

# Modeling and Analysis of SAGE Libraries

Zailong Wang and Shili Lin

October 20, 2005

---

---

Zailong Wang is Postdoctoral Researcher of Mathematical Biosciences Institute, The Ohio State University, 231 West 18th Avenue, Columbus, OH, 43210 (Email: [zlwang@mbi.osu.edu](mailto:zlwang@mbi.osu.edu)). Shili Lin is Professor, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210 ([shili@stat.ohio-state.edu](mailto:shili@stat.ohio-state.edu)). SL's work was supported in part by NSF grant DMS-0306800. The authors would like to thank Professor Andrej Rotter and Dr. Magdalena Popesco for introducing us to SAGE and kindly sharing their data with us. The supplementary material and the programs implementing the algorithms may be accessed at <http://www.stat.ohio-state.edu/~statgen/SOFTWARE/DE-SAGE/>

## ABSTRACT

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 with 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. 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, Metropolis-Hastings algorithm, 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 “close” 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 errors, a small proportion of the tags do 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 various 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 (Popesco et al. 2005). 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 similarly naming scheme was applied to the three aged mice, P810(1), P810(2), and P840. Note that two of the aged mice were both sacrificed at postnatal day of 810. The total number of tags in each of these six libraries varies, as shown on the first row of table 1, while the numbers of unique tags are given on the second row of the same table. 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 are less than three. Since interests are usually focused on “high abundance” tags (Popesco et al. 2005), 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 with counts greater than or equal to 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) introduced a beta-binomial model and suggested a modified t-statistic, while Baggerly *et al.* (2004) 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 software accompanying this paper allows for user selected threshold value to choose a set of putative DE tags for further experimental validation.

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. Section 4 reports the analysis and results of our mouse data, while a few concluding remarks are given in Section 5. Technical details are deferred to the appendix.

## 2 METHODS

### 2.1 Hierarchical Bayesian Modeling

Let  $X = \{X_{kig}, k = 1, \dots, K; i = 1, \dots, n_k; g = 1, \dots, G\}$  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.

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)  $k, i = 1, \dots, n_k$ . 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.  $p_{1g} = p_{2g} = \dots = p_{Kg} := p_g$ . 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:

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

where  $f_j(\cdot|\theta_{jg})$  is a given parametric family of densities indexed by a vector parameter  $\theta_{jg}$ , and  $w_{jg}, j = 1, 2$  denote the mixing proportions of the gene being differentially expressed or not. Specifically, the parameter vector in the two component densities are  $\theta_{1g} = \{p_{1g}, \dots, p_{Kg}\}$ , and  $\theta_{2g} = \{p_g\}$ , respectively.

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.,  $p_{k_1g} \neq p_{k_2g}$  for at least two different groups  $k_1$  and  $k_2$ . The remaining tags which 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  $\theta_M = \theta_1 \cup \theta_2 = (\cup_{g \in S_1} \{p_{1g}, \dots, p_{Kg}\}) \cup (\cup_{g \in S_2} \{p_g\})$ , 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  $\theta_M$  under hyperparameter vector  $\delta_M$ , which is in turn specified by a hyperprior with known parameters. The joint posterior distribution for all the parameters is then factored into

$$P(M, \theta_M, \delta_M|X) \propto L(\theta_M, M)P(\theta_M|\delta_M, M)P(\delta_M|M)P(M). \quad (1)$$

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

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

where  $\bar{N}_k = n_k^{-1} \sum_{i=1}^{n_k} N_{ki}$  and  $\bar{N} = n^{-1} \sum_{k=1}^K \sum_{i=1}^{n_k} N_{ki}$ . The hyperparameter vector  $\delta_M = \{\alpha_{kg}, k = 1, \dots, K, g \in S_1; \alpha_g, g \in S_2\}$  are themselves assumed to be independently distributed as truncated Gamma's:

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

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

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

and we may control the parameter  $\lambda$  by setting  $E\|M\| = G\lambda \stackrel{\text{say}}{=} 50$ . Equivalently, we may obtain the value of parameter  $\lambda$  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  $\theta_M$  and  $\delta_M$  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 the Appendix. In what follows, we present two algorithms for combining the two types of MCMC updates. These two algorithms differ in whether one tag or multiple tags are 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  $\theta_M$  under model  $M$ : The parameters in  $\theta_1$  and  $\theta_2$  will be updated in parallel according to the M-H algorithms described in 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 (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. In the following algorithm, we describe an algorithm that considers all tags in each iteration to ascertain their potential switch of memberships in the two complementary sets.

**Algorithm: All-Tag (AT)**

- Steps 1 and 2. Same as for the OT algorithm.
- Step 3. Update model  $M$ : Perform the following tasks for each tag  $g$  in  $S_1$  and  $S_2$ .
  - (a). If  $g \in S_2$  ( $S_1$ ), calculate the acceptance probability of adding (deleting) it to (from) the DE set  $S_1$ .
  - (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) until convergence.

### **2.3 Computational time consideration, convergence diagnostics, and decision making**

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 the relative performances of the two algorithms, including their rates of convergence. To make the comparisons as fair as possible, we suggest estimating the running time per iteration for the OT and the AT algorithms, and then setting the numbers of iterations of the two algorithms to be inversely proportional to their per iteration running time.

Three types of simple diagnostic plots are utilized for evaluating the absolute and relative performances of the algorithms. 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 set of plots are convergence diagnostics for each gene. They are again trace plots of the posterior probability 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 table 2. With the prior odds set to be 0.1, we ran both the OT and AT algorithms for 80,000 and 1,000 iterations, respectively. Since OT runs about 80 times faster than AT for this dataset, these numbers of iterations render the computational times roughly equal for the two algorithms, which were about 24 minutes on a Pentium 4, 2.4GH PC with 512MB of RAM. As can be seen from the table, both algorithms perform comparably, with over 95% power for identifying tags that are in the DE set (49/50 for OT and 48/50 for AT), and small false positive rates (3/546 for OT and 2/546 for AT). The top half of Figure 1 shows the estimated posterior probabilities in  $S_1$  (green) and  $S_2$  (yellow) for each of the 596 tags, arranged in descending order according to their posterior probabilities in  $S_1$ . The posterior probabilities for the false negatives (dashed lines) and the false positives (solid lines) are indicated in blue in the plots. We can see from the plots that the false positives are among the positives with the smallest probabilities while the false negatives from the AT algorithm are close to making the cutoff as positives.

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 Table 3. 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 either the OT or the AT algorithms. 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  $L$  ( $L=800$  for OT and 10 for AT) 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 plotted them on the top row of Figure 2. Both algorithms seemed to have provided consistent posterior probability estimates after  $20*L$  iterations, although AT appears to have converged slightly faster. In terms of the number of tags declared positive, both again converged rather fast, with AT outperforming OT again (second row of Figure 2). Using a representative tag from  $S_1$  and one from  $S_2$ , we show on the last row of Figure 2 the trace plots of the respective posterior probabilities. Not surprisingly, this set of plots show that the posterior probability settled down slightly faster for the AT algorithm than for the OT algorithm, consistent with the information contained in the other two sets of plots. Trace plots for all the other tags show similar patterns and are thus omitted here. Overall, both algorithms behave rather similarly after adjusting for differential computing times.

## 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.

Both the OT and the AT algorithms were applied to this three-group simulated dataset. We ran 80,000 and 1,000 iterations for the OT and AT algorithms, respectively, which took about 28 minutes each (using the same computer as described before). The outcomes are given in Table 4. They are comparable to those from the two group data, with high powers and low false positive rates. The posterior probabilities of each tag being in  $S_1$  are plotted on the bottom row of Figure 1, with the false positives and false negatives identified by blue. As can be seen from Figure 1, both the OT and the AT algorithms performed similarly in terms of posterior probability estimates for this three-group setting. Diagnostic plots as those in Figure 2 were also generated; they convey essentially the same information as those given by the two group scenario and are thus omitted here.

### 3.3 Sensitivity Analysis

Since results from our Bayesian formulation are dependent on the choice of priors for the model parameter vector  $\theta_M = (\cup_{g \in S_1} \{p_{1g}, \dots, p_{Kg}\}) \cup (\cup_{g \in S_2} \{p_g\})$ , we study the degree of sensitivity of our method by considering a class of beta priors for these parameters. Specifically, 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 priors used in our simulation study in the previous two subsections, we also considered  $t = 0.5$  and  $t = 1.5$  for both simulated datasets ( $K = 2$  and  $K = 3$ ) using the OT algorithm. Table 5 gives the numbers of tags that are identified as DE under several BF cutoff values. The columns with header ‘‘Common’’ give the numbers of tags commonly selected by using all three different priors. These results indicate that, for the simulated data sets, the method is quite robust to the specification of the priors, especially for BF cutoff values of at least 20.

## 4 RESULTS

We now return to the real SAGE libraries in the mouse cerebellum study described in Section 1. We applied both the OT and AT algorithms (with the same numbers of iterations as those

for the simulated data) to the data 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 and 19 tags were selected by OT and AT, respectively, with 17 in common in both lists. Should more or fewer genes are desired by an investigator for further analysis, then one can adjust the BF cutoff to achieve that, for example, from Figure 3, which plots the posterior probabilities of being DE or SE. Nine of these tags, displayed in the first half of table 6, 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 6, 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.

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 specific 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 4, all GO categories that were identified to be significantly enriched (raw  $p < 0.01$ ; with category names in red), together with their ancestral categories (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. The full GO tree can be found in the supplementary material from our website. 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). A cutoff of 0.05 for the adjusted p-values lead to a number of enriched categories no longer being enriched (shaded ones in Figure 4). 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 enriched list. A full list of the adjusted p-values can be found in the supplementary material.

As can be seen from Figure 4, several of the enriched categories were “oxygen” related. This observation lends itself to detailed annotation of the genes involved. For instance, the two genes in the enriched category “oxygen transport” are Hbb-b1 (hemoglobin, beta adult major chain), Hbb-b2 (hemoglobin, beta adult minor chain), which are highly expressed in the aged cerebella. This finding confirms the current understanding of the biological process of aging. As we process oxygen, we also produce toxic molecules, called free radicals, along the way. These can damage DNA, proteins, and mitochondria, the so-called “powerhouses” of cells. Free radicals (and the oxidative damage they cause) are believed to be major factors in much of the cellular and tissue deterioration that occurs with aging (<http://www.infoaging.org/b-cal-10-r-pop.html>). It is interesting to note that these genes are the same ones as those in the enriched categories of “oxygen binding”, “oxygen transporter activity”, and “hemoglobin complex”.

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). This is again consistent

with current biological understanding, since some of the physiological manifestations of aging are related to the effects of declining hormone levels. Brown-Borg *et al* (1996) suggested a possible role for Prl in the development of the immune system. They also suggested that Prl is among several factors necessary to coordinate developmental activities. This implies that the decrease of Prl expression is related to the aging process. More importantly, genes that are involved in each enriched categories are all either up-regulated or down-regulated in the aged group.

## 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 satisfactory. 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 sensible results. In particular, genes that are identified to be more highly expressed in the aged cerebella are involved in producing toxic molecules associated with aging, while genes that have reduced expression levels are involved in coordinating activities. These findings are consistent with the hypothesis that the cerebellum becomes progressively dysfunction during aging, leading to deficiencies in balancing and motor coordination. 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. 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. The programs that implement both of these algorithms can be downloaded freely from our website provided in the authors footnote section.

Finally, we note that although our method was proposed 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.

## 6 APPENDIX

In this appendix, we provide technical details of our MCMC sampling algorithms, including the posterior distributions for each of the parameters, and the M-H algorithms for updating these parameters under a specified model,  $M$ . Specifically, the univariate full conditional distribution for each parameter associated with tags in either the DE set or the SE set is given in A.1. The M-H algorithms for updating the two types of parameters (priors or hyperpriors) are provided in A.2. Finally, in A.3, we supply the details of the reversible jump MCMC algorithm for updating the model  $M$ , including the acceptance probabilities for adding/deleting.

## A.1. Conditional distributions of parameters for a given $M$

Since conditional on the model  $M$ , the tags within a library are assumed to be independently distributed, we can consider updating the parameters associated with each tag separately. For  $g \in S_1$ , denote  $\theta_g = \{p_{1g}, \dots, p_{Kg}\}$  for the associated parameters and  $\delta_g = \{\alpha_{1g}, \dots, \alpha_{Kg}\}$  for the hyperparameters. Using the priors and hyperpriors as specified in the main text, we can easily derive the full conditional distributions (up to a normalizing constant) for these parameters. For  $k = 1, \dots, K$ ,

$$\begin{aligned}\alpha_{kg} | \delta_{-kg}, \theta_g, M &\sim \frac{\alpha_{kg}^{\alpha_{kg}-1} e^{-b_k \alpha_{kg}}}{\Gamma(\alpha_{kg}) \Gamma(\bar{N}_k - \alpha_{kg})} p_{kg}^{(\alpha_{kg}-1)} (1 - p_{kg})^{(\bar{N}_k - \alpha_{kg} - 1)} := f(\alpha_{kg}); \\ p_{kg} | \theta_{-kg}, \delta_g, M &\sim p_{kg}^{\alpha_{kg} + \sum_{i=1}^{n_k} X_{kig} - 1} (1 - p_{kg})^{\bar{N}_k - \alpha_{kg} - 1} e^{-(\sum_{i=1}^{n_k} N_{ki}) p_{kg}} := f(p_{kg});\end{aligned}$$

where  $\delta_{-kg}$ , and  $\theta_{-kg}$  are the parameter vectors  $\delta_g$  and  $\theta_g$  without the element  $\delta_{kg}$  or  $\theta_{kg}$  respectively.

Similar to that given above, we can derive the conditional distributions for the parameter and hyperparameter for each gene  $g$  in  $S_2$ , which are (again up to a constant):

$$\begin{aligned}\alpha_g | p_g, M &\sim \frac{\alpha_g^{\alpha_g-1} e^{-b \alpha_g}}{\Gamma(\alpha_g) \Gamma(\bar{N} - \alpha_g)} p_g^{(\alpha_g-1)} (1 - p_g)^{(\bar{N} - \alpha_g - 1)} := f(\alpha_g); \\ p_g | \alpha_g, M &\sim p_g^{\alpha_g + \sum_{i=1}^n X_{ig} - 1} (1 - p_g)^{\bar{N} - \alpha_g - 1} e^{-(\sum_{i=1}^n N_i) p_g} := f(p_g).\end{aligned}$$

Since these distributions are not from any well known families of distributions and are only known up to a constant, we use M-H algorithms to sample from them as detailed next.

## A.2. M-H Algorithms for updating parameters

**Updating the  $\alpha$  hyperparameters:**  $\{\alpha_{1g}, \dots, \alpha_{Kg}, \alpha_g\}$ .

The following scheme works for any  $\alpha$  parameter in the set. Suppose that the current value for  $\alpha$  is  $\alpha^{(n)}$ . Set the proposal distribution as  $\alpha^* | \alpha^{(n)} \sim \text{Unif}(\alpha^{(n)} / d\alpha, \min\{\tilde{N}, \alpha^{(n)} \cdot d\alpha\})$ , where  $d\alpha > 1$  is a pre-specified constant (say  $d\alpha = 2$ ). Also,  $\tilde{N} = \bar{N}$  for  $\alpha = \alpha_g$  and  $\tilde{N} = \bar{N}_k$  for  $\alpha = \alpha_{kg}$ . Using this proposal distribution, the importance ratio is

$$\begin{aligned}
r &= \frac{f(\alpha^*) q(\alpha^{(n)}|\alpha^*)}{f(\alpha^{(n)}) q(\alpha^*|\alpha^{(n)})} \\
&= \frac{\Gamma(\alpha^{(n)}) \Gamma(\bar{N} - \alpha^{(n)})}{\Gamma(\alpha^*) \Gamma(\bar{N} - \alpha^*)} \left(\frac{\alpha^*}{\alpha^{(n)}}\right)^{a-1} \left(\frac{p e^{-b}}{1-p}\right)^{\alpha^* - \alpha^{(n)}} \frac{d\alpha \min\{\tilde{N}, \alpha^{(n)} d\alpha\} - \alpha^{(n)}}{d\alpha \min\{\tilde{N}, \alpha^* d\alpha\} - \alpha^*}.
\end{aligned}$$

Hence, with probability  $\min\{1, r\}$ ,  $\alpha^{(n+1)} = \alpha^*$ . Otherwise  $\alpha^{(n+1)} = \alpha^{(n)}$ .

**Updating the  $p$  parameters:**  $\{p_{1g}, p_{Kg}, p_g\}$ .

The following algorithm applies to each of the  $p$  parameters in the list. Let  $p^*|p^{(n)} \sim \text{beta}(2, \frac{2}{p^{(n)}} - 2)$  where  $p^{(n)}$  is the current value for the  $p$  parameter. The importance ratio is

$$\begin{aligned}
r &= \frac{f(p^*) q(p^{(n)}|p^*)}{f(p^{(n)}) q(p^*|p^{(n)})} \\
&= \frac{(2-p^*)}{(2-p^{(n)})} \frac{p^{*(\alpha + \sum X_i - 4)} (1-p^*)^{(\bar{N} - \alpha + 3 - 2/p^{(n)})}}{p^{(n)(\alpha + \sum X_i - 4)} (1-p^{(n)})^{(\bar{N} - \alpha + 3 - 2/p^*)}} e^{-\sum N_i(p^* - p^{(n)})}.
\end{aligned}$$

We set  $p^{(n+1)} = p^*$  with probability  $\min\{1, r\}$ . Otherwise  $p^{(n+1)} = p^{(n)}$ .

### A.3. Reversible jump MCMC

Suppose tag  $g$  is being considered for deletion from the DE set  $S_1$ . The new parameters associated with this tag,  $p_g$  and  $\alpha_g$ , can be obtained from the following two equations, based on the values of the current parameters, as follows:

$$\sum_{k=1}^K p_{kg} = K p_g, \quad \sum_{k=1}^K \frac{\alpha_{kg}}{\bar{N}_k} = K \frac{\alpha_g}{\bar{N}}.$$

In addition, we need to generate  $2(K-1)$  new parameters for matching the dimensions (Green 1995). Specifically, we set  $\lambda_{kg} = a_{kg}/b_k$ , where  $a_{kg}$  and  $b_k$  are as defined in the main text, and generate these parameters using the following schemes:

$$\begin{aligned}
u_{kg} &\sim \text{Beta}(\lambda_{kg}, \bar{N}_k - \lambda_{kg}) \mathbf{1}_{\{\max\{0, (Kp_g - 1)/(K-1)\}, \min\{1, Kp_g/(K-1)\}\}}, \\
v_{kg} &\sim \text{Gamma}(a_{kg}, b_k) \mathbf{1}_{\{\max\{0, \bar{N}_k(K\alpha_g/\bar{N} - 1)/(K-1)\}, \min\{\bar{N}_k, \bar{N}_k K\alpha_g/((K-1)\bar{N})\}\}}, \\
&k = 1, \dots, K-1.
\end{aligned} \tag{3}$$

The proposal of adding a tag  $g$  to the DE set, the counterpart of deleting a tag from the set as defined above, can be achieved by setting the new parameters associated with the tag  $g$  being considered for membership in  $S_1$  as follows:

$$p_{kg} = u_{kg}, \quad k = 1, \dots, K-1, \quad p_{Kg} = Kp_g - \sum_{k=1}^{K-1} u_{kg},$$

$$\alpha_{kg} = v_{kg}, \quad k = 1, \dots, K-1, \quad \alpha_{Kg} = \bar{N}_K \left( \frac{K\alpha_g}{\bar{N}} - \sum_{k=1}^{K-1} \frac{v_{kg}}{\bar{N}_k} \right),$$

where the values for  $u_{kg}, v_{kg}, k = 1, \dots, K-1$  are from the sampling values from the corresponding deleting move, as specified in (3).

The acceptance probability for adding is  $\min\{1, A\}$ , where

$$A =$$

$$P_\lambda \cdot K^2 \left( \prod_{k=1}^K p_{kg}^{\sum_{i=1}^{n_k} X_{kig}} \right) p_g^{-\sum_{i=1}^n X_{ig}} \exp \left( \sum_{i=1}^n N_i p_g - \sum_{k=1}^K \sum_{i=1}^{n_k} N_{ki} p_{kg} \right)$$

$$\times \frac{\Gamma(\bar{N}_K + 1) \Gamma(a_g) \Gamma(\alpha_g) \Gamma(\bar{N} - \alpha_g) I\Gamma(b\bar{N}, a_g) \prod_{k=1}^{K-1} \Gamma(\lambda_{kg}) \Gamma(\bar{N}_k - \lambda_{kg})}{\Gamma(\bar{N} + 1) \Gamma(a_{Kg}) \prod_{k=1}^K \Gamma(\alpha_{kg}) \Gamma(\bar{N}_k - \alpha_{kg}) I\Gamma(b_k \bar{N}_k, a_{kg})}$$

$$\times \frac{b_K^{a_{Kg}} \prod_{k=1}^K \alpha_{kg}^{a_{kg}-1} p_{kg}^{\alpha_{kg}-1} (1-p_{kg})^{\bar{N}_k - \alpha_{kg}-1}}{b^{a_g} \alpha_g^{a_g-1} p_g^{\alpha_g-1} (1-p_g)^{\bar{N} - \alpha_g - 1} \prod_{k=1}^{K-1} u_{kg}^{\lambda_{kg}-1} (1-u_{kg})^{\bar{N}_k - \lambda_{kg} - 1} v_{kg}^{a_{kg}-1}}$$

$$\times \prod_{k=1}^{K-1} \left\{ I\beta(\min\{1, Kp_g/(K-1)\}; k) - I\beta(\max\{0, (Kp_g - 1)/(K-1)\}; k) \right\}$$

$$\times \left\{ I\Gamma(b_k \cdot \min\{\bar{N}_k, \bar{N}_k K \alpha_g / ((K-1)\bar{N})\}; a_{kg}) \right.$$

$$\left. - I\Gamma(b_k \cdot \max\{0, \bar{N}_k (K \alpha_g / \bar{N} - 1) / (K-1)\}; a_{kg}) \right\}$$

$$\times \exp \left( b\alpha_g + \sum_{k=1}^{K-1} b_k v_{kg} - \sum_{k=1}^K b_k \alpha_{kg} \right).$$

For the corresponding deleting proposal, the acceptance probability is  $\min\{1, A^{-1}\}$  with the same expression for  $A$  as above. Here  $P_\lambda = \frac{P(M||M||=s)}{P(M||M||=s-1)} = \frac{\lambda}{1-\lambda}$  is the prior ratio for model  $M$  (see equation (2));  $I\beta(x, k) = \int_0^x t^{\lambda_{kg}-1} (1-t)^{\bar{N}_k - \lambda_{kg} - 1} dt / B(\lambda_{kg}, \bar{N}_k - \lambda_{kg})$  is an incomplete beta function; and  $I\Gamma(x, a) = \int_0^x t^{a-1} e^{-t} dt / \Gamma(a)$  is an incomplete gamma function.

## 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. and 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. and Aldaz, C. M. (2004) "Overdispersed logistic regression For SAGE: Modelling multiple groups and covariates," *BMC Bioinformatics*, <http://www.biomedcentral.com/1471-2105/5/144>.
- 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>.

- 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.
- 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.
- Popesco, M., Wang, Z., Frostholm, A., Friedman, L., Lin, S., and Rotter, A. (2005) Serial Analysis of Gene Expression rprofiles in the adult and aged mouse cerebellum. Preprint.
- 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.
- Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) “Serial analysis of gene expression,” *Science*, 270, 484–487.
- 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: Tag counts in the three adult and three aged SAGE libraries

Mouse	Adult			Aged		
	P92	P150	P300	P810(1)	P810(2)	P840
Total tags	16,430	18,103	10,578	18,581	8,528	7,630
Unique tags	7,144	8,420	6,416	10,544	4,989	3,716

Table 2: Results for the two-group simulated data: MCMC approach. The positive tags are those with  $\text{BF} > 55$ .

simulated	Algorithm OT			Algorithm AT		
	$S_1$	$S_2$	Total	$S_1$	$S_2$	Total
Positive	49(98%)	3(0.55%)	52	48(96%)	2(0.37%)	50
Negative	1(2%)	543(99.45%)	544	2(4%)	544(99.63%)	546
Total	50	546	596	50	546	596

Table 3: Results for the two-group simulated data: the t-test and the Z-test suggested by Kal *et al* (1999). The positive tags are those with  $p < 0.01$ .

simulated	t-test			Z-test		
	$S_1$	$S_2$	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 4: Results for the three-group simulated data: MCMC approach. The positive tags are those with  $\text{BF} > 55$ .

simulated	Algorithm OT			Algorithm AT		
	$S_1$	$S_2$	Total	$S_1$	$S_2$	Total
Positive	99(99%)	2(0.4%)	101	98(98%)	1(0.2%)	99
Negative	1(1%)	494(99.6%)	495	2(2%)	495(99.8%)	497
Total	100	496	596	100	496	596

Table 5: Results from our sensitivity analysis with both the two-group and the three-group simulated data. Several BF cutoff values, with the corresponding posterior probabilities (PP) also shown in the table, were entertained. The “Common” column gives the number of tags commonly selected by using different priors.

$BF$	PP	Two Group Data (K=2)				Three Group Data (K=3)			
		t=0.5	1.0	1.5	Common	t=0.5	1.0	1.5	Common
100	0.909	49	51	49	49	100	101	101	100
55	0.846	50	52	50	49	100	101	102	100
20	0.667	56	57	59	53	104	106	109	104
10	0.500	62	66	64	60	112	124	122	111

Table 6: 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 4.

Tag	Unigene-ID	Gene symbol (or name)
GGCATCTCTT	314 / 319830*	Galnt4 / -Transcript sequence*
TGTATAAAAA	87773 / 246377	Tra1 / Tubb2
ATAATACATA	200362	Cybb
AAAAAAAAAAA	292145* / 272120	Gypc* / Gad1
TAAAAAAAAAA	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
AGCAAAGGCC	217311*	-Transcript sequence*
TGTGTGAGGA	258927 / 334078*	Eef1d / Agpat3*
TGTGTTGTGT	220038	Ddx5
AATTCGCGGA	2108	Ttr
ACCAATGAAC	218473	Tde1

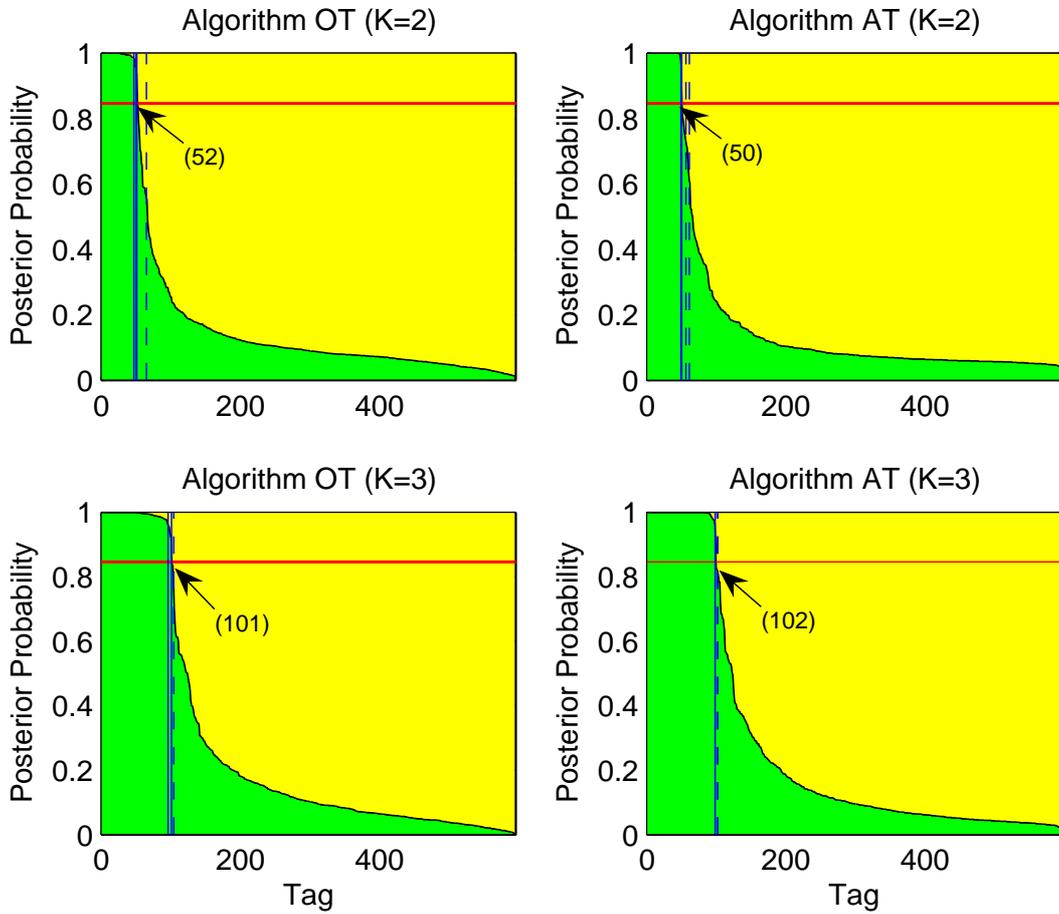


Figure 1: Estimated posterior probabilities (PP) of being DE (green) or SE (yellow) for each tag. The tags are arranged in descending order according to PP of DE. The red line segment in each plot is the threshold used (BF=55, or equivalently, PP=0.846 for prior odds of 0.1) to determine whether a tag should be flagged as differentially expressed. The number in the parentheses gives the number of positive tags. False positive (FP: solid line) or false negative (FN: dashed line) tags are identified by the blue line segments. The top and bottom panels give the results for the two-group and the three-group simulated data, respectively.

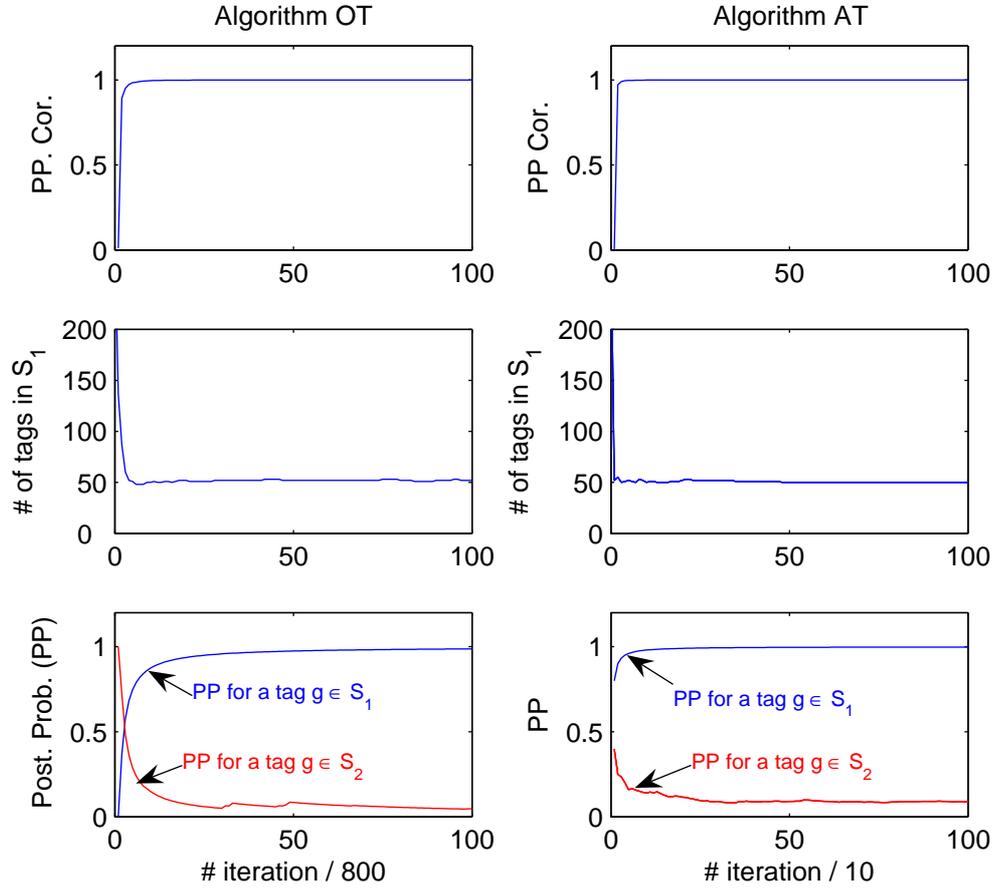


Figure 2: Three types of diagnostic plots with respect to every  $L$  (800 for OT and 10 for AT) MCMC iterations. Top row: correlation between consecutive posterior probability estimates; middle row: number of tags in DE set  $S_1$ ; bottom row: trace plots of posterior probabilities in  $S_1$  of two representative tags. The left column shows the results from the OT algorithm while the right column gives the corresponding ones from the AT algorithm.

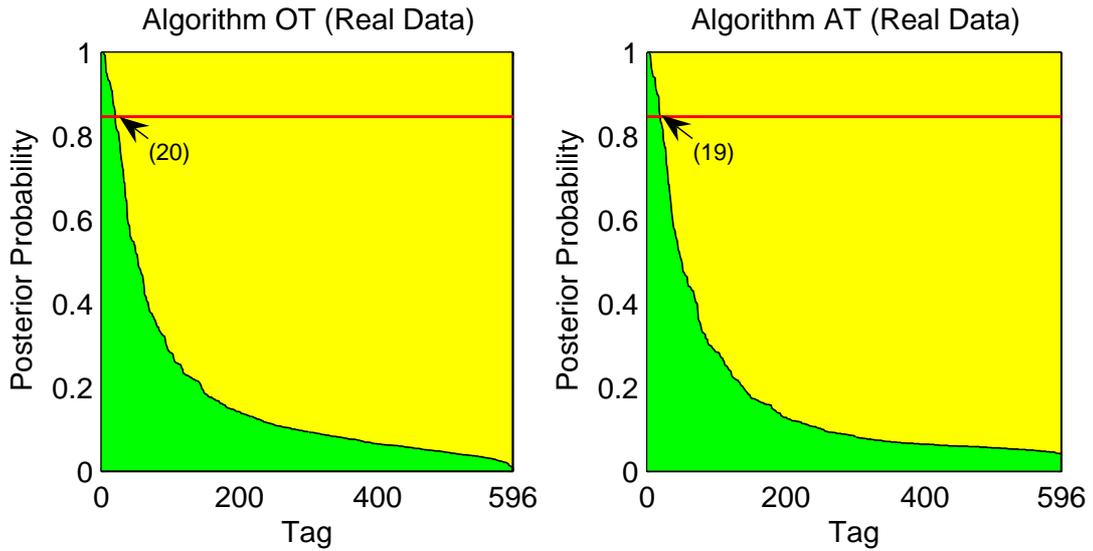


Figure 3: Estimated posterior probabilities (PP) of being DE (green) or SE (yellow) for each tag. The tags are arranged in descending order according to PP of DE. The red line segment in each plot is the threshold used (BF=55, or equivalently, PP=0.846 for prior odds of 0.1) to determine whether a tag should be flagged as differentially expressed. The number in the parentheses gives the number of positive tags.

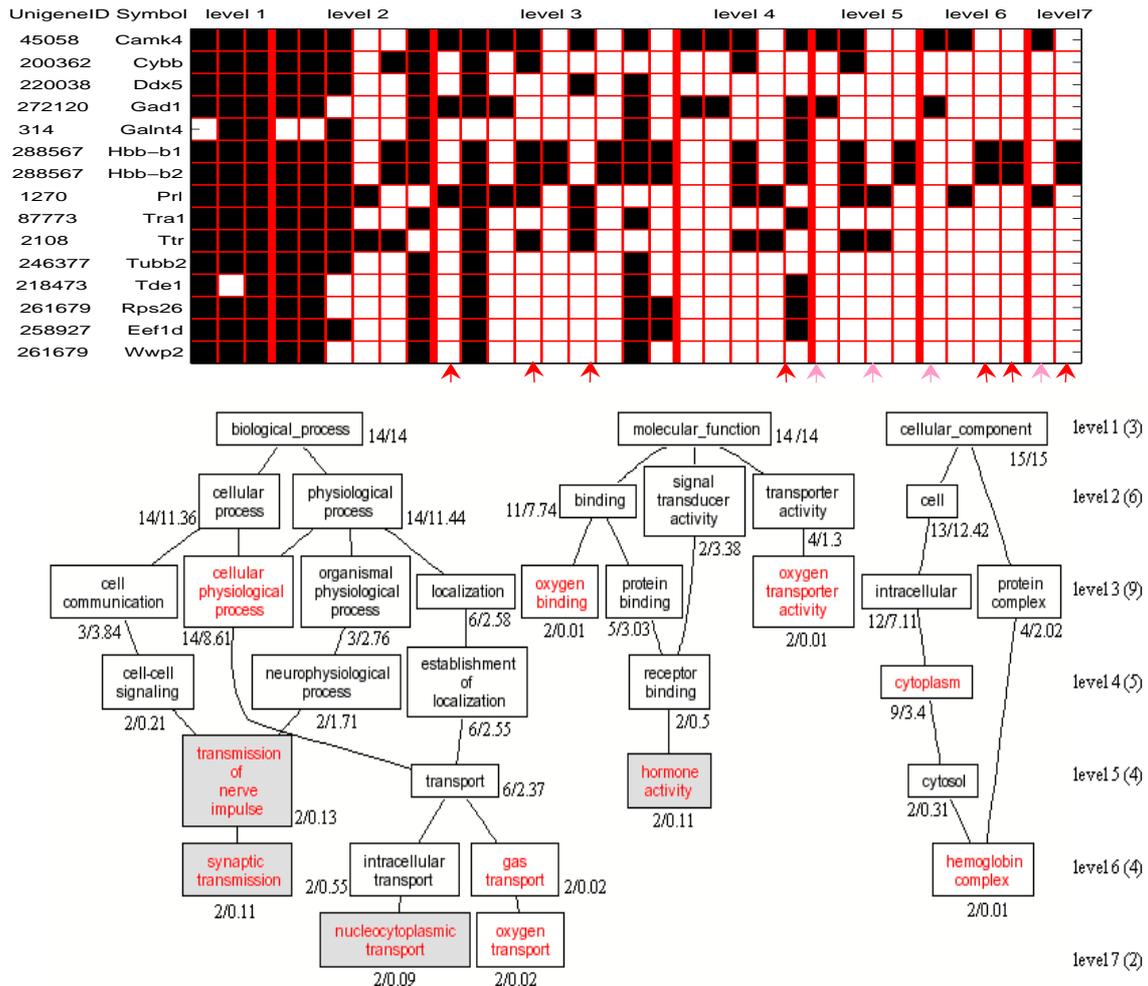


Figure 4: Top panel: Gene list for each category in the bottom panel. Each row represents one gene, with a black square denoting the involvement of the gene in the corresponding column (category). Each column stands for one category ordered first by levels. Within each level, the categories are ordered from left to right according to the display in the bottom panel. The read arrows point to the categories being enriched even after the multiplicity adjustment, while the pink 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 red (without shading) are enriched GO categories (with raw p-value  $< 0.01$  and FDR  $p < 0.05$ ) while those with shading are enriched GO categories meeting only the raw p-value criterion. The black 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.