Modeling and Analysis of Multi-library, Multi-group SAGE Data with Application to a Study of Mouse Cerebellum

Zailong Wang[1], Shili Lin[1,2,*], Magdalena Popesco[3], Andrej Rotter[1,3]

[1]Mathematical Biosciences Institute, [2]Department of Statistics, and [3]Department of Pharmacology, The Ohio State University, Columbus, OH 43210, USA.

[*]Correspondence to: Shili Lin, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, USA. Email: shili@stat.ohio-stat.edu.

SUMMARY. A Serial Analysis of Gene Expression (SAGE) library is a collection of thousands of small DNA "tags", each of which represents a distinct mRNA transcript. Existing methods have been proposed for analyzing single library data (i.e., one library per group) or one tag at a time. The practice of lumping all libraries together (in a multi-library setting) to form a "mega" library for each group is obviously unsatisfactory, but nonetheless performed frequently due to the lack of alternative methods. Since the tag counts within each library are inter-related as they are drawn from a multinomial distribution, analyzing thousands of tags one at a time is undoubtedly inadequate. Not only does such a practice ignore the dependency, but it also faces the multiple testing adjustment issue. This article is an attempt to address both of these issues so that all tags from multi-library groups can be analyzed together. The methods proposed also gears toward multi-group data. Focusing on the problem of identifying genes that are differentially expressed, a Bayesian formulation is established. Under this formulation, the problem of separating the differentially expressed genes from the majority of similarly expressed ones is treated as a model selection problem, and the reversible jump Markov chain Monte Carlo method is adapted for this purpose. The method is applied to a set of mouse libraries to uncover genes that are associated with the process of aging in the cerebellum. Our Gene Ontology (GO) analysis of the genes selected classifies them into several GO categories, which appear to be functionally relevant to aging.

KEY WORDS: Bayesian hierarchical modeling, Reversible jump Markov Chain Monte Carlo, Mouse library, Cerebellum, Serial Analysis of Gene Expression (SAGE), Gene Ontology (GO).

1 INTRODUCTION

The characteristics of an organism are determined by the genes expressed within it. Serial Analysis of Gene Expression (SAGE) has been introduced as a tool for quantifying expressions of tens of thousands of genes simultaneously (Velculescu *et al.* 1995, Madden *et al.* 1997). This is a method for multiplex gene expression screening that depends on short sequences ("tags"; 10 to 14 bp) located at specific sites. The basis is that these short tags are sufficiently long to enable the gene that codes for the mRNA to be uniquely identified with extremely high probability. Therefore SAGE provides a quantification of the mRNA population in a cell without prior selection of the genes to be studied. It has been used to study a wide range of biological systems (Zhang *et al.* 1997, Blackshaw *et al.* 2001, Abba *et al.* 2004, etc.).

Different from microarray technology (Schena *et al.* 1995), SAGE does not require prior knowledge of the transcripts. Instead, it provides estimates of the absolute abundance of the transcripts in the entire genome. In a nutshell, SAGE can be regarded as an "open" system since it can potentially reveal expression levels of all genes, whereas microarrays are "closed" because they can only track the expressions of the genes spotted on the array. Furthermore, SAGE is a much more accessible method since it does not require any sophisticated equipment to track gene expressions, although it can be more difficult to perform, which requires excellent skills of a technician.

A SAGE library is a collection of thousands of tags and their corresponding counts, each of which represents a distinct mRNA transcript. However, due to sequencing and/or PCR errors, a small proportion of the tags may not represent real genes, which alters the estimates

of the numbers of transcripts observed. Thus it is of importance to perform statistical modeling and corrections of errors due to the sequencing step in SAGE (Beißbarth *et al.* 2004) prior to any statistical analysis to provide answers to questions of scientific interest. One type of scientific problems that SAGE has been used to address is the identification of genes with differential expression levels under different conditions through comparing the numbers of tags found in libraries generated under these conditions. Our specific problem as described next falls into this category.

1.1 Study of mouse cerebellum and SAGE data

Age related changes, such as cytological alterations and neuronal losses, have been well documented in the cerebellum. The cerebellum is essential for the control of balance (equilibrium), posture, and motor coordination. During normal aging, the cerebellum can become progressively dysfunctional, which may be attributed to alterations of specific molecular components. Since progressive dysfunction of the cerebellum can lead to life-threatening accidents, it is of importance to use mouse as an animal model to study its cerebellum to identify genes whose expression levels change during aging by comparing adult and aged mice.

In our data set, we have six SAGE libraries from the cerebella of six male mice. These libraries were constructed by Dr. Magdalena Popesco in Professor Andrej Rotter's laboratory at the Ohio State University. They were divided into the adult and the aged groups. The three mice in the adult group were sacrificed at postnatal days of 92, 150, and 300, and will be referred to, respectively, as the P92, P150, and P300 mice. A similar naming scheme was applied to a group of three aged mice, P810(1), P810(2), and P840, in which two were both sacrificed at postnatal day of 810, and the third was at day 840. The total number of tags (unique tags) in each of these six libraries is, in the order given above, 16,430 (7,144), 18,103 (8,420), 10,578 (6,416), 18,581 (10,544), 8,528 (4,989), and 7,630 (3,716), respectively. The

total number of unique tags across all six libraries is more than 26,000, but the majority of the tag counts in each of the libraries is less than three. Since "high abundance" tags are of greatest interest, we pre-process the data to extract the most relevant ones. The tags that are included in our analysis must be present in two of the libraries, each with counts greater than five after normalization (that is, after bringing the total number of tags up to that (18581) of the largest library, P810(1)). This filtering step reduces the number of tags to 596.

1.2 Analysis methods

A popular method among experimental scientists for comparing two-group SAGE library data is that of P-chance from the SAGE2000 software suite (Velculescu *et al.* 1995; Zhang *et al.* 1997). This method is simulation based, which provides Monte Carlo estimates of the p-values for each tag based on normalized summed tag counts of the two groups. The most attractive feature of this method is its conceptual simplicity, but since such an analysis is based on combined libraries, it ignores normal variations between libraries within the same group. Furthermore, the method is only applicable to the two-group setting. Several other methods have also been developed for comparing the relative abundance of mRNAs between two single-library groups. Examples include the eSAGE program (Margulies and Innis 2000) and those discussed in Madden (1997), Michiels *et al.* (1999) and Man *et al.* (2000). The usual Z-test for comparing two population proportions (with pooled data) is such an example. In fact, the P-chance method is a Z-test but with Monte Carlo p-values.

Recognizing the problems associated with data pooling in multi-library/group situations, such as potentially overstating the significance of a difference, methods have been proposed to take into account of within group inter-library variability. For example, Ryu *et al.* (2002) used a series of filters to deal with groups of multiple pancreatic libraries. Baggerly *et al.* (2003) and Vencio *et al.* (2004) both used a beta-binomial model and suggested the use of a

modified t-statistic or a Bayesian error rate, respectively, to select genes that are differentially expressed. Whereas in Baggerly *et al.* (2004), the authors used an overdispersed logistic regression approach to model groups of multi-libraries. However, despite their ability of accounting for between library variations, tags are still being analyzed one at a time, which ignores dependencies among tags within a library and also leads to the issue of adjusting for multiple testing.

In this paper, we develop a statistical method that is amenable to analyzing multi-library, multi-group SAGE data as well as all tags simultaneously. Under a Bayesian hierarchical modeling framework, we cast the issue of separating tags that are differentially expressed (DE) from those that are similarly expressed (SE) as a model selection problem. The reversible jump Markov chain Monte Carlo (MCMC) method is used for this purpose. The posterior probability of each tag being differentially expressed is calculated at the end of the MCMC process, and a criterion based on the Bayes Factor (BF) is used to classify tags into the DE or SE sets.

The rest of this paper is organized as follows. Section 2 presents a hierarchical Bayesian modeling framework and the associated parameter distributions, the MCMC samplers and algorithms, and simple diagnostics and decision rules. This is followed in Section 3 by a simulation study to evaluate the proposed method under two different settings and sensitivity analyses. Section 4 reports the analysis and results of our mouse data, while a few concluding remarks are given in Section 5. Technical details are available from Web Appendix A. The Matlab code that implements the algorithms with instructions are available from our website provided in the Supplementary Materials Section.

2 METHODS

2.1 Hierarchical Bayesian Modeling

Let $X = \{X_{kig}, k = 1, \cdots, K; i = 1, \cdots, n_k; g = 1, \cdots, G\}$ denote the gene expression data from SAGE experiments. Here K is the number of groups (conditions) of SAGE libraries, n_k is the number of SAGE libraries in group k, and G is the total number of unique tags across all libraries. So the total number of libraries is $n = \sum_{k=1}^{K} n_k$. The goal is to identify tags whose expression levels are not all equal among all the K groups, where $K > 2$ and $n_k > 1$ correspond to the multi-group, multi-library scenario, the focus of the current paper, but the method is applicable to situation where $K = 2$ and/or $n_k = 1$ for some of the $k's$.

For a gene g whose expression levels are different among the groups, we assume that the tag count X_{kig} follows a distribution (yet to be defined) with parameter p_{kg}, which represents the abundance of the tag in population (condition) $k, i = 1, \cdots, n_k$. On the other hand, for a gene whose expression levels are the same among all K populations, the abundance parameters are assumed to be equal, i.e. $p_{1g} = p_{2g} = \cdots = p_{Kg} := p_g$. Under this parametrization, the tag count of a gene, without a priori knowledge of whether its expression levels are different among different conditions, can be regarded as following a two-component mixture distribution:

$$X_{kig} \sim \sum_{j=1}^{2} w_j f_j(\cdot|\theta_{jg}),$$

where $f_j(\cdot|\theta_{jg})$ is a given parametric family of densities indexed by a vector parameter θ_{jg}, and $w_j, j = 1, 2$, denote the mixing proportions of the genes being differentially expressed or not, and sum to 1. The parameter vector in the two component densities are $\theta_{1g} = \{p_{1g}, \cdots, p_{Kg}\}$, and $\theta_{2g} = \{p_g\}$, respectively, which are independent of the $w's$ and are the main parameters of interest.

Under this formulation, each tag is postulated to be drawn from a heterogeneous population consisting of two sets, the DE set S_1 and the SE set S_2. Each potential division

of the genes into the two sets is a possible model in the total model space $\mathcal{M}$, that is, $M = \{S_1, S_2\} \in \mathcal{M}$. Note that the size of the model space is $\|\mathcal{M}\| = 2^G$. All tags in S_1 have differential expression levels among the groups, i.e., $p_{k_1 g} \neq p_{k_2 g}$ for at least two different groups k_1 and k_2. The remaining tags that fall into S_2 have the same abundance for all the groups.

Our purpose can then be regarded as choosing an appropriate model M from the space $\mathcal{M}$ given data X. Given a model $M \in \mathcal{M}$ and its associated parameter vector $\theta_M = \theta_1 \cup \theta_2 = (\cup_{g \in S_1}\{p_{1g}, \cdots, p_{Kg}\}) \cup (\cup_{g \in S_2}\{p_g\})$, the likelihood function can be simply written as

$$L(\theta_M, M) = \prod_{g \in S_1} \prod_{k=1}^{K} \prod_{i=1}^{n_k} f_1(X_{kig}|p_{kg}) \times \prod_{g \in S_2} \prod_{k=1}^{K} \prod_{i=1}^{n_k} f_2(X_{kig}|p_g),$$

assuming that the tag counts are independent conditional on the specific model M and the individual group parameters.

To facilitate learning about the model M and its associated parameters, we cast the problem into a hierarchical modeling framework. We introduce prior distributions for θ_M under hyperparameter vector δ_M, which is in turn specified by a hyperprior with known parameters. The joint posterior distribution for all the parameters is then factored into

$$P(M, \theta_M, \delta_M|X) \propto L(\theta_M, M \mid X) P(\theta_M|\delta_M, M) P(\delta_M|M) P(M). \tag{1}$$

It remains to specify the distribution for the data and the prior distributions for the parameters. The process of SAGE experiments naturally leads to the assumption that the tag counts of a library follow a multinomial distribution. Thus each tag count can be modeled as coming from a binomial distribution, which is well approximated by a Poisson distribution for SAGE data (Cai *et al.* 2004). Therefore, we assume that $X_{kig} \sim \text{Poisson}(N_{ki} p_{kg})$ or $X_{kig} \sim \text{Poisson}(N_{ki} p_g)$, depending on whether $g \in S_1$ or $g \in S_2$, where $N_{ki} = \sum_{g=1}^{G} X_{kig}$ is the total number of tags in library i within group k. The prior distributions for the

parameters in $\theta_M = (\cup_{g\in S_1}\{p_{1g}, \cdots, p_{Kg}\}) \cup (\cup_{g\in S_2}\{p_g\})$ are assumed to be independent:

$$p_{kg} \sim \beta(\alpha_{kg}, \bar{N}_k - \alpha_{kg}), \quad \text{for } g \in S_1, \quad k = 1, \cdots, K;$$
$$p_g \sim \beta(\alpha_g, \bar{N} - \alpha_g), \quad \text{for } g \in S_2,$$

where $\bar{N}_k = n_k^{-1}\sum_{i=1}^{n_k} N_{ki}$ and $\bar{N} = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k} N_{ki}$. One may note that the above modeling is similar to that of beta-binomial of Baggerly et al. (2003) and Vencio et al. (2004) if conditioning on only the hyperparameter. The hyperparameter vector $\delta_M = \{\alpha_{kg}, k = 1, \cdots, K, g \in S_1; \alpha_g, g \in S_2\}$ are themselves assumed to be independently distributed as truncated Gamma's:

$$\alpha_{kg} \sim \Gamma(a_{kg}, b_k)\mathbf{1}_{[0,\bar{N}_k]} \quad \text{with} \quad a_{kg} = \sum_{i=1}^{n_k} X_{kig} \text{ and } b_k = n_k;$$

$$\alpha_g \sim \Gamma(a_g, b_k)\mathbf{1}_{[0,\bar{N}]} \quad \text{with} \quad a_g = \sum_{k=1}^{K}\sum_{i=1}^{n_k} X_{kig} \text{ and } b = n.$$

Finally, the prior distribution for $M \in \mathcal{M}$ is set such that the expected number of genes in S_1 can be controlled by investigators. Specifically, let $\|S_1\|$ denote the number of genes in S_1 under model M. Then

$$P(M \mid \|S_1\| = s) = \lambda^s(1-\lambda)^{G-s}, \tag{2}$$

and we may control the parameter λ by setting $E\|M\| = G\lambda = c$ for a pre-determined constant c, which may be set by the scientist based on information from prior knowledge or budgetary consideration. Equivalently, we may obtain the value of parameter λ by controlling the prior odds, $\lambda/(1+\lambda)$, for each tag to be in S_1.

2.2 MCMC Samplers and Algorithms

The MCMC methods used for sampling from the posterior distribution (1) include a mixture of Metropolis-Hastings (M-H) algorithms for updating the parameters under model M and the reversible jump MCMC method of Green (1995) for updating the model M itself (that

is, for tag movement between the DE and the SE sets). More specifically, we use the M-H algorithms to update the parameters in θ_M and δ_M for a given model M, whereas reversible jump MCMC is used to add tags to S_1 (i.e., delete tags from S_2) or vice versa. These two types of MCMC updates are combined to identify tags that are differentially expressed. Details of the M-H and the reversible jump MCMC algorithms, as adapted for our analysis, are given in Web Appendix A. In what follows, we present an algorithm for combining these two types of MCMC updates in which one tag is selected for adding/deleting from the DE set in each iteration (cycle) of parameter updating.

Algorithm: One-Tag (OT)

- Step 1. Initialization: Create an initial model for M (e.g., by randomly separating all tags into two sets, S_1 and S_2, of equal sizes), and initialize all the other parameters under the initial model $M = \{S_1, S_2\}$.

- Step 2. Update the parameters in θ_M under model M: The parameters in θ_1 and θ_2 will be updated in parallel according to the M-H algorithms described in Web Appendix A.2.

- Step 3. Update model M:

  (a). Choose one tag, g, randomly from the entire gene set. If $g \in S_2$ (S_1), calculate the acceptance probability of adding (deleting) it to (from) the DE set S_1. Note that this updating step involves changes in the parameter space and consequently its dimension, and thus a reversible jump MCMC algorithm is used (Web Appendix A.3) to guarantee dimension matching.

  (b). If the proposed move is accepted, then update the model M to reflect the successful move.

- Repeat the two updating steps (2 and 3) as many times as needed until convergence.

Note that in the above algorithm, only one tag is randomly selected for potential switch of set membership. This results in a small change in the model, M, even if the move is successfully accepted. Larger steps in the model space can be taken with an alternative updating scheme. For instance, several tags may be selected for potential movement between S_1 and S_2 during each cycle. To this end, we have also explored an algorithm that considers all tags (AT algorithm) in each iteration to ascertain their potential switch of memberships in the two complementary sets. It is anticipated that, with respect to the number of iterations, the AT algorithm, which has larger movement in the model space $\mathcal{M}$, will lead to faster convergence compared to the OT algorithm. However, the computational time for the former is expected to be longer than the latter for each iteration. Thus, it is important to take computational time into account when assessing their relative performances. Since these two algorithms performed similarly with the same amount of computational time from our experience, only results from the OT algorithm are reported in the simulation study below.

2.3 Convergence diagnostics and decision making

Three types of simple diagnostic plots are utilized for evaluating the performance of the algorithm. The first is based on the correlation between two lag L estimates of the posterior probability vector. The second type of plots is based on the number of tags selected to be in the DE set S_1 in each iteration. In other words, they are trace plots of the sizes of the the DE set against iterations. The third provides convergence diagnostics for each gene. They are again trace plots of the posterior probabilities of a gene being classified as differentially expressed against the number of iterations.

One way to identify tags as from S_1 is via the Bayes Factor (BF), which is the posterior odds over the prior odds. Theoretically, BF is independent of the chosen priors (Richardson and Green, 1998), and according to Raftery (1996), a BF between 10 and 100 is considered as strong evidence for $H_1 : g \in S_1$ against $H_0 : g \in S_2$. In our simulation study and

application, we use $BF > \frac{10+100}{2} = 55$ as our decision rule for declaring a tag to be in S_1. This corresponds to a posterior probability of 0.846 or grater with a prior odds of 0.1. Other cutoff values for BF are also explored in our sensitivity analysis.

3 SIMULATION STUDIES

3.1 Two groups

In order to test our method, we simulated two groups of data as follows so that their characteristics resemble those of the real SAGE data as described in the Introduction Section. For each (g) of the 596 tags in our pre-filtered mouse data set, we computed the between group to within group variations ratio (BW_g) as well as the average proportions of tag counts in each of the two groups $\{\bar{p}_{1g}, \bar{p}_{2g}\}$, where $\bar{p}_{kg} = (\sum_{i=1}^{n_k}(X_{kig}/N_{ki}))/n_k, k = 1, 2$, using the notation defined in Section 2. Among the set of tags for which $\bar{p}_{1g} > \bar{p}_{2g}$ (the up-regulated group), those corresponding to the largest 25 BW ratios were selected as belonging to the DE set S_1. Similarly, 25 tags among the down-regulated set ($\bar{p}_{1g} < \bar{p}_{2g}$) were selected to be included in S_1. The remaining 546 tags were treated as coming from the SE set S_2. To complete our simulation setting, we assume the underlying common parameter for each tag g in S_2 to be $p_g = (\bar{p}_{1g} + \bar{p}_{2g})/2$. For each tag in S_1, on the other hand, the underlying parameters in the two groups are set to be $\{p_{1g} = \bar{p}_{1g},\ p_{2g} = \bar{p}_{2g}\}$ up to a scale. The common scale parameter for those in the up-regulated group and that among the down-regulated genes were set to satisfy the constraints that the probability vector in each group adds up to 1. Based on these parameter values, we simulated each library according to the multinomial distribution, mimicking the experimental process of generating a real SAGE library. The number of libraries in each group was set to match our real data, while the size of each library was 10 times of that of our data so that they are more in line with typical SAGE library sizes as reported in the literature (Ruijter *et al.* 2002).

The results for this simulated data set are shown in the left panel of segment I of Table 1. With the prior odds set to be 0.1, we ran the OT algorithm for 80,000 iterations, which took 24 minutes on a Pentium 4, 2.4GH PC with 512MB of RAM. As can be seen from the table, the algorithm performs reasonably well, with over 95% power for identifying tags that are in the DE set (49/50) and a small false positive rate (3/546). The left plot of Figure 1 shows the estimated posterior probabilities in S_1 (dark gray) and S_2 (light gray) for each of the 596 tags, arranged in descending order according to their posterior probabilities in S_1. The posterior probabilities for the single false negative (dashed vertical line) and the false positives (solid vertical lines) are indicated in the plot. We can see from it that the false positives are among the positives with the smallest probabilities while the false negative is also closer to the boundary compared to the majority of the negatives.

To compare the results from our method with those from traditional methods, we applied the two-sample t-test and Z-test suggested by Kal *et al.* (1999) to the same simulated dataset. The results with a per-comparison error rate of 0.01 are reported in segment II of Table 1. As can be seen from these results, the t-test has a lower power (86%) and both tests have higher numbers of false positives (7 and 10 for the t- and Z-test respectively) when compared to results from the OT algorithm. After adjusting for multiple testing based on a false discovery rate of 0.05, the numbers of false positives fall down to about the same levels as ours (3 for both of the tests), but the power for the t-test drops down to only 74%, although that for Z-test remains the same.

Lag L (L=800) correlations between two posterior probability (PP) vector estimates were calculated as a way of monitoring convergence. In other words, we calculated the correlation between every two consecutive estimates of the posterior probability vector, estimated after every L iterations, and showed them in the top left plot of Figure 2. The algorithm seemed to have provided consistent posterior probability estimates after 20*L iterations. In terms of the number of tags declared positive, the algorithm again converged rather fast, as shown

in the middle left plot of Figure 2. Using a representative tag from S_1 and one from S_2, we show on the left of the last row of Figure 2 the trace plot of the respective posterior probabilities. Trace plots for all the other tags show similar patterns and are thus omitted here. Overall, the monitored estimates all appeared to have converged; the diagnostic plots do not reveal any feature that may cause serious concerns.

3.2 Three Groups

For testing our approach with more than two groups of libraries, we simulated a third group with three libraries. For the 50 tags in S_1 in our earlier simulation setting, we assumed the parameter for the third group to be $p_{3g} = p_{1g} + p_{2g}$ ($p_{1g} \neq p_{2g} \neq p_{3g}$). In addition, we selected another 50 tags from S_2 (in which $p_{1g} = p_{2g}$) in the two-group setting and let $p_{3g} = p_{1g}/2$ ($= p_{2g}/2$). These 50 tags are the first 50 of the remaining 549 tags arranged in alphabetical order of the tags. Under this new simulation setting, we have 100 tags belonging to the DE set S_1 and 496 in the SE set S_2. All libraries were again simulated from the multinomial distributions. Note that the probabilities for the tags in S_1 were again scaled to make the multinomial probability vectors each summing to 1.

We ran the OT algorithm for 80,000 iterations, which took 28 minutes to complete using the same computer as described above. The same prior odds of 0.1 as in the two-group setting was used. The outcomes are given in the right panel of segment I of Table 1. They are comparable to those from the two group data, with a high power and a low false positive rate. The posterior probabilities of each tag being in S_1 are given in the right plot of Figure 1, with the two false positives and the single false negative identified by solid and dashed vertical lines, respectively. As can be seen from the figure, the false positives and the false negatives are all close to the boundary for declaration of significance, similar to the results from the two-group setting. Diagnostic plots as those in the left panel of Figure 2 were shown in the right panel of the same figure. Again, none of the plots reveal any unusual

feature that requires further investigation.

3.3 Sensitivity Analysis

Since results from our Bayesian formulation are dependent on the choice of priors for the model parameter vector $\theta_M = (\cup_{g\in S_1}\{p_{1g}, \cdots, p_{Kg}\}) \cup (\cup_{g\in S_2}\{p_g\})$, and the model M itself, we studied the degree of sensitivity of our method to the specifications by considering a class of beta priors for the θ_M and various prior odds (through the λ parameter as in Section 2.1) for M. In addition, we also studied the sensitivity of the method to the pre-processing step (normalization and filtering as described in 1.1). Since this step is only applied to the mouse data, we defer discussion on it until the Results Section of the data analysis.

For studying the potential influence of prior odds on the resulting tags labeled as differentially expressed, in addition to setting $\lambda/(1+\lambda) = 0.1$, as in our simulation study for the two-group setting, we also considered the following prior odds: 0.01, 0.05, 0.15, and 0.5. The results are given in segment I of Table 2, with those based on the prior odds of 0.1 included in the same table for ease of comparisons. As can be seen from the table, the powers are high and the Type I error rates are low regardless of the wide range of prior odds. This is indicative of the robustness of the method to the choice of the prior for M. Results from the three-group setting would be similar, as setting the prior odds equal to 0.1 in that case already constitutes a marked deviation from the expected number of differentially expressed tags.

With regard to θ_M, the p_{kg} and p_g parameters are assumed to follow $\beta(t\alpha_{kg}, t(\bar{N}_k - \alpha_{kg}))$ and $\beta(t\alpha_g, t(\bar{N} - \alpha_g))$, respectively. In addition to $t = 1$, the value used in our simulation study in the previous two subsections, we also considered $t = 0.5$ and 1.5 for both simulated datasets ($K = 2$ and $K = 3$) using the OT algorithm. The results for the two-group setting is given in segment II of Table 2, in which the "Common" column gives the numbers of tags commonly selected by using all three different priors. Using a BF cutoff of 55 as in

the simulation study, the number of tags identified under the different t values deviates by at most two, with around 95% of the tags being common. Using a much bigger cutoff of BF=100, the results stay very much the same, as can be seen from the table, leading to the conclusion that the method is robust to the specification of the priors for the θ_M parameters. For the three-group setting, the number of tags selected as differentially expressed ranges from 100 to 102 (with 100 of them in common) for the different t values and BF cutoffs. Since conclusion drawn from these results does not deviate from that for the two-group setting qualitatively, the detailed results are omitted.

In summary, the above results indicating insensitivity of the methods to the priors are not surprising, and are largely due to the choice of the BF for inference. As pointed out by Richardson and Green (1998), an attractive feature of BF is that it is theoretically independent of the priors; inference under BF does not need to reference the priors used and thus should be insensitive to the specific choices.

4 RESULTS

We now return to the real SAGE libraries in the mouse cerebellum study described in Section 1.1. We ran the OT algorithm for 80,000 iterations to identify tags that are differentially expressed in the aged group versus the adult group. Using a prior odds of 0.1 and a BF cutoff value of 55, 20 tags were selected. For this real data analysis, we also ran the AT algorithm for 1,000 iterations (since OT is roughly 80 times faster than AT for each iteration), which resulted in the identification of 19 DE tags, of which 17 were in common with those selected by the OT algorithm. This largely consistent findings instill initial confidence in the tags identified, especially in those that are commonly predicted. Nine of these tags, displayed in the first half of Table 3, are up-regulated (that is, more highly expressed) in the aged cerebella. Furthermore, seven of them correspond to known unigene-IDs, whose gene symbols/names are also listed in the table. The remaining eight of the common genes, shown

in the second half of Table 3, were down-regulated in the aged cerebella, whose corresponding unigene IDs and gene symbols/names are given in the table as well. As can be seen from the table, six of the tags correspond to two unigene IDs each, reflecting in part the imperfect system of gene naming convention and repository of data. Diagnostic plots (not shown) as those in Figure 2 did not reveal any unusual feature to require any further investigation.

As described in Section 1.1, all the libraries were pre-processed by normalizing the counts and filtering out low abundance tags. To study whether the tag identification method is robust to this pre-processing step, in addition to normalizing to the size of the largest library (18,581 tags) as done in the above analysis, we also considered normalizing all the libraries to 50,000 tags. With this new normalizing count and the same original filtering step (count of 5 in two of the libraries), the number of tags surviving jumped to 3,237, which is 443% over 596 tags. The use of this new dataset and the OT algorithm led to the identification of 27 DE tags. Given that this is only 35% over 20 with a dataset that is more than five times larger, the result is encouraging. When the threshold in the filtering step is raised to 10, the number of remaining tags is 1,030, which still almost doubles that of 596 tags, but the number labeled as DE this time remained the same at 20. These results suggest that the method is not overly sensitive to the preprocessing method, especially when comparable combinations of the normalizing and filtering steps are chosen. We note that the requirement of at least two libraries passing the threshold for the tag to be filtered through is most appropriate for the multi-library per group setting, which has the effect of guarding against outlying observations. In the case that there is a single library in some of the groups, a different filtering scheme is warranted.

To discern whether the genes selected as differentially expressed are meaningful biologically, the 21 unigene IDs were used for further analysis using the Gene Ontology (GO) Tree Machine (GOTM; http://genereg.ornl.gov/gotm) to annotate their functions and classify them into functional categories. Using all genes in the mouse genome as our reference gene

set, we were interested in identifying GO categories that are being enriched in our set of 21 genes. In other words, we wanted to identify functional categories in which there are more genes in our list belonging to them than expected if the genes were randomly selected from the mouse genome. For a given category, under the null hypothesis of random selection, the number of genes from our list falling into that particular category follows a hypergeometric distribution, leading to a simple test for the hypothesis. In Figure 3, all GO categories that were identified to be significantly enriched (raw $p < 0.01$; with category names in black, boxes shaded or not), together with their ancestral categories (with category names in gray; up to the top level with three main categories: biological process, molecular function, and cellular component), were displayed as a directed acyclic graph (bottom panel). The numbers below or next to a category are the observed/expected gene numbers for that category. Also displayed in the figure (top panel) are the genes involved in each GO category shown in the bottom panel. The GO categories pointed to by arrows correspond to those identified as enriched.

From the raw p-values of GOTM, we calculated the adjusted p-values to correct for multiple testing using the FDR method (Benjamini and Hochberg, 1995) implemented in SAS (www.sas.com). A cutoff of 0.05 for the adjusted p-values led to a number of enriched categories no longer being enriched (non-shaded with names in black in Figure 3). If we would use either the step-down Bonferroni method of Holm (1979) or the step-up Bonferroni method of Hochberg (1988), both of which control for family wise error rate, then two additional categories, "cellular physiological process" and "cytoplasm" would be dropped out from the multiplicity adjusted enriched list.

As can be seen from Figure 3, several of the enriched categories were "oxygen" related. This observation lends itself to further annotation of the genes involved. Oxygen binding heme proteins (neuroglobin, hemoglobin and myoglobin) may protect neurons from hypoxic-ischemic injury in vitro and in vivo (Sun et al, 2005). While neuroglobin expression in the

cerebellum decreases with age, our finding suggests that this process may be counteracted by an increased expression of hemoglobin mRNA (Hbb-b1 and Hbb-b2). An elevated presence of this oxygen-binding transporter protein may serve to protect cerebellar cells from age-dependent neurodegeneration.

The two genes involved in the "hormone activity" category are decreasingly expressed (down-regulated) in the cerebella of the aged mice. We note that this category would have been labeled as FDR enriched (FDR p=0.076; raw p=0.006) had we used a less stringent cutoff. These two genes are Prl (prolactin) and Ttr (transthyretin). The function of transthyretin, a lipophilic molecule binding protein, is to transport thyroid hormones, such as thyroxine and retinoids; it also acts as a chelation agent for the neurotoxic beta-amyloid peptide, thus preventing its deposition in the brain (Schwarzman et al, 1994). Within the brain, transthyretin is synthesized exclusively in the choroid plexus (including that associated with the cerebellum) and is secreted into the cerebrospinal fluid (Chen et al., 2005). It has been suggested that thyroxine contributes to age-related cognitive decline (Hulbert, 2000), and it is plausible that the decrease in transthyretin synthesis observed in the present study reflects a decreased capacity to protect the cerebellum against neurodegenerative insults. The presence of prolactin-like immunoreactivity (Emanuele et al., 1987) and prolactin mRNA in the cerebellum (Emanuele et al, 1992) is well established. Mice lacking prolactin live longer than their normal siblings and exhibit many symptoms of delayed aging (Bartke and Brown-Borg, 2004). The observed decrease in prolactin message in aged cerebellum may reflect a mechanism that optimizes cell survival.

5 DISCUSSION

In this paper, we propose a statistical method for analyzing multi-library, multi-group SAGE data with all tags considered simultaneously. The results from our simulation studies indicate that the method is able to identify tags (genes) that do not have the same expression levels

across all groups while keeping the false positive rates low. Compared to standard analysis methods in situations where such methods are applicable, our proposed approach is certainly competitive. For the three-group simulated data, our methods also performed satisfactorily. This represents a step forward in enriching the tools capable of analyzing more complex SAGE library data. More importantly, application of the method to the mouse cerebellum data yields biologically interesting and meaningful results. Together with results from other studies, this may help uncover the specific molecular changes during normal aging.

Throughout our simulation study and analysis of the SAGE mouse data, we used, respectively, 80,000 and 1,000 MCMC iterations for the OT and the AT algorithms, which took less than 30 minutes to compute on a typical PC, although we presented results for the OT algorithm only in the simulation study since they were comparable and would lead to the same conclusion. Our simple diagnostic plots indicate that similar results would have been achieved had we executed much shorter runs. For all the analysis carried out in the current paper, since the algorithms converged rather quickly while our runs were fairly long, our results were based on all iterations without entertaining a burn-in period. However, in general, it is advisable to delete the initial portion (say 10%) before the additional iterations are used for making inferences. Our exploration reveals that the AT algorithm that considers all tags for switching their set membership in each iteration appears to be slightly more efficient with the same amount of computational time, although both algorithms lead to satisfactory results. Therefore, either one should be a reasonable choice. One possibility is to run both algorithms and use either the intersection of the two lists (as we have done for the mouse data to be conservative) or their union for further analysis.

Finally, we note that although our method was motivated by the SAGE mouse data, the general scheme of the methodology is applicable to other types of biological data from a number of platforms, such as the DHM (differential hypermethylation) 12K CpG islands arrays. The differences in handling the different types of data lie in the distributional assumptions,

which affect the likelihood component as well as the choice of appropriate priors.

Supplementary Materials

Web Appendix A referenced in Sections 1.2 and 2.2 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics. The mouse cerebellum data analyzed in Section 4 are available at http://www.ncbi.nlm.nih.gov/geo/query /acc.cgi?acc=GSE1090. The Matlab code implementing the OT an AT algorithms is available at http://www.stat.ohio-state.edu/∼statgen/SOFTWARE/DE-SAGE/.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants DMS-0112050 (to the MBI) and DMS-0306800 (to S.L.). The authors would like to thank Drs. Adrienne Frostholm and Lynn Friedman for stimulating discussions, and an Associate Editor and two anonymous referees for constructive comments and suggestions.

REFERENCES

Abba, M. C., Drake, J. A., Hawkins, K. A., Hu, Y., Sun, H., Notcovich, C., Gaddis, S., Sahin, A., Baggerly, K. and Aldaz, C. M. (2004) "Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression," *Breast Cancer Res.*, 6, R499-R513.

Baggerly, K. A., Deng, L., Morris, J. S. nad Aldaz, C. M. (2003) "Differential expression in SAGE: accounting for normal between-library variation," *Bioinformatics*, 19, 1477-1483.

Baggerly, K. A., Deng, L., Morris, J. S. nad Aldaz, C. M. (2004) "Overdispersed logistic regression For SAGE: Modelling multiple groups and covariates," *BMC Bioinformatics*,

http://www.biomedcentral.com/1471-2105/5/144.

Bartke, A., and Brown-Borg, H. (2004) Life extension in the dwarf mouse. *Current Topics in Developmental Biology*, 63, 189-225.

Beißbarth, T., Hyde, L., Smyth, G. K., Job, C., Boon, W., Tan, S., Scott, H, S., and Speed, T. P. (2004) "Statistical modeling of sequencing errors in SAGE libraries," *Bioinformatics*, 20, 131-139.

Benjamini, Y., Hochberg, Y. (1995) "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *J. R. Statist. Soc.* B. 57, 289-300.

Blackshaw, S., Fraioli, R. E., Furukawa, T., and Cepko, C. L. (2001) "Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes," *Cell*, 107, 579-589.

Brown-Borg, H. M., Zhang, F. P., Huhtaniemi, I., Bartke, A. (1996) "Developmental aspects of prolactin receptor gene expression in fetal and neonatal mice," *Eur. J. Endocrinol.* 134(6), 751-757.

Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C., and Wong, W. H. (2004) " Clustering analysis of SAGE data using a Poisson approach," *Genome Biology, http://genomebiology.com/2004/5/7/R51.*

Chen, R. L., Athauda, S. B. P., Kassem, N. A., Zhang, Y., Segal, M. B., and Preston, J. E. (2005) Decrease in transthyretin synthesis at the blood-cerebrospinal fluid barrier of old sheep. *Journal of Gerontology: Biological Sciences*, 60A, 852-858.

Emanuele, N. V., Metcalfe, L., Wallock, L., Tentler, J., Hagen, T. C., Beer, C. T., Martinson, D., Gout, P.W., Kirsteins, L., and Lawrence, A.M. (1987) Extrahypothalamic brain prolactin: characterization and evidence for independence from pituitary prolactin. *Brain Research*, 421, 255-62.

Emanuele, N. V., Jurgens, J. K., Halloran, M. M., Tentler, J. J., Lawrence, A. M., Kelley, M. R. (1992) The rat prolactin gene is expressed in brain tissue: detection of normal and alternatively spliced prolactin messenger RNA. *Molecular Endocrinology*, 6, 35-42.

Green, P. J. (1995) "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hochberg, Y. (1988) "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, 75, 800802.

Holm, S. (1979) "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

Hulbert, A. J. (2000) Thyroid hormones and their effects: a new perspective. Biological reviews of the Cambridge Philosophical Society, 75, 519-631.

Kal, A. J., van Zonneveld, A. J., Benes, V., vanden Berg, M., Koerkamp, M. G., Albermann, K., Strack, N., Ruijter, J. M., Rochter, A., Dujon, B. *et al.* (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell*, 10, 1859-1972.

Madden, S., Galella, E., Zhu, J., Bertelsen, A. and Beaudry, G. (1997) "SAGE transcript profiles for p53-dependent growth regulation," *Oncogene*, 15, 1079-1085.

Man, M. Z., Wang, X. and Wang, Y. (2000) "POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, 16, 953-959.

Margulies, E. H. and Innis, J. W. (2000) "eSAGE: managing and analyzing data generated with serial analysis of gene expression (SAGE)," *Bioinformatics*, 16, 650-651.

Michiels, E. M. C., Oussoren, E., van Groenigen, M., Pauws, E., Bossuyt, P. M. M., Voûte, P. A. and Baas, F. (1999) "Genes differentially expressed in medulloblastoma and fetal brain," *Physiol. Genomics*, 1, 83-91.

Raftery, A. (1996) "Hypothesis testing and model selection, In Markov Chain Monte Carlo in Practice," Chapman and Hall.

Richardson, S. and Green, P. J. (1998) "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *J. R. Statist. Soc. B*, 59, 731-792.

Ruijter, J. M., Kampen, A. H. C., and Baas, F. (2002) "Statistical evaluation of SAGE libraries: consequences for experimental design," *Phy. Genomics*, 11, 37-44.

Ryu, B., Jones, J., Blades, N. J., Parmigiani, G., Hollingsworth, M. A., Hruban, R. H. and Kern, S. E. (2002) "Relationships and differentially expressed genes among pancreatic cancers examined by large-scale serial analysis of gene expression," *Cancer Res.*, 62, 819-826.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995) "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Scien*, 270, 467-470.

Schwarzman, A. L., Gregori, L., Vitek, M. P., Lyubski, S., Strittmatter, W. J., Enghilde, J. J., Bhasin, R., Silverman, J., Weisgraber, K. H., Coyle, P. K., Zagorski, M. G., Talafous, J., Eisenberg, Saunders, A. M., Roses, A. D., and Goldgaber, D. (1994) Transthyretin sequesters amyloid beta protein and prevents amyloid formation. *Proceedings of the National Academy of Sciences USA*, 91, 8368-8372.

Sun, Y., Jin, K., Mao, X. O., Xie, L., Peel, A., Childs, J. T., Logvinova, A., Wang, X,, and Greenberg, D. A. (2005) Effect of aging on neuroglobin expression in rodent brain. *Neurobiology of Aging*, 26, 275-278.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) "Serial analysis of gene expression," *Science*, 270, 484–487.

Vencio, R.Z.N., Brentani, H., Patrao, D.F.C., Pereia, C.A.B. (2004) "Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE)", *BMC Bioinformatics*, 5, 119.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. (1997), "Gene expression profiles in normal and cancer cells," *Science*, 276, 1268-1272.

Table 1: Results for simulated data. Segment I gives the results from the OT algorithm, with those for the two-group setting on the left and the three-group setting on the right; the positive tags are those with BF > 55. Segment II shows the results from the t-test (left) and the Z-test (right) by Kal *et al.* (1999) under the two-group setting; the positive tags are those with $p < 0.01$.

I		Two Group (OT)			Three Group (OT)		
	simulated	S_1	S_2	Total	S_1	S_2	Total
	Positive	49(98%)	3(0.55%)	52	99(99%)	2(0.4%)	101
	Negative	1(2%)	543(99.45%)	544	1(1%)	494(99.6%)	495
	Total	50	546	596	50	546	596

II		Two Group (t-test)			Two Group (Z-test)		
	simulated	S_1	S_2	Total	S_1	S_2	Total
	Positive	43(86%)	7(1.28%)	50	49(98%)	10(1.83%)	59
	Negative	7(14%)	539(98.72%)	546	1(2%)	536(98.17%)	537
	Total	50	546	596	50	546	596

Table 2: Results from sensitivity analysis based on the two-group simulated data. Segment I gives the powers and type I error rates (%) with five different prior odds ($\lambda/(1+\lambda)$). Segment II gives the numbers of tags individually and commonly ("Common" column) identified as differentially expressed using three different t parameter values in the priors for θ_M. Two different BF thresholds with the corresponding posterior probabilities (PP) were entertained.

I	Prior Odds	0.01	0.05	0.1	0.15	0.5
	Power	98	96	98	96	98
	Type I Error	0.55	0.55	0.55	0.92	0.92

II	BF	PP	t=0.5	t=1.0	t=1.5	Common
	100	0.909	49	51	49	49
	55	0.846	50	52	50	49

Table 3: Selected tags and the corresponding unigene IDs and gene symbols. The top half gives tags up-regulated in the aged mice while the bottom half shows those down-regulated. The genes with an * are those not involved in any of the displayed GO categories in Figure 3.

Tag	Unigene-ID	Gene symbol (or name)
GGCATCTCTT	314 / 319830*	Galnt4 / -Transcript sequence*
TGTATAAAAA	87773 / 246377	Tra1 / Tubb2
ATAATACATA	200362	Cybb
AAAAAAAAAA	292145* / 272120	Gypc* / Gad1
TAAAAAAAAA	286177* / 299512*	Serf1* / Igh-1a*
ATTTTCAGTT	unknown	unknown
TCCCTATTAA	unknown	unknown
TGGATCCTGA	288567	Hbb-b1/b2
GAAAATGCAT	40059*	A030001O10Rik*
CTTGGGTGCA	1270	Prl
TACAATGTGA	45058	Camk4
TAAAGAGGCC	324741* / 261679	-Transcript sequence* / Rps26/Wwp2
AGCAAAAGCC	217311*	-Transcript sequence*
TGTGTGAGGA	258927 / 334078*	Eef1d / Agpat3*
TGTGTTGTGT	220038	Ddx5
AATTCGCGGA	2108	Ttr
ACCAATGAAC	218473	Tde1

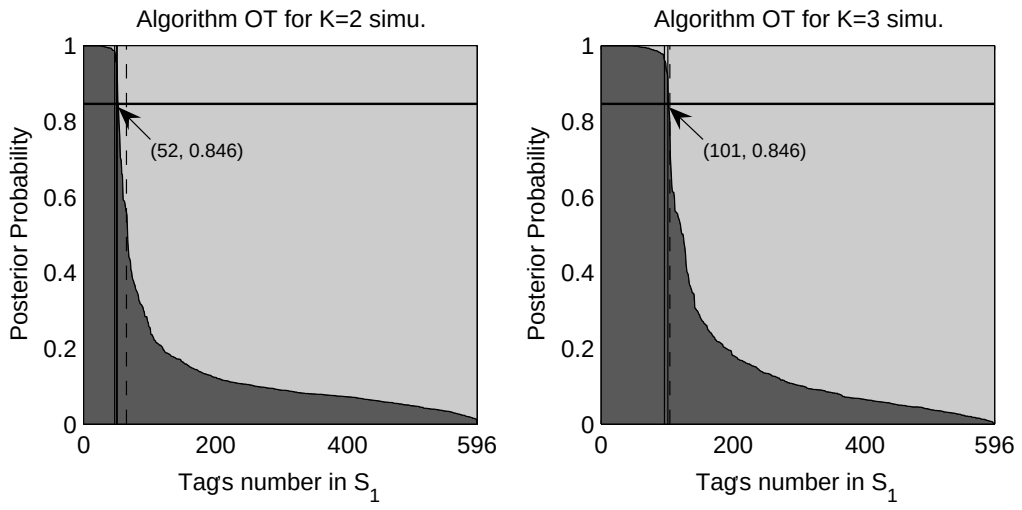

Figure 1: Estimated posterior probabilities (PP) of being DE (dark gray) or SE (light gray) for each tag using the OT algorithm. The tags are arranged in descending order according to PP of DE. The horizontal line segment in each plot is the threshold (BF=55, or equivalently, PP=0.846 for prior odds of 0.1) used to determine whether a tag should be flagged as differentially expressed. The number in the parentheses gives the number of positive tags. False positive (solid line) or false negative (dashed line) tags are identified by the vertical line segments. The left and right plots are for the two-group and the three-group simulated data, respectively.

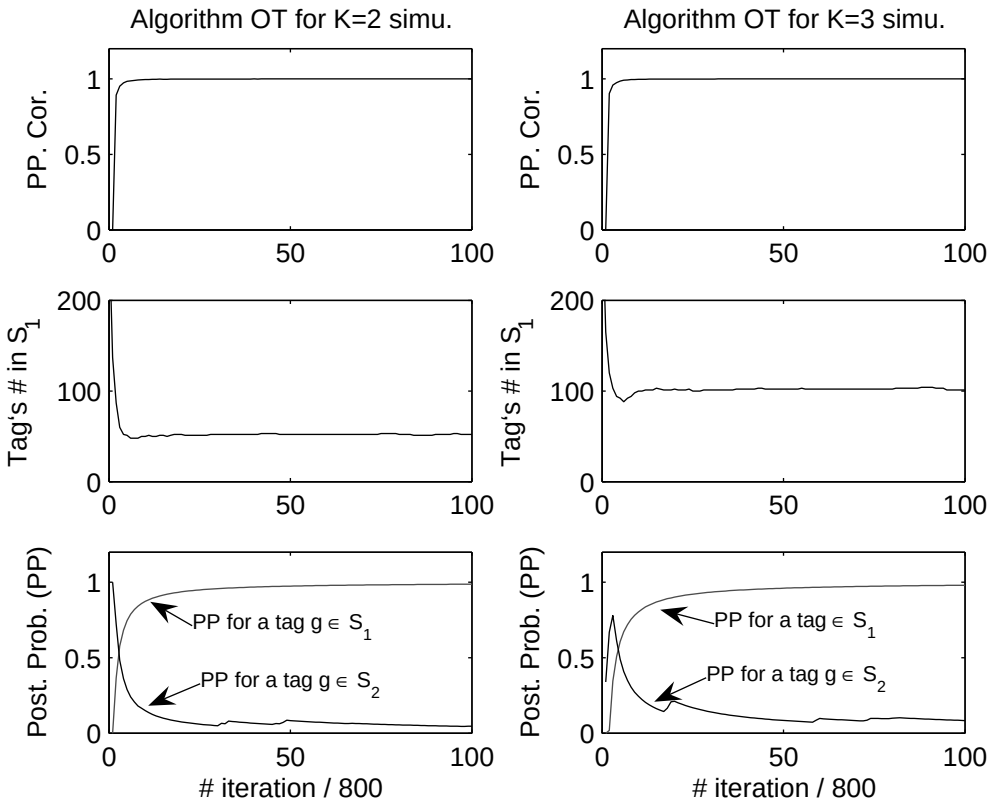

Figure 2: Three types of diagnostic plots with respect to every L (800) MCMC iterations using the OT algorithm. Top row: correlation between consecutive posterior probability estimates; middle row: number of tags in DE set S_1; bottom row: trace plots of posterior probabilities in S_1 or S_2 of two representative tags. The left column shows the results for the two-group simulated data while the right column gives the corresponding ones for the three-group data.

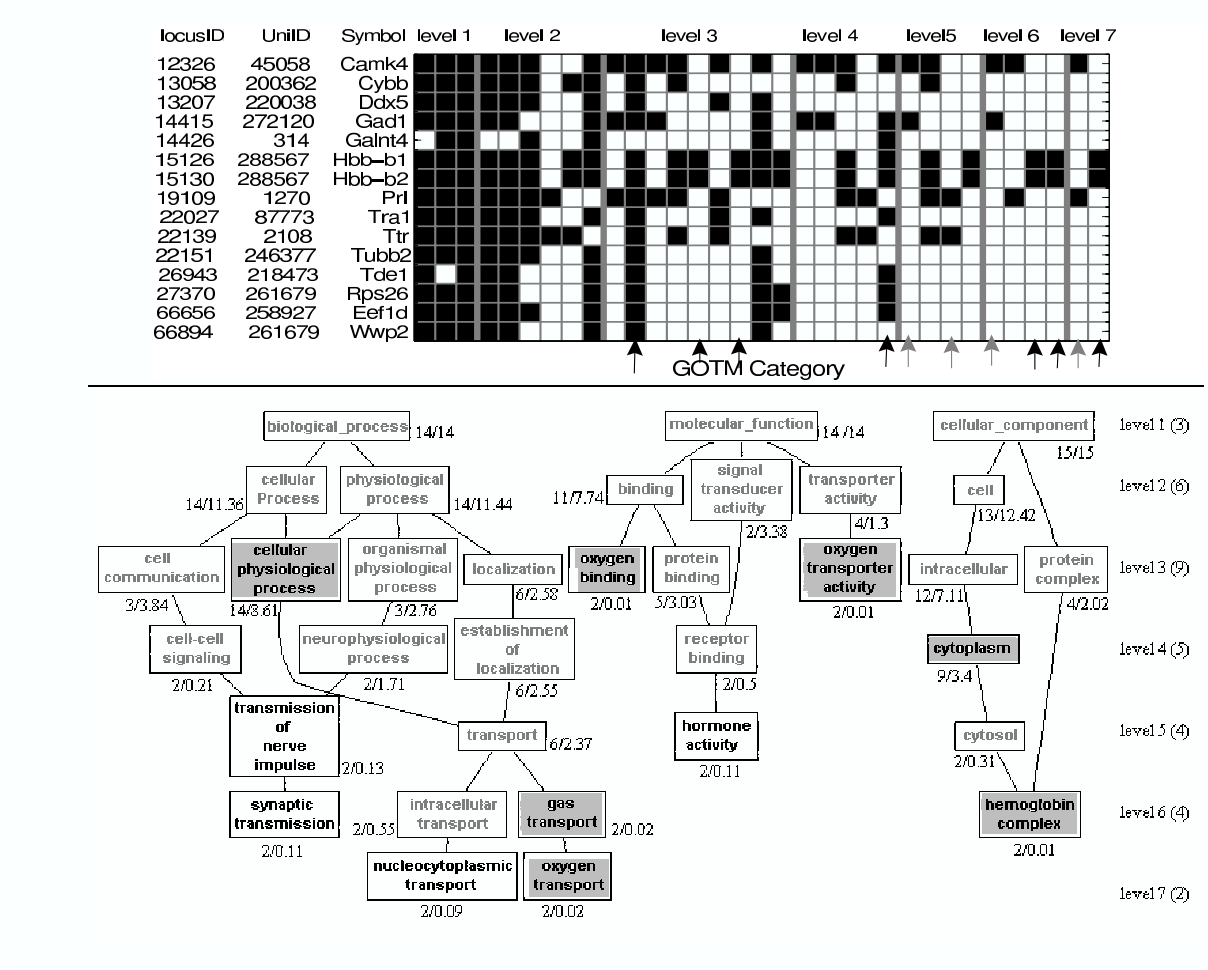

Figure 3: Top panel: Gene list for each category in the bottom panel. Each row represents one gene, with a black square denoting the involvement of the gene in the corresponding column (category). Each column stands for one category ordered first by levels. Within each level, the categories are ordered from left to right according to the display in the bottom panel. The black arrows point to the categories being enriched even after the multiplicity adjustment, while the light gray ones indicate those enriched only with the raw p-values. Bottom panel: A directed acyclic graph view of the enriched GO categories in our list of selected genes. The GO categories in black and with shading are enriched GO categories (with raw p-value < 0.01 an FDR $p < 0.05$) while those in black but without shading are enriched GO categories meeting only the raw p-value criterion. The gray categories are their non-enriched ancestors up to the top level. The numbers around each category are the observed/expected gene numbers for that category. The number of categories within each level is in the parentheses.