Bayesian Analysis (2006) **1**, Number 1, pp. 105–120

Bayesian Identification of Differential Gene Expression Induced by Metals in Human Bronchial Epithelial Cells

Leanna L. House* and Merlise A. Clyde† Yuh-Chin T. Huang‡

Abstract. The study of genetics continues to advance dramatically with the development of microarray technology. In light of the advancements, interesting statistical challenges have arisen. Given that only one observation can be made from each gene on a single array, statisticians are faced with three issues: analysis with more genes than arrays, separating true differential expression from noise, and multiple hypothesis testing for regulation. Within this study, we model the expression of 1185 genes simultaneously in response to five chemical constituents of particulate matter; arsenic, iron, nickel, vanadium, and zinc. Taking advantage of a hierarchical Bayesian mixture model with latent variables, we compare multiple treatments to a control and estimate noise across arrays without assuming equal treatment means for housekeeping genes. To account for model uncertainty and hyperparameter specification, model averaging, MCMC, and Rao-Blackwell estimation are utilized.

Keywords: Bayesian, latent variables, MCMC, differential expression, hierarchical model, microarray, macroarray, toxicology, model selection

1 Introduction

Documented by observed or in vitro studies, adverse effects occur when humans are exposed to high or chronic doses of heavy metals (Vahter et al. 2002; Clayton et al. 2002). The scope and/or statistical power of such experiments are limited by ethical issues concerning the number of people and the amount of exposure to toxic matter. With the recent development of microarray technology however, researchers may perform safe, controlled studies that target specific information pertaining to how, at the genetic level, toxins may inhibit or interrupt natural functionings of the human body (Chi et al. 2002). While most current toxicology research aims to assess gene expression when insulted at the population level, understanding the expression profile provided by one individual is imperative for further research in the development and utility of macro/microarrays.

*Institute of Statistics and Decision Sciences, Duke University, Durham, NC, http://www.isds.duke.edu/~house/

†Institute of Statistics and Decision Sciences, Duke University, Durham, NC, http://www.isds.duke.edu/~clyde/

‡National Health and Environmental Effects Research Laboratory, Environmental Protection Agency, Research Triangle Park, NC, huang.tony@epa.gov

Specifically, we aim to identify differentially expressed genes from one individual when exposed to either arsenic, iron, nickel, vanadium, or zinc. Given a non-replicated human profile consisting of 1185 spots, only one data point exists per gene for analyzing the effect of a given toxin (in comparison to a control). Without multiple arrays per treatment, fundamental statistical issues including sample size, dimensionality, separating signal from noise, and multiple testing arise. Most statistical methods developed to date cannot adequately address these issues (Yang and Speed 2002). Hence, we propose a hierarchical multiple response model with latent variables.

Given that multiple treatments were administered to each gene, we consider the gene-specific parameters to belong to an "experimental population" rather than an array specific population. We model the observed up or down gene regulation as a function of a 5x1 vector containing the true means for differential expression plus noise (resulting from experimental preparation, implementation, or random error) across the six arrays. The method by which we estimate precision differs significantly from current work in that we do not use pre-specified housekeeping genes. Rather, the hierarchical structure and latent variables induce an internal mechanism for normalizing the data.

Since our efforts are dually motivated by statistical and biological advancements, Section 2 provides a brief description of the experimental protocol. Section 3 describes a Bayesian framework that addresses a multiple treatment study design and deciphers noise from pollutant effect given a limited sample size. The results of the analysis are summarized in Section 4, where we identify which genes are differentially expressed and rank them according to their posterior probabilities of up and/or down regulation under all, none, or one of the pollutants. Section 5 concludes our efforts with a discussion summarizing the analysis and possible future directions.

2 Study design

Primary bronchial cells were obtained from one individual by bronchoscopy. Clontech Laboratories, Inc (Palo Alto, CA), utilizing 32P-labeled cDNA for hybridization, provided six gene array filters (Atlas Human cDNA Expression Arrays cat no. 7850) including one control and 5 treatment arrays. Each array measures the expression of 1185 genes; 9 housekeeping genes and 1176 genes of interest. Following standard procedures (http://www.clontech.com/techinfo/manuals/), average array-specific background levels were calculated and subtracted from the corresponding gene intensity levels. The treatment arrays profile the expression of genes when exposed to the following concentrations of metal ions: 50 mM zinc(II), 1 mM iron(II), 50 mM vanadium(IV), 25 mM arsenic(III), and 100 mM nickel(II).

3 Methods

3.1 Statistical Model

Consider a multiple response model that describes the intensity levels, after adjusting for background[1], observed on the six Clontech matrices. Let $\mathbf{Y}_g = [y_{g0}, y_{g01}..., y_{gP}]'$ denote the vector of log intensities, y_{gp}, for gene g and pollutant p where $g \in \{1, 2, ..., 1185\}$ and $p \in \{0, 1, ..., P\}$. The number of pollutants, P, equals 5 and $P = 0$ corresponds to the control. Conditional on the true gene log intensity levels, μ_{gp}, we assume that the observed expressions, $\mathbf{Y}_g$, are independent across arrays

$$\mathbf{Y}_g = \begin{bmatrix} y_{g0} \\ y_{g1} \\ \vdots \\ y_{gP} \end{bmatrix} = \begin{bmatrix} \mu_{g0} \\ \mu_{g1} \\ \vdots \\ \mu_{gP} \end{bmatrix} + \begin{bmatrix} e_{g0} \\ e_{g1} \\ \vdots \\ e_{gP} \end{bmatrix} \tag{1}$$

with independent, identically distributed measurement errors. As inferred from investigative plots and past microarray literature (Broberg 2002; Speed and Group 2000), the data appear log-normal and are transformed accordingly. Furthermore, we consider $\mathbf{Y}_g$ to belong to an "experiment-wide" population where the variance of $e_{gp} = 1/\phi$ remains constant within arrays and between arrays. In the Discussion section we reflect upon relaxing this assumption.

To focus on the difference between control and treatment, we transform and re-parameterize the previously stated model resulting in

$$\left[\begin{array}{c} S_g \\ \hline \mathbf{D}_g \end{array}\right] = \left[\begin{array}{c} S_g \\ \hline d_{g1} \\ d_{g2} \\ \vdots \\ d_{gP} \end{array}\right] = \left[\begin{array}{c} \alpha_g \\ \hline \tau_{g1} \\ \tau_{g2} \\ \vdots \\ \tau_{g5} \end{array}\right] + \left[\begin{array}{c} \varepsilon_{g0} \\ \hline \varepsilon_{g1} \\ \varepsilon_{g2} \\ \vdots \\ \varepsilon_{gP} \end{array}\right], \tag{2}$$

where $\mathbf{D}_g$ is a 5x1 vector of $d_{gp} = y_{gp} - y_{g0}$ and $S_g = \sum_{i=0}^{5} y_{gi}$. The parameter τ_{gp} refers to the true differential expression for gene g under pollutant p, whereas α_g represents the gene's true total expression across treatments. Because we consider the genes independent (conditional on μ_{gp} and ϕ), the re-parameterization induces a block structure in the covariance matrix across all genes. Each block corresponds to the covariance matrix for gene g equal to $\text{cov}(\boldsymbol{\varepsilon}_g)$ where

$$\boldsymbol{\varepsilon}_g \sim N_6\left(\left[\begin{array}{c} 0 \\ \hline \mathbf{0} \end{array}\right], \left[\begin{array}{c|c} 6 & \mathbf{0} \\ \hline \mathbf{0} & \Sigma_\epsilon \end{array}\right] \phi^{-1} \right) \quad \text{and} \quad \Sigma_\epsilon = \begin{bmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{bmatrix}. \tag{3}$$

[1]From this point forward, any discussion relating to the modeling of gene expression refers to measurements where the background has previously been subtracted.

Given only one observation per gene and pollutant, all model parameters are unidentifiable without further information. To separate noise, $\boldsymbol{\varepsilon}_{gp}$, from signal, τ_{gp}, we introduce a latent binary variable, δ_{gp}, as an indicator of whether pollutant p causes differential expression of gene, g. Precisely, δ_{gp} indicates whether or not τ_{gp} differs from zero. For example, in cases where differential expression does not occur, $\delta_{gp} = 0$, hence $\tau_{gp} = 0$. Whereas, if $\delta_{gp} \neq 0$ then the parameter, τ_{gp} does not equal zero and is worthy of further investigation/estimation. Notice, for each gene, there are 2^5 different expression profiles or *Latent Expression Models*, noted as $M_g = [\delta_{g1},\ \delta_{g2}, ..., \delta_{g5}]$. Using the binary expansion of values 0-31, we may list all possible model configurations for gene g. The vector [0 0 0 0 0], the null latent expression model (Model #0), corresponds to none of the pollutants causing differential expression, whereas [1 1 1 1 1], the full latent expression model (Model #31), refers to all of the pollutants causing differential expression.

In addition to maintaining model identifiability and flagging differentially expressed genes, $\boldsymbol{\delta}$ provides an internal mechanism for calculating precision. When $\delta_{gp} = 0$, the observed measurement from the macroarray, d_{gp}, reflects experimental noise rather than a possible treatment effect of pollutant p. Hence, all observations where $\delta_{gp} = 0$ contribute to the estimate of ϕ. This dramatically differs from common practice where estimates for array variability are derived from predetermined housekeeping genes, genes assumed to be unaffected by treatment. Our approach avoids controversial assumptions including the choice, number, and true immunity to treatment effects of housekeeping genes (Chi et al. 2002; Smyth et al. 2003), and uses the information gathered from housekeeping genes as though it were collected from genes of interest.

One approach for estimating ϕ, as well as $\boldsymbol{\tau}_g$ and α_g, entails selecting the profile with the highest likelihood, and completing maximum likelihood estimation. This method however, ignores the uncertainty of the expression profile. Instead, we extend the model by placing prior distributions on all uncertain quantities, including the latent expression model. In turn, we may state the posterior probabilities of $M_g = m$ and $\delta_{gp} = 1$. The following section describes the priors chosen and the process by which the posterior probabilities are calculated.

3.2 Prior Specification

Specifying subjective prior distributions, especially in situations with small sample sizes, is ideal (Berger and Pericchi 1997). However, the task of developing informative priors for parameters $\boldsymbol{\tau}$, $\boldsymbol{\alpha}$, and ϕ under all possible expression configurations, is daunting, at best. Thus, to avoid overwhelming the information gained from the data and to keep the number of incorporated hyperparameters to a minimum, we select objective prior distributions.

Often, improper prior distributions are chosen for model parameters in hopes of re-

maining "non-informative". However, with few exceptions, such default priors will skew the calculation of Bayes Factors by arbitrarily weighting one model over another. For example, improper distributions generally cause erroneous conclusions when placed on non-orthogonal elements (Kass and Raftery 1995). Model specific constants for parameters that do not occur in each model will not cancel when deriving Bayes Factors and posterior model probabilities. Hence, improper prior distributions will illegitimately weight models with and without such arbitrary constants. In our case, α_g is indeed independent of the remaining parameters, but ϕ is only orthogonal to those in the likelihood distribution and τ_{gp} does not occur under all latent expression models.

Group-theoretic invariance arguments, enable Jeffrey's independent prior to remain an acceptable choice for α and ϕ (Berger et al. 1998),

$$f(\alpha_g, \phi) \quad \propto \quad 1/\phi, \tag{4}$$

but cannot extend to τ_{gp}. Thus, we place a Zellner's g-prior on τ_g, conditional on M_g and ϕ,

$$f(\tau_g|\phi, \gamma, M_g) = \mathrm{N}_k(\mathbf{0}, \phi^{-1}\gamma\Sigma_{M_g}). \tag{5}$$

This choice is a proper distribution, preserves the correlation structure of the likelihood in the prior, and ensures that the distribution of $\boldsymbol{\tau}_g$, under different expression profiles, is conditionally compatible (Dawid and Lauritzen 2000). The parameter γ represents the constant in the g-prior and the covariance of τ_g, Σ_{M_g}, is determined by Σ_ϵ and $M_{gp} = [\delta_{gp}...\delta_{g1}]$. Since Σ_ϵ is uneffected by any permutation of its rows, we may write Σ_ϵ as a partitioned matrix,

$$\Sigma_\epsilon = \begin{bmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{bmatrix} = \begin{bmatrix} \Sigma_{\delta\delta} & \Sigma_{\delta(\delta)} \\ \Sigma_{(\delta)\delta} & \Sigma_{(\delta)(\delta)} \end{bmatrix},$$

where (δ) refers to the set of τ's in the vector $[\tau_{g1}\ \tau_{g2}\ ...\ \tau_{gP}]'$ that equal zero or where $\delta_{gp} = 0$, and δ refers to those that remain. Conditional on the set $\boldsymbol{\tau}_{g,(\delta)}$, the covariance of $\boldsymbol{\tau}_{g,\delta}$, denoted by $\Sigma_{\delta\delta.(\delta)}$, is $\Sigma_{\delta\delta} - \Sigma_{\delta(\delta)}(\Sigma_{(\delta)(\delta)})^{-1}\Sigma_{(\delta)\delta}$, and Σ_{M_g} equals

$$\Sigma_{M_g} = \begin{bmatrix} \Sigma_{\delta\delta.(\delta)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Notice the distribution in equation (5) is degenerate at 0 given the null model, $M_g = 0$.

The hyperparameter γ is often prespecified as a constant value; e.g. γ equals sample size. In this case, deciding upon an appropriate constant is unclear. Should γ equal the number of subjects, treatments or genes? Furthermore, using a large value for γ (which implies the specification of a diffuse prior on $\boldsymbol{\tau}$) would inevitably lead to the "Bartlett's Paradox" (Kass and Raftery 1995), favoring the null expression model, regardless of the

data. Motivated by Strawderman (1971), we use a Beta prior on $\frac{\gamma}{1+\gamma}$, the shrinkage factor, and suggest the proper prior,

$$f(\gamma) \quad = \quad \frac{1}{2}(1+\gamma)^{-3/2}\mathbf{1}_{(\gamma>0)}. \tag{6}$$

We complete the hierarchical model by specifying the distribution of M_g. We follow the suggestions of George and McCulloch (1993) for calculating prior expression profile probabilities. Let $\omega_p = \mathrm{P}[\delta_{.p} = 1]$ for $p = \{1, 2, ...P\}$) represent the prior probability that pollutant p causes differential expression, regardless of gene. Assuming expression across pollutants is conditionally independent, the prior probability of each expression profile is

$$P(M_g = m|\boldsymbol{\omega}) \quad = \quad \prod_{p=1}^{P} \omega_p^{\delta_{gp}}(1-\omega_p)^{1-\delta_{gp}}. \tag{7}$$

The prior probability of all expression profiles for the experiment, $\mathbf{M}$, equals the product of $P(M_g = m|\boldsymbol{\omega})$ across the 1185 genes of interest, $P(\mathbf{M}|\boldsymbol{\omega}) = \prod_{g=1}^{G} P(M_g = m|\boldsymbol{\omega})$. Additionally, we place a uniform prior on each ω_p.

Notice apriori most parameter distributions are conditionally independent of one another. However, a posteriori this is not the case. The next section describes in detail the steps taken to obtain the marginal distributions for each parameter.

3.3 Posterior Analysis

Provided the goal for our effort is to pin-point differentially expressed genes, the primary parameter of interest is $\mathbf{M}$, which encompasses $\delta_{g1}, \delta_{g2}, ..., \delta_{gP}$ for all genes. Thus, we formulate 32 Bayes Factors by comparing each model to the full model in order to simplify the calculation of the posterior probabilities of $M_g = m$. In using posterior model probabilities to assess expression rather than pure Bayes Factors, we avoid incoherence issues as described by Lavine and Schervish (1999) and directly quantify our degree of belief for observing significant up or down regulation to individual pollutants.

Analytical Integration

When dealing with hierarchical models one strategy for diminishing computation time and minimizing variability in parameter estimates, is to numerically integrate out parameters when feasible. Integrating the marginal distributions over α and $\boldsymbol{\tau}_g$, we derive the Bayes Factors for comparing different expression profiles, $M_g = m$, to the full expression profile, $M_g = 31$, $BF_{gm} \equiv f(\mathbf{D}_g, S_g|\gamma, \phi, M_g = m, \boldsymbol{\omega})/f(\mathbf{D}_g, S_g|\gamma, \phi, M_g = 31, \boldsymbol{\omega})$:

$$\text{if } m_g = 31 \text{ (full model)}, \quad BF_{gm} = 1$$

$$\text{if } m_g \in [0, 30], \quad BF_{gm} = (1+\gamma)^{(5-k_g)/2} \exp\left\{ -\frac{\phi}{2}\frac{\gamma}{1+\gamma}\mathbf{D}'_{g(\delta)}\Sigma^{-1}_{(\delta)(\delta)}\mathbf{D}_{g(\delta)} \right\}.$$

We eliminate the remaining parameters upon which the Bayes Factors are conditional and obtain $\mathrm{P}[M_g = m|\mathbf{D}, \mathbf{S}]$ using stochastic integration and Rao-Blackwell techniques.

Stochastic Integration

To estimate the marginal posterior distributions of ϕ, γ, $\mathbf{M}$, and $\boldsymbol{\omega}$, we stochastically integrate the full joint distribution using MCMC. The process is a straight forward Gibbs sampler (Gelfand and Smith 1990). Note that by analytically marginalizing over $\boldsymbol{\tau}$ we effectively avoid the problem of "changing dimensions" within a sampler. A detailed derivation of the full conditional for M_g is in the appendix. Rather than estimating $\mathrm{P}[M_g = m|\mathbf{D}_g, S_g]$ via Monte Carlo frequency from the MCMC, we derive Rao-Blackwellized estimators of $\mathrm{P}[M_g = m|\mathbf{D}_g, S_g]$ and $\mathrm{P}[\delta_{gp} = 1|\mathbf{D}_g, S_g]$. All other full conditionals follow from standard distribution theory: given all remaining parameters, ϕ and $\frac{1}{1+\gamma}$ come from Gamma distributions and ω has a Beta distribution. The explicit distributions are provided in the Appendix.

The next section summarizes the results after running the Markov Chain for 100,000 iterations, removing the first 5000 observations to account for burn-in and selecting every tenth observation to minimize autocorrelation within samples.

4 Results

Table 1 summarizes the posterior, empirical means and credible intervals for ϕ, γ, and $\boldsymbol{\omega}$. Incidentally, recall the question surrounding the prespecification of a fixed γ. The posterior mean of γ is clearly closer to the number of arrays than the number of genes, but does not equal either quantity.

Table 1: Mean and credible intervals for ϕ, γ, and $\boldsymbol{\omega}$

Parameter	Posterior Mean	Credible Interval
ϕ	4.338	(3.855, 4.902)
γ	7.750	(6.841, 8.734)
ω_1	0.788	(0.723, 0.851)
ω_2	0.263	(0.211, 0.318)
ω_3	0.407	(0.354, 0.463)
ω_4	0.466	(0.402, 0.533)
ω_5	0.566	(0.502, 0.630)

4.1 Relative Gene Behavior

One advantage for using the proposed method is that we may rank the genes according to the posterior probabilities of differential expression. Sorting the posterior means of P_{mg} and/or the $\mathrm{P}[\delta_{gp} = 1|\mathbf{D}, \mathbf{S}]$, provides an ordered list, where, quite simply, genes with high posterior probabilities of expression will rank above those with low posterior probabilities of expression. A histogram of the posterior probabilities displays the number of genes that fall within pre-specified intervals. Figure 1 graphs the posterior probabilities of the null ($P_{0,g}$) and full ($P_{31,g}$) expression models across all genes. From the plots, we see that only a few genes require the full model with probability 1, whereas no genes require the null model with probability greater than .25. Similarly, the effects of the individual pollutants are graphed in Figure 2. Arsenic causes a greater than .50 probability of differential expression a posteriori for all genes, whereas for approximately 40% of the genes, the posterior probability of differential expression when exposed to iron is less than .20.

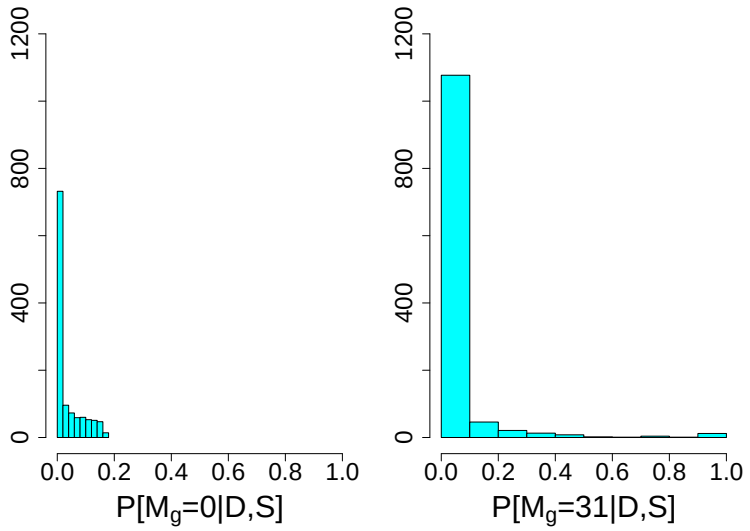

Figure 1: Frequency histograms of posterior model probabilities, across all genes; Model #0 (left): Null model indicating no differential expression and Model #31 (right): Full model indicating all pollutants cause differential expressions.

To obtain a general idea of the active locations on the macroarrays, we recreate a map of the slides and plot the posterior odds of expression. For this summary only, we depict the posterior odds of expression rather than the posterior probability of expression. For example, the scaling in Figure 3 is chosen to comply with Jeffreys' suggestion

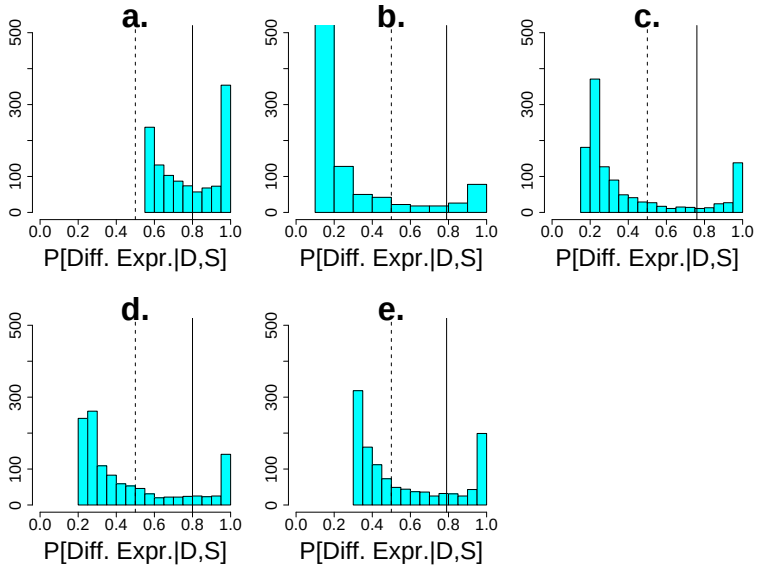

Figure 2: Frequency histograms of posterior probabilities of differential expression across all genes per pollutant, arsenic(a), iron(b), nickel(c), zinc(d) and vanadium(e); controlling the false discovery rate to .05, genes falling to the right of the solid line (β^*) are declared differentially expressed.

to the use of half- units on the log scale (Kass and Raftery 1995) to determine strong or weak evidence against a null hypothesis (Table 2). Figure 3 shows consistent activity in the bottom right quadrant of each slide.

Table 2: Purpose of scaling in Figure 3

Posterior Odds	Strength of Evidence
< 1	*None*
1-3.2	*Not worth more than a bare mention*
3.2-10	*Substantial*
10 - 100	*Strong*
> 100	*Decisive*

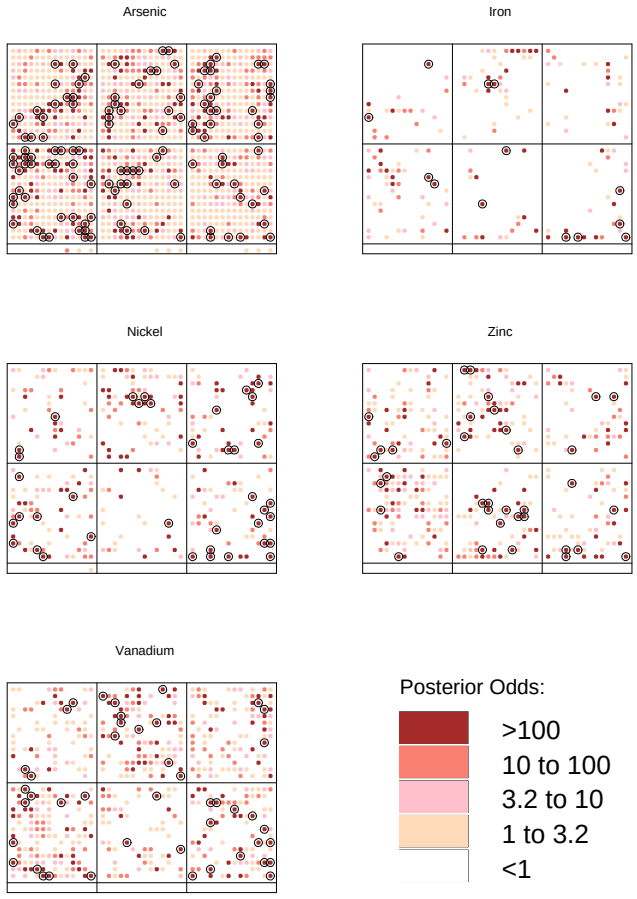

Figure 3: Map posterior odds of expression; circled genes indicate $\mathrm{P}[\delta_{gp} = 1|\mathbf{D}, \mathbf{S}] = 1$.

4.2 Declare Differential Expression

To declare a gene as significantly regulated, we must decide upon an appropriate cutoff β, so that if $\mathrm{P}[\delta_{gp} = 1|\mathbf{D}, \mathbf{S}] > \beta$ then gene g is differentially expressed under pollutant p. Provided an equal cost for false positives as false negatives, Bayesians often choose $\beta = 0.5$. In most cases, however, there exists a need to control the false discovery or non-discovery rate. Limiting the type II error to .05, we choose β^* as the smallest β such that the average posterior probability of acceptance equals .05 (Genovese and Wasserman 2002),

$$\boldsymbol{\beta}^* = \inf \left\{ \boldsymbol{\beta} : \frac{\sum_{i=1}^{n} I_{(\hat{\delta}_{gp} > \boldsymbol{\beta})} (1 - \hat{\delta}_{gp})}{\sum_{i=1}^{n} I_{(\hat{\delta}_{gp} > \boldsymbol{\beta})}} \leq .05 \right\}.$$

Using the posterior probabilities of expression, Figure 2 and Table 3 depict the effects of β^*. Table 3 lists the calculated values for β^* and documents the difference between the number of genes declared differentially expressed when controlling and not controlling the false discovery rate. Figure 2 displays this difference by plotting β and β^* on the histograms. Needless to say, across all pollutants, the number of declared genes dramatically decreases when β^* is utilized.

Table 3: Number of genes declared (Decl.) differentially expressed.

Pollutant	$\boldsymbol{\beta^*}$		$\boldsymbol{\beta = .5}$
	Value	**No. Decl.**	**No. Decl.**
Arsenic	0.80	552	1185
Iron	0.79	104	162
Nickel	0.76	211	297
Zinc	0.80	214	379
Vanadium	0.76	304	521

4.3 Up/Down Regulation

Thus far, we only addressed whether genes exhibit any change from the baseline, regardless of an increase or decrease in intensity. In order to calculate the posterior probability of up or down regulation, the posterior marginal distribution of τ is needed. Even though we analytically integrate out $\boldsymbol{\tau}$ to complete the MCMC, we may still use other sampled parameters at each MCMC iteration to calculate the Rao-Blackwellized estimates, $\hat{\tau}_{gp}$, and $P[\tau_{gp} > 0|\mathbf{D}, \mathbf{S}]$. Focusing on the genes that we declare differentially expressed in Table 3 (when using β^*), we plot by pollutant the distribution of $\hat{\boldsymbol{\tau}}$ across all genes and the posterior probabilities of up regulation in Figures 4 and 5 respectively. In Figure 4, we see most of the mass in each distribution lies outside the interval [log(1/2),log(2)]. This implies that the genes we flag as differentially expressed are centered around a two fold change or greater. Genes depicted on the right hand side of the plots in Figure 5 indicate a high probability of up regulation, whereas those that occur on the left hand side display a low probability of up regulation or high probability of down regulation. More genes are likely to experience down regulation than up when exposed to arsenic, zinc and vanadium.

5 Discussion

We propose a theoretical model and practical approach for assessing differential expression when only one data point exists per gene and pollutant. Given the extremely limited amount of data, all conclusions that we make reflect the person from whom

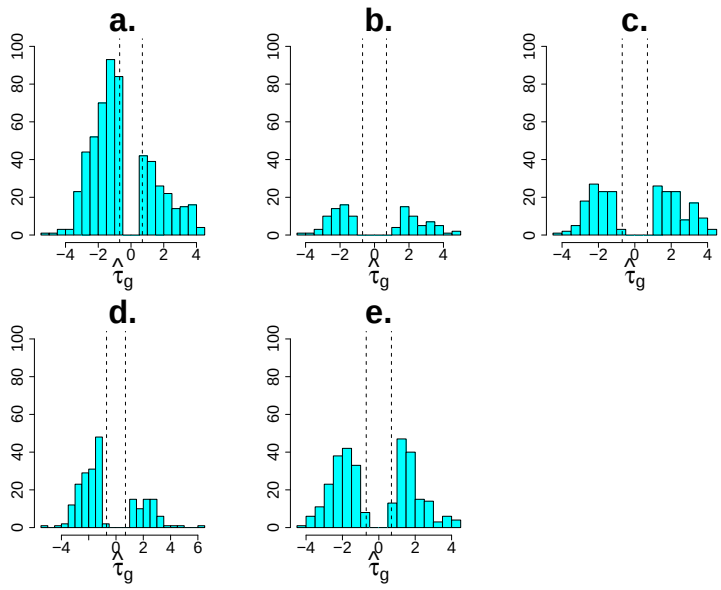

Figure 4: Frequency histogram of Rao-Blackwellized estimates for τ_g by pollutant: arsenic(a), iron(b), nickel(c), zinc(d) and vanadium(e). The dotted lines are placed at log(1/2) and log(2).

the data were collected, and do not extrapolate to larger populations. Additionally, small samples of data imply heavy reliance on accurate technology. Using hierarchical modeling, this individual specific model may certainly extend to borrow strength in hyperparameters from a population of subjects or repeated measures. Nevertheless, we maintain that thoroughly understanding an individual's macro/microarray profile is a vital component to genomic research. Since our approach does not require replicate data, we conceivably have a cost efficient and effective tool for summarizing one person's genomic profile.

By considering the observations as multivariate responses (5x1 vectors) driven by latent variables, we design a hierarchical process that enables us to explore all of the possible expression models, as well as to internally standardize the data without relying on housekeeping genes to estimate variance. At the completion of running MCMC, we use Rao-Blackwellized estimates to summarize parameters of interest. We note that originally an Empirical Bayes approach similar to that of Newton et al. (2001) and Clyde and George (2000) was pursued. Doing so, however, required the enumeration of 32^{1185} models. MCMC avoids this impossible task and supplies an easy method to stochastically integrate the full joint posterior distribution over all possible models. Furthermore, upon assuring chain convergence we may, unlike frequentist analyses, rank the studied genes according to the posterior probabilities of general gene expression, up

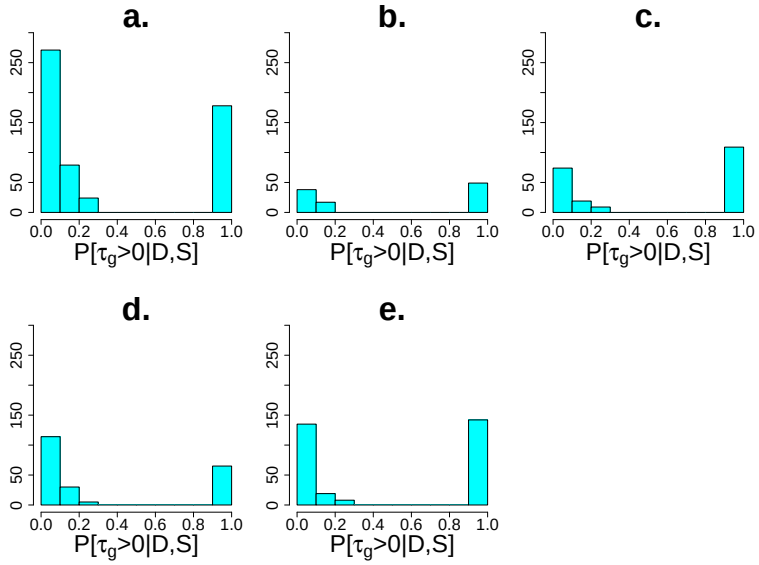

Figure 5: Frequency histograms of posterior probabilities of up regulation, across genes declared differentially expressed by pollutant: arsenic(a), iron(b), nickel(c), zinc(d) and vanadium(e).

regulation, or down regulation. For the studied individual, arsenic had the most impact on the investigated genes.

Aside from possible complications of the Neyman-Scott problem, we have no reason to believe that the proposed method cannot accommodate larger arrays (i.e. microarrays). However, we make certain assumptions that could be relaxed with further computational effort. For example, we assume that all six arrays have the same degree of random variation and consider two sources of variability: treatment effects and random error. Incorporating either gene-specific (ϕ_g) or a treatment-specific (ϕ_p) precision parameters might model the data better than using a single ϕ, as we propose. In fact, some analyses suggest that the log transformation does not entirely alleviate inconsistent variance within arrays. MCMC and Rao-Blackwell techniques can, in principle, adapt to the model inclusion of multiple precision parameters. Incorporating gene-specific precision parameters that are constant across arrays, $\boldsymbol{\phi_g}$, would merely require another level in the proposed hierarchical model. Using the model as presented here, no data would be available to assess a gene-specific variance when $M_g = [1, 1, ..., 1]$. Adding another level in the hierarchical model could enable genes to "borrow" information about the precision from other genes. A model that assumes the precision to remain constant within but not across arrays, $\boldsymbol{\phi_p}$, suggests a Reversible Jump MCMC (RJMCMC) algorithm. Since the block-covariance structure as described in section 3.1 is lost when $\boldsymbol{\phi_p}$ is

included, the analytical integration of $\boldsymbol{\tau}$ is no longer tractable. RJMCMC is necessary to effectively explore the distribution of models or expression profiles.

6 Appendix

The Gibbs sampler implemented for parameter estimation entailed sampling from gamma, beta and multinomial distributions. Here, we derive the multinomial full conditional posterior distribution for the primary parameter of interest, M_g, and state full conditionals for ϕ, $1/(1+\gamma)$, and ω.

The 32 posterior latent expression model probabilities for each gene, P_{mg}, are obtained via the posterior latent expression model odds, O_{gm} (in comparison to the full model, M_g=31). If we let $\boldsymbol{P_g}$=$[P_{g0}, P_{g1}, ..., P_{g31}]$ then we have

$$f(M_g|\omega,\gamma,\phi,\mathbf{D},\mathbf{S}) = \text{Multinomial}(1,\boldsymbol{P_g}) \text{ and}$$

$$P_{gm} = P[M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g] = \frac{O_{gm}}{\sum_{g=1}^{G} O_{gm}},$$

where

$$\begin{aligned} O_{gm} &= \text{Odds}(M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g) \\ &= \frac{f(\mathbf{D}_g,S_g|\phi,\gamma,M_g=m,\boldsymbol{\omega},)}{f(\mathbf{D}_g,S_g|\phi,\gamma,M_g=31,\boldsymbol{\omega})}\frac{f(M_g=m|\boldsymbol{\omega})}{f(M_g=31|\boldsymbol{\omega})} \\ &= BF_{gm}\frac{\prod_{p=1}^{5}\omega_p^{\delta_{mp}}(1-\omega_p)^{1-\delta_{mp}}}{\prod_{p=1}^{5}\omega_p}. \end{aligned}$$

Ultimately, we provide Rao-Blackwellized estimates for M_g and δ_{gp}. At each iteration i, the expected values for M_g and δ_{gp} are calculated conditional upon ϕ, γ, $\boldsymbol{\omega}$, $\mathbf{D}_g$, and S_g,

$$\begin{aligned} \text{E}[M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)} &= \text{P}[M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)} \quad \text{and} \\ \text{E}[\delta_{gp}|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)} &= \sum_{m=1}^{32}\mathbf{1}_{\{\delta_p=1\}}\text{P}[M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)}. \end{aligned}$$

Upon the completion of the MCMC, the expected values are averaged across all I iterations- i.e. the Rao-Blackwellized estimates for M_g and δ_{gp} are

$$\begin{aligned} \text{P}[M_g = m|\mathbf{D}_g,S_g] = \text{E}[M_g = m|\mathbf{D}_g,S_g] &\approx \frac{\sum_{i=1}^{I}\text{E}[M_g = m|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)}}{I} \quad \text{and} \\ \text{P}[\delta_{gp} = 1|\mathbf{D}_g,S_g] = \text{E}[\delta_{gp} = 1|\mathbf{D}_g,S_g] &\approx \frac{\sum_{i=1}^{I}\text{E}[\delta_{gp} = 1|\phi,\gamma,\boldsymbol{\omega},\mathbf{D}_g,S_g]^{(i)}}{I}. \end{aligned}$$

The full conditional distributions for the remaining parameters are

$$f(\phi|\mathbf{D},\mathbf{S},M,\gamma)$$

$$\sim \text{Gamma}(\frac{1185(5)}{2}, \mathbf{1}_{M_g\in[1,31]}\frac{.5}{1+\gamma}\mathbf{D}_g^{'}({\Sigma_\varepsilon}^{-1}+\gamma{\Sigma_{(\delta)(\delta)}}^{-1})\mathbf{D}_g+\mathbf{1}_{M_g=32}\mathbf{D}_g^{'}{\Sigma_\varepsilon}^{-1}\mathbf{D}_g),$$

$$f(\tfrac{1}{1+\gamma}|\mathbf{D},\mathbf{S},M,\phi)$$

$$\sim \text{Gamma}\left(\frac{3+\sum_{gp}^{1185(5)}\delta_{gp}}{2}-1, \frac{\phi}{2}(\sum_{i=1}^{1185}\mathbf{D}_g^{'1}\Sigma_\varepsilon\mathbf{D}_g+\mathbf{1}_{M_g\in[2,31]}\mathbf{D}_{g(\delta)}^{'}{\Sigma_{(\delta)(\delta)}}^{-1}\mathbf{D}_{g(\delta)})\right), \text{ and}$$

$$f(\omega_p|\mathbf{D},\mathbf{S},M,\phi)\sim \text{Beta}(1+\textstyle\sum_{g=1}^{1185}\delta_{gp}, 1185+1-\sum_{g=1}^{1185}\delta_{gp}).$$

Bibliography

Berger, J. and Pericchi, L. (1997). "Objective Methods for Model Selection: Introduction and Comparison." NSF (grants DMS-9303556 DMS-9802261) and CONICIT-Venezuela g-97000592, Duke University and Universidad Simon Bolivar. 108

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). "Bayes Factors and Marginal Distributions in Invariant Situations." *Sankhya, Series A, Indian Journal of Statistics*, 60:307–321. 109

Broberg, P. (2002). "Ranking Genes with Respect to Differential Expression." *Genome Biology*, 3(9):preprint0007.1–0007.23. 107

Chi, C., Rao, D., Stormo, G., and Hicks, C. (2002). "Role of Expression Microarray Analysis in Finding Complex Disease Genes." *Genetics Epidemiology*, 23:37–56. 105, 108

Clayton, C., Pellissari, E., and Quackenboss, J. (2002). "National Human Exposure Assessment Survey: Analysis of Exposure Pathways and Routes for Arsenic and Lead in EPA Region 5." *Journal of Expo. Anal. Environ. Epidemiology*, 12(1):29–43. 105

Clyde, M. and George, E. (2000). "Flexible Empirical Bayes Estimation for Wavelets." *Journal of the Royal Statistical Society, Series B, Methodological*, 62(4):698–698. 116

Dawid, A. P. and Lauritzen, S. L. (2000). "Compatible Prior Distributions." In George, E. I. (ed.), *Bayesian Methods with Applications to Science Policy and Official Statistics. The sixth world meeting of the International Society of Bayesian Analysis*, 109–118. 109

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85:398–409. 111

Genovese, C. and Wasserman, L. (2002). "Operating Characteristics and Extensions of the False Discovery Rate Procedure." *Journal of the Royal Statistical Society, Series B, Methodological*, 64(3):499–517. 114

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423):881–889. 110

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90:773–795. 109, 113

Lavine, M. and Schervish, M. J. (1999). "Bayes Factors: What they are and what they are not." *The American Statistician*, 53:119–122. 110

Newton, M. A., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes from Microarray Data." *Journal of Computational Biology*, 8:37–52. 116

Smyth, G., Yang, Y.-H., and Speed, T. (2003). *Functional Genomics: Methods and Protocols*, volume 224 of *Methods in Molecular Biology*, chapter Statistical Issues in cDNA Microarray Data Analysis, 111–136. Totowa, NJ: Humana Press. 108

Speed, T. and Group (2000). "Hints and Prejudices." website: http://stat-www.berkeley.edu/users/terry/zarray/Html/log.html. 107

Strawderman, W. E. (1971). "Proper Bayes Minimax Estimators of the Multivariate Normal Mean." *The Annals of Mathematical Statistics*, 42:385–388. 110

Vahter, M., M., B., Akesson, A., and C., L. (2002). "Metals and Women's Health." *Environ. REs.*, 88(3):145–155. 105

Yang, Y.-H. and Speed, T. (2002). "Design Issues for cDNA Microarray Experiments." *Nature Reviews, Genetics*, 3:579–588. 106

Acknowledgments

This material was based upon work supported by the National Science Foundation (NSF) under Agreement No. DMS–9733013 and the Environmental Protection Agency (EPA)-Duke training agreement. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or policies of the NSF or the EPA. Specifically, the research described in this article has been reviewed by the Health Effects and Environmental Research Laboratory of the United States EPA and has been approved for publication. Approval does not imply that the mention of trade names or commercial products constitute endorsement or recommendation for use.