\alpha_{\boldsymbol{\gamma} j}, \phi_{\boldsymbol{\gamma} j})$, where $\pi(\alpha_{\boldsymbol{\gamma} j}, \phi_{\boldsymbol{\gamma} j} | \mathcal{M}_{\boldsymbol{\gamma} j}) \propto 1/\phi_{\boldsymbol{\gamma} j}$, and a Cauchy prior on the remaining parameters, $\boldsymbol{\beta}_{\boldsymbol{\gamma} j}$. The multivariate Cauchy prior can then be represented as a mixture of g-priors with an Inv-gamma$(1/2, n/2)$ prior on g, that is,

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma} j} | \phi_{\boldsymbol{\gamma} j}, \mathcal{M}_{\boldsymbol{\gamma} j}) \propto \int N\left(\boldsymbol{\beta}_{\boldsymbol{\gamma} j} | \mathbf{0}, \frac{g}{\phi_{\boldsymbol{\gamma} j}} (\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right) \pi(g) \, dg,$$

with

$$\pi(g) = \frac{\sqrt{n/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.$$

The Bayes factor in equation (3.2) can be written in closed form as

$$BF(\mathcal{M}_{\boldsymbol{\gamma}j} : \mathcal{M}_{\mathbf{0}j}) = \int_0^\infty (1+g)^{(n-1-\rho_{\boldsymbol{\gamma}j})/2} \times [1 + (1 - R^2_{\boldsymbol{\gamma}j})g]^{-(n-1)/2} \pi(g)\, dg,$$

where $\rho_{\boldsymbol{\gamma}j}$ denotes the number of covariates included in $\mathcal{M}_{\boldsymbol{\gamma}j}$ and $R^2_{\boldsymbol{\gamma}j}$ is the ordinary coefficient of determination of this model. This quantity can be obtained through direct numerical integration or through the Laplace approximation.

In addition to the prior $\pi(\boldsymbol{\Theta}_{\boldsymbol{\gamma}j}|\mathcal{M}_{\boldsymbol{\gamma}j})$ on model parameters, one must also choose a prior on the models themselves, which relates directly to multiplicity. Scott and Berger (2010) discussed several prior model probability choices regarding their effects on multiplicity-control for multiple models in a conventional Bayesian model selection/averaging setting involving one outcome variable. With the high throughput data, typically, the prior model probabilities should reflect our prior belief about the distribution of the models among the genes in the transcriptome, which can be difficult to quantify. A uniform prior assumed equal probabilities of the models can be problematic when thousands of genes are evaluated simultaneously because it puts an unrealistically low weight to the null model. When the resulting posterior model probabilities are used to estimate the posterior expected FDR (*pe*FDR) [Newton et al. (2004)], great underestimation can occur [Sartor et al. (2006); Cao et al. (2009)]. Correctly estimating FDR under the Bayesian framework remains an active research field [Efron (2008)]. Recent discussions and attempts have largely been focused on statistics derived from single model approaches [Müller, Parmigiani and Rice (2007); Cao and Zhang (2010)]. In our case, proper control for multiplicity derived from multiple genes and multiple models becomes even more challenging.

We believe that the prior should lead to a reasonably well calibrated posterior model probability that measures the model's ability for describing the data. We propose an empirical approach to obtain estimates for the prior model probabilities, $p(\mathcal{M}_{\boldsymbol{\gamma}j})$, under the assumption that the prior probabilities of a given model are the same across genes, that is, $p(\mathcal{M}_{\boldsymbol{\gamma}j}) = p(\mathcal{M}_{\boldsymbol{\gamma}})$. We first estimate the proportion of DE genes described by a nonnull model $\boldsymbol{\gamma}$, $\omega_{\boldsymbol{\gamma}}$, using Bayes factors. Since $BF(\mathcal{M}_{\boldsymbol{\gamma}} : \mathcal{M}_{\mathbf{0}}) > c$, $c \geq 1$ suggests evidence against the null model [Kass and Raftery (1995)], we can estimate $\omega_{\boldsymbol{\gamma}}$ as follows:

$$\omega_{\boldsymbol{\gamma}} = \frac{1}{J} \sum_j \mathbf{1}_{[BF(\mathcal{M}_{\boldsymbol{\gamma}j} : \mathcal{M}_{\mathbf{0}j}) = \max(BF_j)]} \cdot \mathbf{1}_{[BF(\mathcal{M}_{\boldsymbol{\gamma}j} : \mathcal{M}_{\mathbf{0}j}) > c]},$$

where BF_j is a vector of null-based Bayes factors for gene j. Therefore, $\omega_{\boldsymbol{\gamma}}$ represents the proportion of genes for which model $\boldsymbol{\gamma}$ is the best model in terms of Bayes factors. Given that Bayes factors based on the Zellner–Siow prior are consistent for model selection whether or not the true model is null [Liang et al. (2008)], this estimator is a consistent estimator of the proportion of genes expressing in a pattern specified by the model. In our simulation studies, we found that fixing c at 1 resulted in $\omega_{\boldsymbol{\gamma}}$ being close to the truth in most settings. Second, we argue that if the prior model probabilities, $p(\mathcal{M}_{\boldsymbol{\gamma}})$, result in the equality between the overall peFDR under $\mathcal{M}_{\boldsymbol{\gamma}}$ and $1-\omega_{\boldsymbol{\gamma}}$, reasonable calibration of the posterior model probabilities can be achieved. This suggests the following relationship between $p(\mathcal{M}_{\boldsymbol{\gamma}})$ and $\omega_{\boldsymbol{\gamma}}$, that is,

$$\omega_{\boldsymbol{\gamma}} = \frac{1}{J}\sum_j \frac{BF(\mathcal{M}_{\boldsymbol{\gamma}j}:\mathcal{M}_{\mathbf{0}j})p(\mathcal{M}_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}'} BF(\mathcal{M}_{\boldsymbol{\gamma}'j}:\mathcal{M}_{\mathbf{0}j})p(\mathcal{M}_{\boldsymbol{\gamma}'})}.$$

Hence, $p(\mathcal{M}_{\boldsymbol{\gamma}})$ can be obtained by iteratively updating the following equation:

$$p^{(l)}(\mathcal{M}_{\boldsymbol{\gamma}}) = \frac{\sum_j \mathbf{1}_{[BF(\mathcal{M}_{\boldsymbol{\gamma}j}:\mathcal{M}_{\mathbf{0}j})=\max(BF_j)]}\cdot \mathbf{1}_{[BF(\mathcal{M}_{\boldsymbol{\gamma}j}:\mathcal{M}_{\mathbf{0}j})>c]}}{\sum_j [BF(\mathcal{M}_{\boldsymbol{\gamma}j}:\mathcal{M}_{\mathbf{0}j})/\sum_{\boldsymbol{\gamma}'} BF(\mathcal{M}_{\boldsymbol{\gamma}'j}:\mathcal{M}_{\mathbf{0}j})p^{(l-1)}(\mathcal{M}_{\boldsymbol{\gamma}'})]}$$

under the constraint $\sum_{\boldsymbol{\gamma}} p^{(l)}(\mathcal{M}_{\boldsymbol{\gamma}}) = 1$, where l denotes the iteration step. In our experience, 30 iterations were adequate to result in convergence. At present stage, theoretical justification for this prior choice for multiplicity control is still lacking. We resort to the simulation study to show that this prior choice led to improved performance in both the ranking of the genes and in direct FDR estimation compared with the uniform prior.

4. Simulation study. Simulation studies were designed to compare the single model approaches with and without covariate adjustment and the "gold standard" multiple-model approach with the correct covariate adjustments, as well as the performance of BMA over single-model approaches when a multi-model approach is appropriate. Bias and efficiency in each approach and sensitivity to the choice of prior on the set of models also will be discussed.

4.1. *Simulation of microarray data.* The microarray data were simulated to mimic an observational study for identifying genes associated with a binary variable, for example, the smoking status (s), in a sample with two potential confounders, gender (g) and heavy alcohol drinking (d) which are also binary. A detailed data generation scheme for the subject characteristics is provided in Zhou, Liu and Dannenberg (2012). Marginally, half of the subjects are assumed to be females, smokers or heavy drinkers. We also assume complex correlation among these covariates. First, s is correlated with both g and d. Specifically, in smokers, 75% are males and 80% are heavy drinkers; while in nonsmokers, 25% are males and 20% are heavy drinkers. Second, g is also correlated with d. Specifically, 75%

of male subjects are heavy drinkers, while 25% of females are heavy drinkers. Proportions of subjects in groups defined by categories of the covariates are provided in Zhou, Liu and Dannenberg (2012). Each microarray data set consists of the expression of 10,000 genes from n subjects. Gene expression for each subject was simulated based on the following model:

$$y_{ij} = \beta_{1j} s_i + \beta_{2j} g_i + \beta_{3j} d_i + \varepsilon_{ij},$$

where $\beta_{\cdot j}$ takes either 0 or nonzero values generated from normal distributions with variances generated following procedures similar to that described by Smyth (2004). Detailed procedures for generating the simulated microarray data are provided in Zhou, Liu and Dannenberg (2012). Each simulation setting was characterized by values of the following parameters: f_s, f_g and f_d, the proportion of genes affected by smoking (s), gender (g) or heavy drinking (d), respectively, and n, the sample size. Both moderate and relatively large sample sizes were considered, $n = 40$ and $n = 80$. For each setting, we simulated 10 microarray data sets. The reported results were averaged over the results obtained for each data set.

4.2. *Performance of the single model approaches*. In this section we compare the performances of three single model approaches that differed by covariate adjustment, that is, without covariate adjustment (SM_1), with adjustment to g and d (SM_2), and with adjustment to surrogate variables of g and d (*SVA*) [Leek and Storey (2007)], and that of the gold standard multi-model approach (*MM*) where the DE genes were fit with their respective true models, that is, the adjustment for g and/or d is applied only to genes truly affected by g and/or d. The sensitivity and FDR corresponding to the ranking statistic, p-value of s, were obtained for each method. To show the interplay of bias and efficiency on these performance measures, we also quantified the contribution to these measures from genes not associated with g and d, denoted as $g0d0$ genes.

Table 1 shows the performance difference between the single and multi-model approaches among top ranked genes identified with a p-value cutoff of 0.001. We can see that, as discussed in Section 2, SM_1 led to large increase in total FDR compared to *MM*. The magnitude of difference increased with the sample size, the proportion of the genes associated with the confounder and the number of the confounders. On the other hand, the difference in FDR contributed from the $g0d0$ genes remained small. Hence, the results suggested that bias in effect estimation among genes associated with the confounders was the main cause for the FDR increase. SM_2 and *SVA* showed slightly greater FDR compared to *MM*. This increase came mainly from $g0d0$ genes and suggested that the effects of the efficiency loss could have a negative impact on the total FDR, particularly in small sample size settings. A more notable limitation of SM_2 and *SVA* was the loss of sensitivity. Compared to *MM*, the magnitude of sensitivity loss increased slightly with sample size and the number of confounders.

TABLE 1

False discovery rate (FDR) and sensitivity (S), in %, among the top smoking related genes identified with a p-value cutoff of 0.001 *using ranking statistics based on the single model approach without covariate adjustment (SM_1), the single model approach with covariate adjustment (SM_2), the surrogate variable analysis approach (SVA) and the "gold standard" multi-model approach (MM). FDR and sensitivity arising from g0d0 genes (i.e., genes not associated with d and g) were included. Microarray data sets were simulated based on various settings defined by proportion of genes associated with each covariate: f_s, f_g, f_d, and the sample size n*

Methods	$n = 40$				$n = 80$			
	FDR_{g0d0}	FDR_{total}	S_{g0d0}	S_{total}	FDR_{g0d0}	FDR_{total}	S_{g0d0}	S_{total}
$f_s = 0.10$, $f_g = 0.05$, $f_d = 0$								
SM_1	4.2	6.5	14.1	14.9	2.3	8.2	28.5	30.2
SM_2	6.1	6.5	10.0	10.4	2.2	2.5	23.1	24.3
SVA	6.2	6.7	9.3	9.7	2.2	2.3	22.7	24.0
MM	4.4	4.8	14.1	14.5	2.4	2.6	28.5	29.7
$f_s = 0.05$, $f_g = 0.10$, $f_d = 0$								
SM_1	8.5	18.0	12.9	14.5	3.6	22.9	26.6	29.8
SM_2	10.5	11.7	9.4	10.6	5.1	5.7	21.0	23.4
SVA	10.3	11.6	9.0	10.2	5.4	6.2	20.7	23.1
MM	9.5	9.5	12.9	14.1	4.5	5.0	26.6	29.0
$f_s = 0.1$, $f_g = 0.05$, $f_d = 0.05$								
SM_1	4.1	8.5	13.4	15.0	2.6	12.8	26.4	29.5
SM_2	6.8	7.4	8.4	9.3	3.0	3.2	19.9	22.1
SVA	7.1	8.0	8.2	9.1	2.9	3.1	19.3	21.5
MM	4.4	4.9	13.4	14.3	3.0	3.1	26.4	28.8
$f_s = 0.05$, $f_g = 0.10$, $f_d = 0.10$								
SM_1	4.7	19.4	12.6	15.7	3.4	36.9	25.1	31.3
SM_2	9.9	12.6	7.9	9.6	4.5	6.1	18.4	22.6
SVA	10.7	13.1	7.7	9.4	4.6	6.0	18.0	22.4
MM	5.9	8.2	12.6	14.4	5.4	6.5	25.1	29.6

4.3. *Performance of the BMA approach.* In this section we examine the performance of the proposed BMA approach in comparison with the single model and the gold standard multi-model approaches. To evaluate the effects of prior choice on the performance of the BMA approach, we considered three prior model probability choices: the proposed empirical prior obtained using the two step approach (BMA_1), the uniform prior (BMA_2), and the true proportion of genes for each model (BMA_3). The posterior inclusion probability of s was used as the ranking statistics. The number of genes identified by each methods at 5% FDR were compared in Table 2. We can see that the *BMA* approaches had greater power in detecting DE genes compared to the *SM* approaches in general and the performance came close to that of the *MM* approach. In fact, in all the simulated settings, the *BMA* approaches, particularly BMA_1 and BMA_3, showed sensitivity close to the

TABLE 2
Power of different methods for identifying genes differentially expressed between smokers and nonsmokers at 5% FDR under different simulation settings

f_s	f_g	f_d	SM_1	SM_2	SVA	BMA_1	BMA_2	BMA_3	MM
$n = 40$									
0.10	0.05	0	139	96	83	145	119	149	155
0.10	0.05	0.05	126	80	72	137	124	137	150
0.05	0.10	0	42	31	31	51	46	52	52
0.05	0.10	0.10	46	30	26	56	49	57	58
$n = 80$									
0.10	0.05	0	286	294	290	346	335	344	356
0.10	0.05	0.05	239	250	248	317	308	318	334
0.05	0.10	0	94	113	110	147	135	146	152
0.05	0.10	0.10	82	108	106	145	138	142	147

MM approach for a given FDR threshold and greater than the single model approaches. Figure 1 showed the magnitude of performance difference in two representative settings. The *BMA* approaches appeared to be relatively insensitive to the choice of prior model probabilities for gene ranking.

Besides providing proper ranking of the gene, it is often useful to estimate the FDR of the finding and quantifying the proportion of DE genes in the transcriptome. Therefore, we also evaluated how well the FDR could be estimated based on the ranking statistics. For the p-value based approach, FDR and the proportion of DE genes were estimated using the approach by Storey (2002) and Storey

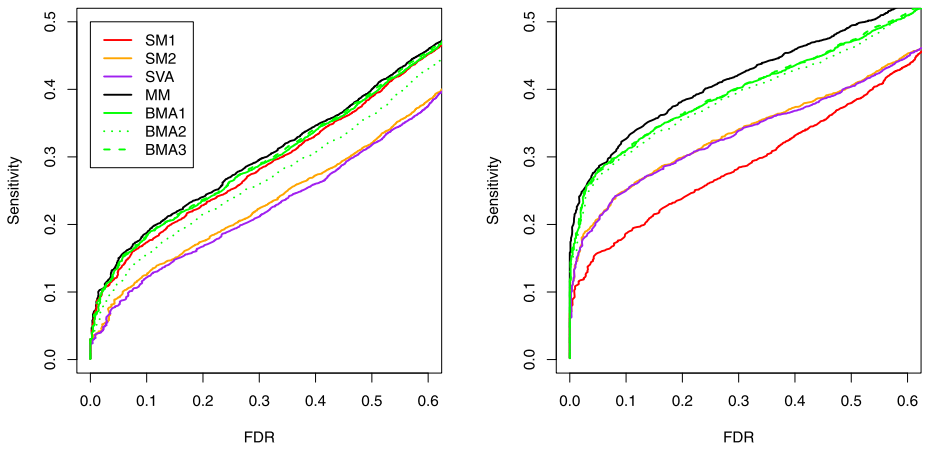

FIG. 1. *Sensitivity vs. FDR curves in two simulation settings.* (a) $f_s = 0.1$, $f_g = 0.05$, $f_d = 0$, $n = 40$. (b) $f_s = 0.05$, $f_g = 0.1$, $f_d = 0.1$, $n = 80$.

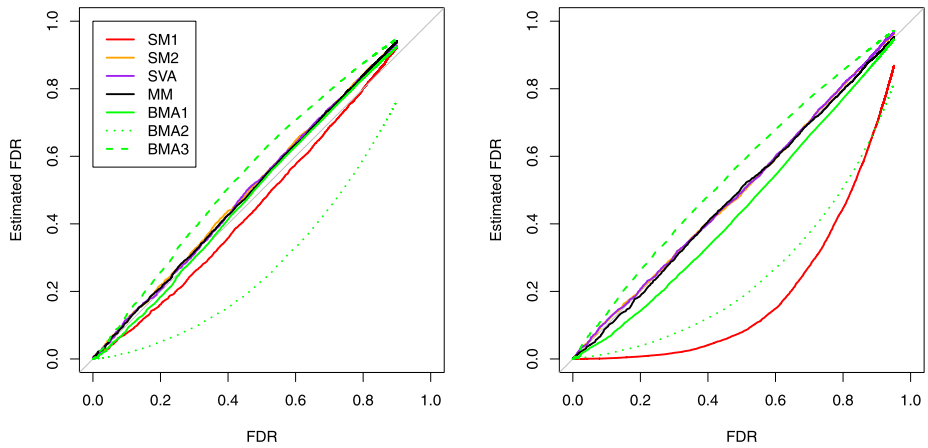

FIG. 2. *Estimated FDR vs. true FDR in two simulation settings.* (a) $f_s = 0.1$, $f_g = 0.05$, $f_d = 0$, $n = 40$. (b) $f_s = 0.05$, $f_g = 0.1$, $f_d = 0.1$, $n = 80$.

and Tibshirani (2003). For the Bayesian model averaging approach, the *pe*FDR was directly estimated based on the posterior inclusion probability [Newton et al. (2004)], that is,

$$peFDR_k(p) = \sum_j (1 - P_{kj}) \cdot \mathbf{1}_{[P_{kj} \le p]} / \sum_j \mathbf{1}_{[P_{kj} \le p]},$$

where $0 < p \le 1$ and P_{kj} is the posterior inclusion probability of variable k for gene j. Figure 2 shows the estimated FDR vs. the true FDR in two representative settings. We can see that using p-values from SM_1 in studies with confounder associated genes, the estimated FDR was smaller than the true FDR. The magnitude of underestimation increased with the sample size and the proportion of the confounder associated genes. On the other hand, the FDR estimated using p-values from SM_2, *SVA* or *MM* was very close to the true FDR. The accuracy of the *pe*FDR, as observed by other researchers, appeared to be sensitive to the prior choice. *pe*FDR obtained based on BMA_2, the Bayesian model averaging approach with uniform prior can greatly underestimate the FDR. *pe*FDR obtained based on BMA_1 showed improved accuracy in FDR estimation. The results from our simulation also suggest that the *pe*FDR based on BMA_1 are close to true FDR in all simulated settings. BMA_3 appeared to result in *pe*FDR that slightly overestimated the FDR. Level of sensitivity of the BMA_1 approach to the choice of c and model space misspecification can be found in Zhou, Liu and Dannenberg (2012).

We also carried out two additional sensitivity analyses related to the BMA approaches using the simulated microarray data [see Zhou, Liu and Dannenberg (2012) for detail]. First, we investigated the sensitivity of the performance of the empirical BMA approach to the choice of the cutoff c. The results suggest that

the BMA approach with empirical prior is relatively robust in gene ranking with respect to the value of c. Second, we investigated the performance of the BMA approach to the misspecification of model space, that is, omission of an important covariate d. As expected, there is a decrease in ranking performance, but the BMA approach still outperforms all the single model approaches. It is possible to avoid the performance loss due to omission of important covariates by introducing the surrogate variables [Leek and Storey (2007)] into the models. However, including the surrogate variables in the BMA approach is not a trivial extension due to model uncertainty, and it is definitely an interesting future research topic.

5. Application to the observational micorarray data sets. We applied the BMA approach to two smoking related observational microarray studies. Through the application, we intended to demonstrate the complex relationship between the gene expression pattern and sample characteristics and the flexibility of the BMA approach in capturing and quantifying such relation in a unified and coherent framework.

5.1. *Microarray study of airway epithelium samples*. The first data set (GSE10006) came from a study with a total of 87 current and never smokers [Carolan et al. (2008)]. The microarray analyses were carried out on airway epithelium samples from these subjects. The data were preprocessed with the Affymetrix MAS method. After excluding gene probe sets whose expression measurements were deemed absent or marginal among all subjects, the remaining data consisted of expression profiles of 44,085 probe sets of genes from the Affymetrix HGU133plus2 chip for each subject. Among these probe sets, 34,614 were annotated for probing the expression of 17,690 genes. About half of these genes were probed by multiple probes. To eliminate the potential dependence issue, average expression measurements were obtained for genes with multiple probe sets. We analyzed the expression data of the 17,690 genes from 60 healthy subjects. Individuals with known lung disease were excluded. Besides smoking status, information on age, gender, race and site of the tissue was available. The samples were heavily unbalanced, the proportion of smokers was greater in female participants than in males (86% vs. 57%), the proportion of large airway samples was slightly larger in females than in males (57% vs. 46%), and the proportion of caucasian participants was larger in females compared to males (43% vs. 37%).

With five covariates, a total of 2^5 models were included in the model space. Interaction terms were ignored. The BMA approach allowed for simultaneous assessment of the association between the gene expression and each of the sample characteristics, and straightforward estimation of both the total proportion of the DE genes in the transcriptome and the proportion of DE genes associated with each covariate based on Bayes factors. The application showed a complex picture of the expression pattern in the epithelium microarray study. A total of 69% of the genes were estimated to be differentially expressed. The estimated proportions

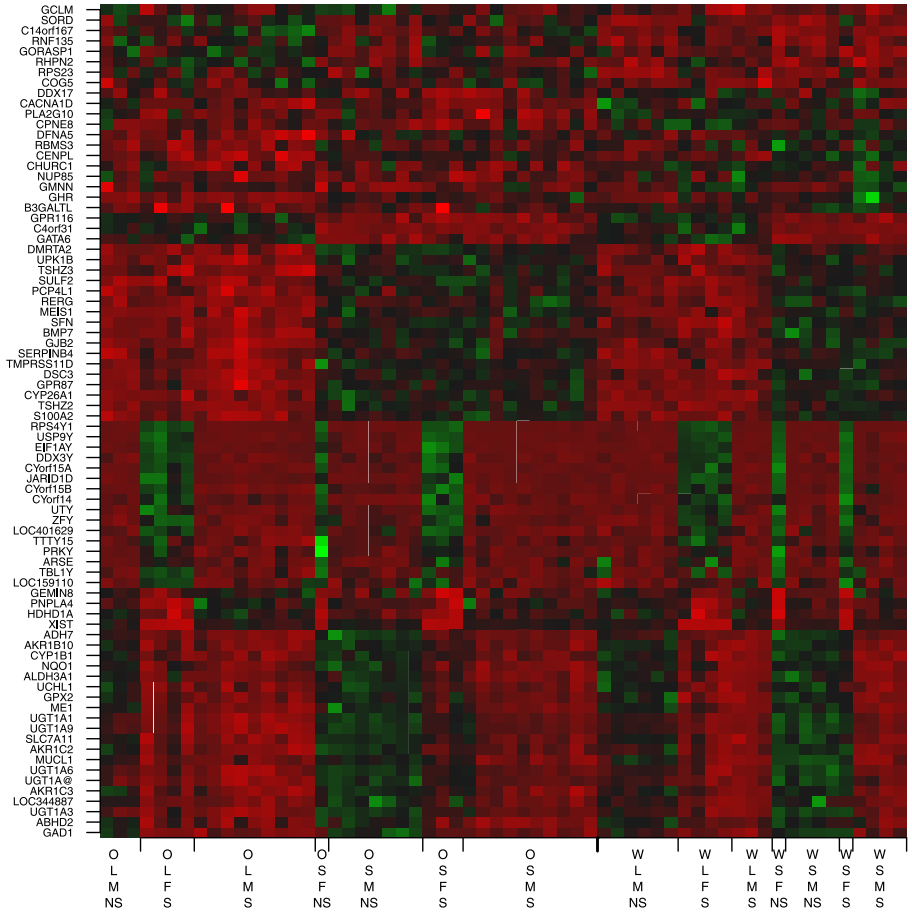

FIG. 3. *Gene expression intensities for the top* 20 *genes associated with each of the four covariates* (*smoking*, *gender*, *site and race*) *identified by using* BMA_1. *Labels along the x-axis show the characteristics of a sample subgroup. From top to bottom, the label represents categories of race* (*Others vs. White*; *O vs. W*), *site* (*Large airway vs. Small airway*; *L vs. S*), *gender* (*Male vs. Female*; *M vs. F*) *and smoking status* (*Non-Smoker vs. Smoker*; *NS vs. S*). *For example "OLMNS" indicates the subgroup with the following characteristics*: *Other races* (*i.e.*, *nonwhite*), *Large airway sample*, *Male*, *Non-Smoker.*

of DE genes for association with *smoking*, *site*, *gender*, *race* and *age* were 19%, 34%, 6%, 6% and 4%, respectively. By controlling the *pe*FDR at 5%, we identified a number of DE genes associated with *smoking* (928), *site* (3089), *gender* (73), *race* (33) and *age* (7). The complex expression patterns were illustrated in Figure 3 where we show the expression pattern of the top 20 genes associated with *smoking*, *gender*, *site* and *race*, respectively.

The results also revealed complex roles of some of these DE genes which were strongly associated with multiple sample characteristics. For example, among the

top 928 *smoking* related DE genes, 343, 18 and 6 of them were also identified as hits for association with tissue *site*, *gender* and *race*, respectively. Additionally, there were 25 genes identified as hits for association with three or more sample characteristics, mostly *smoking*, *site* and *gender*. The BMA approach allows for assessing jointly genes' association with multiple sample characteristics. For example, the joint posterior inclusion probability of *smoking*, *site* and *gender* can be obtained by summing over the posterior probabilities of models containing all three covariates. *pe*FDR can then be derived similarly using this posterior inclusion probability. The analysis identified 6 genes, NRARP, TMEM178, UGT1A@, UGT1A1, UGT1A3 and UGT1A6, as hits for joint association with the three characteristics at 5% *pe*FDR. The existence of such genes suggested a connection between tobacco smoking and the functions of these genes which were partly revealed through their association with the phenotype of the subjects from whom samples were obtained. Results from such analysis offer additional important information that is useful for generating new hypotheses and insights into the effects of tobacco smoke on the transcriptome.

As discussed in the previous sections, given the existence of genes associated with various sample characteristics, single model approaches were subjected to the effects of increased bias or reduced power in unbalanced study design. For the epithelium microarray data, we saw large differences in gene rankings derived from the BMA approach and the three single model approaches. Among the top 1000 smoking related DE genes identified by each method, the agreement was merely 19.7% among all four methods. Specifically, the *SVA* approach produced gene lists that were vastly different from the gene lists produced by the other approaches, where more than half of the top 1000 genes had ranks beyond 1000 by the other three methods [see the Venn diagram in Zhou, Liu and Dannenberg (2012)]. Careful examination of the gene lists produced by the *SVA* approach suggested possible effects of overfitting as the *SVA* approach adjusted for a total of 10 surrogate variables for each gene. The agreement was about 56% for the SM_1, SM_2 and BMA_1 approaches, that is, 56% were ranked within the top 1000 by all three methods. The agreement between BMA_1 and each of the single model approaches (SM_1, SM_2 and *SVA*) was 85%, 64% and 35%, respectively. These differences were driven by the genes whose expression patterns were not properly captured by the single model. The higher agreement between results from BMA_1 and SM_1 reflects the fact that a large proportion of the *smoking* related DE genes are associated with *smoking* only.

5.2. *Microarray study of oral mucosa samples*. The second data set came from a study with a total of 79 age and gender matched healthy smokers and never smokers [Boyle et al. (2010)]. The microarray analyses were carried out on oral mucosa samples obtained from these subjects through buccal biopsies. The preprocessed microarray data consisted of 24,103 probe sets of genes from the Affymetrix HGU133plus2 chip for each subject. Among these probe sets, 22,004

were annotated for probing the expression of 12,911 genes. About 43% of these genes were probed by multiple probe sets. To eliminate the potential dependence issue, average expression measurements again were obtained for these genes. The analysis was carried for the expression data of the 12,911 genes. For subjects recruited for this study, information regarding age, gender and smoking status was available.

The study samples were balanced in terms of gender between smokers and nonsmokers. Therefore, single model approaches with or without adjustment for gender would provide similar results. However, one interesting biological question was whether there were genes affected by smoking differently between the males and females. In this context, direct application of the single model approach could lead to confusing results. For example, at 5% estimated FDR, the single model without adjustment for the interaction term resulted in 944 hits for association with *smoking*, while the model adjusted for both *gender* and *gender* $\times$ *smoking* interaction led to the identification of only 1 gene as hits for association with *smoking* and no genes were identified as hits for *smoking* $\times$ *gender* interaction. Such large difference in DE gene assessment between different models is difficult to reconcile and interpret under the single model framework. Yet, such difference can be expected if there are genes associated with the interaction because the two variables, *smoking* and *smoking* $\times$ *gender* interaction, are correlated. Joint testing of the effects of *smoking* and *smoking* $\times$ *gender* interaction led to the identification of 311 DE genes with the likelihood ratio test. However, this method can not quantify the relative contribution from the two variables. We therefore applied the BMA approach to these data to illustrate the flexibility and usefulness of this approach to handle possible interaction effects.

In this application, the model space consists of a total of 16 models including the null model, three models with *smoking* and/or *gender* as main effects only and 12 models for different patterns that could arise from interaction between *smoking* and *gender*. For the oral mucosa data, our analysis estimated that about 22.5% of the genes are differentially expressed, in which about 12.3%, 1.5% and 8.6% were associated with *smoking*, *gender* and *smoking* $\times$ *gender* interaction, respectively. Controlling the *pe*FDR at 5%, our approach identified a total of 414 genes as hits associated with smoking through either the main effect, the interaction effect or both. Specifically, 222 of these genes were associated with *smoking* primarily through the main effect, 2 were associated with *smoking* primarily through the interaction effect, while for the rest of these genes various degrees of association were contributed from the interaction term.

By comparing the *smoking* related DE genes identified by the single model approaches and the BMA approach, we noted that the difference was mainly from genes that were over/under expressed in only one subgroup of the subjects, female smokers. Neither the model with *smoking* status as the only covariate nor the full model adjusted for both the *gender* and the *smoking* $\times$ *gender* interaction were able to adequately capture the strength of association for this group of genes and properly rank them due to either increased bias or decreased power. Table 3 showed

TABLE 3

Posterior inclusion probabilities of a single covariate, s (for smoking), g (for gender), or s × g interaction, and a composite of covariates, s and/or s × g interaction (denoted as s|s × g), obtained under BMA_1, for a list of DE genes associated with s primarily through s × g interaction. Also shown are the ranks of these genes based on the strength of association with the covariate/s under different methods ($\mathcal{R}^{method}_{covariate/s}$)

GSymbol	Cytoband	P_s	P_g	$P_{s\times g}$	$P_{s\|s\times g}$	$\mathcal{R}_s^{SM_1}$	$\mathcal{R}_s^{SM_2}$	$\mathcal{R}_s^{BMA_1}$	$\mathcal{R}_{s\times g}^{SM_2}$	$\mathcal{R}_{s\times g}^{BMA_1}$	$\mathcal{R}_{s\|s\times g}^{SM_2}$	$\mathcal{R}_{s\|s\times g}^{BMA_1}$
CEACAM7	19q13.2	0.042	0.003	0.969	0.998	205	3366	6703	96	1	66	39
CD177	19q13.2	0.035	0.006	0.933	0.962	1156	6912	8041	327	2	523	191
MARK1	1q41	0.061	0.004	0.928	0.985	485	9489	4967	6	3	83	122
GTF2A2	15q22.2	0.055	0.005	0.904	0.953	997	6777	5367	260	4	425	214
PLA2G2A	1p35	0.092	0.008	0.878	0.963	643	3677	3636	805	5	375	189
AKR1B10	7q33	0.062	0.008	0.875	0.931	1128	5915	4929	608	6	618	278
THYN1	11q25	0.020	0.033	0.869	0.885	3123	10,582	12,884	522	7	1583	384
BMS1	10q11.21	0.117	0.008	0.861	0.970	502	3142	3017	774	8	278	169
CLIC2	Xq28	0.079	0.029	0.858	0.934	910	3395	4095	1929	9	686	265
PRDX5	11q13	0.059	0.061	0.854	0.908	1266	3128	5065	3370	10	1004	331

the posterior inclusion probabilities and ranks based on different approaches for a few of these genes. A large difference in the rankings by different methods can be seen.

6. Discussion. In the past decade, microarray technology has greatly increased our ability to simultaneously interrogate the expression of tens of thousands of genes. Use of this technology has contributed to an improved understanding of the molecular basis of various diseases. As one of the primary tools for such studies, methods for finding DE genes have also been refined over time. Various approaches have been proposed to deal with multiple issues in microarray data. Yet, from the modeling perspective, many approaches have ignored sample heterogeneity, its impact on the analysis results, and the great opportunity it presents. Since Potter (2003) discussed the need for controlling bias and confounding in observational microarray studies, it has been increasingly recognized that the lack of control for sample heterogeneity could be a barrier to the reproducibility of the study findings. In two editorials [Webb et al. (2007); Troester, Millikan and Perou (2009)], improved data analysis methods and better study design have been considered crucial for advancing the field of cancer epidemiology with microarray technology. In particular, Troester, Millikan and Perou (2009) discussed the potential of model selection strategies in the process. Nevertheless, there remain obstacles to fully appreciate the effect of complex sample characteristics on DE gene detection and the value of improving upon current approaches.

In this paper, we proposed a novel concept for high throughput data analysis involving a heterogeneous sample, that is, a multi-model handling is intrinsically needed. We presented the theoretical framework that explains why basing inferences on a single model could be problematic in observational microarray studies. The problem arises from the inadequacy of using a single model to describe the complex expression pattern of genes among a heterogeneous sample, which can result in increased number of false discoveries due to bias when a simple model is used or increased random error due to reduced efficiency when a complex model is used. Such effects of model misspecification are hard to avoid because of the existence of genes being affected by different sets of sample characteristics and/or their interactions. We showed through simulation that the single model approaches have inferior performance in DE gene finding in comparison with a multi-model approach should we know the right model for the right set of genes. The magnitude of effects on false discovery depends on the study design, specific biological system and the mechanism underlying expression variation.

We proposed to use the BMA approach to improve our ability to identify DE genes. This approach utilizes the Zellner–Siow prior for model parameters. The consistency property of this prior is important, as it allows for obtaining a consistent estimate of the distribution of the genes in the model space using Bayes factors. Another choice could be the hyper-g/n prior proposed in Liang et al. (2008).

We proposed to use an iterative procedure to obtain the prior model probabilities so that the estimated distribution of the genes among the model space based on posterior model probabilities matches the estimate based on the Bayes factors. These prior choices allow the efficient computation of the Bayes factors and the posterior inclusion probabilities that does not rely on a MCMC simulation. Our simulation study demonstrated that this approach performed almost as well as the gold standard multi-model approach with true models and better than the single model approaches in gene ranking. The ranking performance was relatively insensitive to a wide range of choice for prior model probabilities. However, accuracy of the FDR directly estimated from the posterior model/inclusion probabilities was sensitive to the prior choice. Our simulation study showed that the proposed empirical prior model probability allowed for reasonably good calibration of posterior model/inclusion probabilities for multiplicity and the estimated FDR was close to the true FDR in settings with moderate to large sample size. In the rare case of a small study with a heterogeneous sample, care needs to be taken when using the empirical prior because the small sample size property of the Zellner–Siow prior is less certain. Nevertheless, it should be pointed out that multiplicity control in the Bayesian modeling framework remains a challenging and active research area. Further studies on the theoretical aspects of the prior choice for multiplicity control across the multiple genes and multiple models are needed. The current BMA approach is developed under the M-complete assumption, that is, the model space contains the true model. Should unknown confounders exist, it is possible to capture the latent confounding factors by introducing the surrogate variables [Leek and Storey (2007)]. We note, however, it would be unwise to directly incorporate the surrogate variables, currently constructed based on residuals derived from a single model fit of the data, into the proposed BMA approaches. Our work relies on the assumption of linear regression models with normal errors, which may be violated in practice. This calls for new approaches that are robust to the normality assumption, which is likely to be particularly useful for studies with small sample sizes. For the analysis of conventional data with one outcome variable, robust Bayesian model selection/averaging approaches have been suggested, for example, the approach by Gottardo and Raftery (2009). Extending such ideas to the observational microarray studies represents an interesting future direction.

Finally, through the application of the BMA approach to an observational mircoarray study with unbalanced study design and one with balanced study design, we showed that complex expression patterns did exist when study samples were heterogeneous. Previous research has demonstrated the complexities of underlying biological mechanisms for gene expression variation. Genes affected by several common factors, such as age [Tan et al. (2008)], gender [Delongchamp et al. (2005); Yang et al. (2006); Tan et al. (2008)], smoking [Spira et al. (2004)] and drinking alcohol [Lewohl et al. (2001)], have been found in different tissue samples. Our study showed that such complexity interfered with the DE gene detection. Notably, the BMA approach was able to avoid missing important genes whose

expression patterns were not adequately captured by a single model approach. As an added value, the BMA approach is found to be a flexible tool that allows for more comprehensive characterization of the association between gene expression and the characteristics of the subjects from whom the samples were obtained. All these can be done within a unified and coherent framework.

Acknowledgments. The authors thank Doctors Jaya Satagopan and Li-Xuan Qin at the Memorial Sloan-Kettering Cancer Center for helpful discussions. The authors are grateful to the Editor, the Associate Editor and four anonymous referees whose comments and suggestions greatly improved this article. *Conflict of Interest:* None declared.

SUPPLEMENTARY MATERIAL

Supplement to "A Bayesian model averaging approach for observational gene expression studies" (DOI: 10.1214/11-AOAS526SUPP; .pdf). Detailed description of the simulation setup and simulation procedure and additional results from the simulation study and application to the airway epithelium microarray study are provided.

REFERENCES

BOYLE, J. O., GUMUS, Z. H., KACKER, A., CHOKSI, V. L., JENNIFER, M. B., ZHOU, X. K., ANTE'S, R. K., HUGHES, D., DU, B., JUDSON, B. L., SUBBARAMAIAH, K. and DANNENBERG, A. J. (2010). Effects of cigarette smoke on the human oral mucosal transcritpome. *Cancer Prevention Reseach* **3** 266–278.

CAO, J. and ZHANG, S. (2010). Measuring statistical significance for full Bayesian methods in microarray analyses. *Bayesian Anal*. **5** 413–427. MR2719658

CAO, J., XIE, X.-J., ZHANG, S., WHITEHURST, A. and WHITE, M. A. (2009). Bayesian optimal discovery procedure for simultaneous significance testing. *BMC Bioinformatics* **10** 5.

CAROLAN, B. J., HARVEY, B. G., DE BISHNU, P., VANNI, H. and CRYSTAL, R. G. (2008). Decreased expression of Intelectin 1 in the human airway epithelium of smokers compared to nonsmokers. *Journal of Immunology* **181** 5760–5767.

CONLON, E. M., SONG, J. J. and LIU, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics* **7** 247.

DELONGCHAMP, R. R., VELASCO, C., DIAL, S. and HARRIS, A. J. (2005). Genome-wide estimation of gender differences in the gene expression of human livers: Statistical design and analysis. *BMC Bioinformatics* **6** Suppl 2 S13.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

EFRON, B. (2010). Sets of cases (Enrichment). In *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics Monographs* **1** 163–184. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. and TIBSHIRANI, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.* **1** 107–129. MR2393843

GOTTARDO, R. and RAFTERY, A. (2009). Bayesian robust transformation and variable selection: A unified approach. *Canad. J. Statist.* **37** 361–380. MR2547204

HELLER, R., MANDUCHI, E. and SMALL, D. S. (2009). Matching methods for observational microarray studies. *Bioinformatics* **25** 904–909.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

HUMMEL, M., MEISTER, R. and MANSMANN, U. (2008). GlobalANCOVA: Exploration and assessment of gene group effects. *Bioinformatics* **24** 78–85.

JEFFERY, I. B., HIGGINS, D. G. and CULHANE, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* **7** 359.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** 1724–1735.

LEWOHL, J. M., DODD, P. R., MAYFIELD, R. D. and HARRIS, R. A. (2001). Application of DNA microarrays to study human alcoholism. *J. Biomed. Sci.* **8** 28–36.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

MÜLLER, P., PARMIGIANI, G. and RICE, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics* 8 (J. M. Bernardo, M. Bayarri, J. Berger, et al., eds.). 349–370. Oxford Univ. Press, Oxford. MR2433200

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

POTTER, J. D. (2003). Epidemiology, cancer genetics and microarrays: Making correct inferences, using appropriate designs. *Trends Genet.* **19** 690–695.

RAO, P. (1971). Some notes on misspecification in multiple regressions. *Amer. Statist.* **25** 37–39.

RAO, P. (1973). Some notes on the errors-in-variables model. *Amer. Statist.* **27** 217–218. MR0348931

ROSENBERG, S. H. and LEVY, P. S. (1972). A characterization on misspecification in the general linear regression model. *Biometrics* **28** 1129–1133. MR0329145

SARTOR, M. A., TOMLINSON, C. R., WESSELKAMPER, S. C., SIVAGANESAN, S., LEIKAUF, G. D. and MEDVEDOVIC, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* **7** 538.

SCHEID, S. and SPANG, R. (2007). Compensating for unknown confounders in microarray data analysis using filtered permutations. *J. Comput. Biol.* **14** 669–681.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SEBASTIANI, P., XIE, H. and RAMONI, M. F. (2006). Bayesian analysis of comparative microarray experiments by model averaging. *Bayesian Anal.* **1** 707–732. MR2282204

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 3, 29 pp. (electronic). MR2101454

SPIRA, A., BEANE, J., SHAH, V., LIU, G., SCHEMBRI, F., YANG, X., PALMA, J. and BRODY, J. S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. USA* **101** 10143–10148.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445 (electronic). MR1994856

TAN, Q., ZHAO, J., LI, S., CHRISTIANSEN, L., KRUSE, T. A. and CHRISTENSEN, K. (2008). Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. *Genomics* **92** 94–100.

TROESTER, M. A., MILLIKAN, R. C. and PEROU, C. M. (2009). Microarrays and epidemiology: Ensuring the impact and accessibility of research findings. *Cancer Epidemiology, Biomarkers & Prevention* **18** 1–4.

WEBB, P. M., MERRITT, M. A., BOYLE, G. M. and GREEN, A. C. (2007). Microarrays and epidemiology: Not the beginning of the end but the end of the beginning. *Cancer Epidemiology, Biomarkers & Prevention* **16** 637–638.

WU, X. L., GIANOLA, D., ROSA, G. J. M. and WEIGEL, K. A. (2010). Bayesian model averaging for evaluation of candidate gene effects. *Genetica* **138** 395–407.

XU, L., CRAIU, R. V. and SUN, L. (2011). Bayesian methods to overcome the winner's curse in genetic studies. *Ann. Appl. Stat.* **5** 201–231. MR2810395

YANG, X., SCHADT, E. E., WANG, S., WANG, H., ARNOLD, A. P., INGRAM-DRAKE, L., DRAKE, T. A. and LUSIS, A. J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16** 995–1004.

YEUNG, K. Y., BUMGARNER, R. E. and RAFTERY, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21** 2394–2402.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics*: *Proceedings of the First International Meeting Held in Valencia* (*Spain*) (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. Valencia Univ. Press, Valencia.

ZHOU, X. K., LIU, F. and DANNENBERG, A. J. (2012). Supplement to "A Bayesian model averaging approach for observational gene expression studies." DOI:10.1214/11-AOAS526SUPP.

X. K. ZHOU
DIVISION OF BIOSTATISTICS AND EPIDEMIOLOGY
DEPARTMENT OF PUBLIC HEALTH
WEILL CORNELL CANCER CENTER
WEILL MEDICAL COLLEGE OF CORNELL UNIVERSITY
NEW YORK, NEW YORK 10065
USA
E-MAIL: kaz2004@med.cornell.edu

F. LIU
DEPARTMENT OF BUSINESS ANALYTICS AND MATHEMATICAL SCIENCES
IBM WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NEW YORK 10598
USA
E-MAIL: feiliu@us.ibm.com

A. J. DANNENBERG
DEPARTMENT OF MEDICINE
WEILL CORNELL CANCER CENTER
WEILL MEDICAL COLLEGE OF CORNELL UNIVERSITY
NEW YORK, NEW YORK 10065
USA
E-MAIL: ajdannen@med.cornell.edu