Published in final edited form as:

Biometrics. 2011 March ; 67(1): 142–150. doi:10.1111/j.1541-0420.2010.01447.x.

Bayesian Hierarchical Modeling and Selection of Differentially Expressed Genes for the EST Data

Fang Yu^{†,*}, Ming-Hui Chen[◇], Lynn Kuo[◇], Peng Huang[‡], and Wanling Yang[§]

[†]Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198-4350

[◇]Department of Statistics, University of Connecticut, Storrs, CT, 06269-4120

[‡]Division of Oncology Biostatistics, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205-2013

[§]Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, China

Summary

Expressed sequence tag (EST) sequencing is a one-pass sequencing reading of cloned cDNAs derived from a certain tissue. The frequency of unique tags among different unbiased cDNA libraries is used to infer the relative expression level of each tag. In this paper, we propose a hierarchical multinomial model with a nonlinear Dirichlet prior for the EST data with multiple libraries and multiple types of tissues. A novel hierarchical prior is developed and the properties of the proposed prior are examined. An efficient Markov chain Monte Carlo algorithm is developed for carrying out the posterior computation. We also propose a new selection criterion for detecting which genes are differentially expressed between two tissue types. Our new method with the new gene selection criterion is demonstrated via several simulations to have low false negative and false positive rates. A real EST data set is used to motivate and illustrate the proposed method.

Keywords

Dirichlet distribution; Gene expression; Mixture distributions; Multinomial distribution; Shrinkage estimators

1. Introduction

Differential gene expression can be investigated via several methods, including EST sequence sampling from cDNA libraries, serial analysis of gene expression (SAGE), and high density DNA microarrays. The EST experiment produces expressed sequence tags (ESTs) (200–300bp) to detect expressed genes. First, mRNA (sequences from expressed genes) was extracted from a given tissue, and reverse transcribed into cDNA. These cDNAs were subsequently subcloned into plasmid vectors and amplified in *E. coli*. The pool of cDNA clones is often called a cDNA library. Each *E. coli* colony represents one original

*fangyu@unmc.edu.

9. Supplementary Materials The Web Appendix referenced in Section 4 is available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

mRNA sequence and each expressed gene could have more than one but different numbers of representative clones. Different numbers of clones for each library were randomly selected and sequenced, and their sequences (EST) were automatically submitted to EST database, which provide rapid identification of the reference sequences for all human genes. A major public EST database (available at NCBI, National Center for Biotechnology Institution) is dbEST, a subsidiary of GenBank. Therefore, by identifying and counting the frequency of appearance of particular gene sequences or EST, it is possible to gain an estimate for relative abundance of each mRNA transcripts. Note that abundant clones are likely to be represented by many ESTs and rare clones are likely to be represented by only a few ESTs. Unlike other gene expression techniques, only the tissue specific transcript frequencies are observed in EST and SAGE. In this paper, we focus on detecting differentially expressed (DE) genes between normal and cancer tissues using the EST data. The statistical methods can also be applied to the SAGE data.

There is a considerable amount of literature on analyzing EST or SAGE data from different libraries. Schmitt et al. (1999) and Romualdi et al. (2001) fit the observed frequency counts with a collapsed 2×2 table and used a Fisher exact test or a χ^2 test to identify DE genes. These approaches are not efficient as a substantial amount of information is lost after collapsing the data. Stekel et al. (2000) employed a likelihood ratio test based on a Poisson model to detect DE genes. Audic and Claverie (1997) and Claverie (1999) evaluated a conditional probability of a frequency count of a gene in one condition given that of the other condition for an equally expressed gene using a Poisson model. They claimed a gene to be DE if its conditional probability is small. To account for the possibility of overdispersion in the tag counts, Baggerly et al. (2004) and Lu et al. (2005) considered an overdispersed logistic regression model and an overdispersed loglinear model, respectively, for the binomial or Poisson sampling assumption for identifying differentially expressed genes using SAGE data. Although EST data provide us rich genetic information, it is difficult to utilize them since only a small proportion of unique tags (genes) are observed under each library. For example, in our study, we consider 9 libraries (6 normal and 3 tumor tissues) sampled from the lymphoreticular or lymph node and observe a total of 8,190 unique genes among these libraries. To obtain a better understanding of the frequency counts in a library, we group the genes if they have the same count and count its size. Then we plot the size (log scale) of each group versus the groups (ordered in count). Figure 1 displays this for a library of size of 1908 tags summed over all groups. We see that the distribution of the group size is highly right-skewed. The largest observed count in this library is 33, while 7,500 (or $\exp(8.923)$) unique genes ($> 91\%$) have zero counts. All other libraries, no matter which tissue type they come from, share a similar pattern. This skewness in the distribution for gene expression counts has also been noted by others including Kuznetsov (2001) and Morris et al. (2003, 2006).

To estimate the expression level of genes which are highly correlated with the frequency count in the EST data, a simple approach is to use the maximum likelihood estimator (MLE) of the true frequency count for each gene. The resulting estimate is the ratio of the observed count on one gene over the corresponding library size. Due to a relatively small library size, a failure to sample certain genes might occur. Hence a “missing” gene does not mean the gene is not expressed at all, and MLE will underestimate them. Also the EST data may

accidentally capture too many copies of some genes which are actually of a low expression level. In this case, the MLE tends to overestimate them. To circumvent these issues of underestimation and overestimation, the following two methods have been proposed recently by Morris et al. (2003). The first method is to fit a multinomial Dirichlet model to the data. Let n denote the library size and G the number of unique tags. The model assumes a $G - 1$ variate symmetric Dirichlet prior distribution with the same parameter θ for the probabilities in the multinomial model. Then it gives a shrinkage estimator with a weight of $\frac{n}{n+G\theta}$ given to the MLE. This approach is simple and easy to implement. However, when the library size n is small relatively to the number of unique genes, which is common in EST data, this estimator works poorly for genes with large counts because too much weights are reassigned from abundant genes (with relatively large observed counts) to scarce genes (with relatively small counts). The second method is to fit a non-linear mixture Dirichlet model (NLMD). It assumes that genes are classified into two classes: abundant and scarce according to their relative frequencies. Two different forms of shrinkage estimators are then obtained for these two types; they not only protect the abundant genes from being shrunk too much, but also steal weights from scarce genes to compensate for the missing genes. Morris et al. (2003) called it “Reverse Robin Hood” and applied it to *only* one-library of a SAGE data.

We extend their method to EST or SAGE data with multiple libraries and multiple tissue types in Sections 2 and 3. We first modify the hyperparameter in their NLMD prior for a single library to be size dependent as it is expected that the value of the total mass increases when the size of the abundant set increases. We then build a hierarchical structure on the hyperparameters of the NLMD to yield more robust posterior estimates. Specifically, at the library level, an NLMD is assumed for each library and the parameters from the different NLMDs share common priors across different libraries so that the information can be borrowed among different libraries within the same tissue type. At the tissue level, an extra layer of priors are assumed to synthesize the information across different tissue types.

In Section 4, an efficient Markov chain Monte Carlo (MCMC) sampling algorithm is developed for carrying out the high-dimensional posterior computation. It also includes several novel computational strategies for reducing computing times and avoiding potential over-floating problems in MCMC sampling. In Section 5, a new gene selection criterion is proposed for detecting DE genes between two different types of tissues. The proposed criterion, easy to implement, has a nice statistical interpretation. Several simulations have been conducted in Section 6. They show that the proposed method has smaller false negative rates than several existing methods while its false positive rates are comparable to the others. In Section 7, a real EST data set is analyzed to illustrate the proposed methodology. A brief discussion is provided in Section 8.

2. Models for EST Data with Multiple Libraries Under Multiple Tissue Types

We apply the NLMD model given by Morris et al. (2003) to each library and use similar notations and parameterizations to theirs, except we add subscripts for libraries and tissues to the abundance probability of each tag in order to explain the hierarchical structure for

sharing information across libraries within each tissue type and then between tissues at a higher level.

Assume each unique tag represents a unique gene. Let G be the number of unique tags (genes) observed in the data. Also let L_t be the total number of libraries of tissue type t for $t = 1, \dots, T$. Then, the observed data can be written as $D = \{X_{tli}, i = 1, \dots, G, l = 1, \dots, L_t, t = 1, \dots, T\}$, where X_{tli} is the number of occurrences of a unique EST tag i in library l from

type t , with its size $n_{tl} = \sum_{i=1}^G X_{tli}$. Further, let $\mathbf{X}_{tl} = (X_{tl1}, \dots, X_{tlG})'$ denote a column vector of frequency for each gene from the library l of type t . Suppose each library size is fixed, then as a consequence of the Poisson model for the frequency count, we can assume that $\mathbf{X}_{tl}$ follows a multinomial distribution with parameters n_{tl} and $\mathbf{p}_{tl} = (p_{tl1}, \dots, p_{tlG})'$, where p_{tli} represents the relative abundance of the i^{th} unique tag in the l^{th} library of type t , and

$\sum_{i=1}^G p_{tli} = 1$, for $\forall l$ and $\forall t$. Then the likelihood function is given by

$L(p_{11}, \dots, p_{TL_T} | D) = \prod_{t=1}^T \prod_{l=1}^{L_t} \binom{n_{tl}}{X_{tl1}, \dots, X_{tlG}} p_{tl1}^{X_{tl1}} \dots p_{tlG}^{X_{tlG}}$. For each library, we use the same NLMD prior as in Morris et al. (2003) for the $\mathbf{p}_{tl}$, except modifying their hyperparameters for the total mass of abundant genes to be size dependent.

To build a hierarchical structure, we assume that the different libraries in the same tissue type share the same indicator functions on whether a unique tag is abundant or scarce. That is, let λ_{ti} be a latent indicator for the i^{th} tag such that $\lambda_{ti} = 1$ if the i^{th} unique tag of type t is abundant and $\lambda_{ti} = 0$ otherwise. Write $\boldsymbol{\lambda}_t = (\lambda_{t1}, \dots, \lambda_{tG})'$. Then, under the same tissue type t , all unique tags are split into two subsets: an “abundant set” $A_t = \{i : \lambda_{ti} = 1\}$ and a “scarce

set” $S_t = \{i : \lambda_{ti} = 0\}$. In addition, let $k_{A_t} = \sum_{i=1}^G \lambda_{ti}$ denote the number of abundant unique tags, and $k_{S_t} = \sum_{i=1}^G (1 - \lambda_{ti})$ the number of scarce unique tags. We further assume that λ_{ti} 's are independent over i and each of them follows a Bernoulli distribution with probability ϕ_t , which is the probability of a tag being abundant under tissue type t . Given the value of $\boldsymbol{\lambda}_t$,

we define $p_{tl}^* = \sum_{i \in A_t} p_{tli}$, which is the total mass of the abundant set in the l^{th} library of type

t . If all unique tags are abundant, then we have $p_{tl}^* = \sum_{i=1}^G p_{tli} = 1$ and there are no scarce genes, i.e., $S_t = \emptyset$. Similarly, we have $p_{tl}^* = 0$ when no unique tags are abundant. In this case,

the abundant set $A_t = \emptyset$. Mathematically, we have $p_{tl}^* = 1$ if $\sum_{i=1}^G \lambda_{ti} = G$, $p_{tl}^* = \sum_{i \in A_t} p_{tli}$ if

$0 < \sum_{i=1}^G \lambda_{ti} < G$, and $p_{tl}^* = 0$ if $\sum_{i=1}^G \lambda_{ti} = 0$. Then we define a column vector $\mathbf{q}_{tl} = (q_{tl1}, \dots, q_{tlG})'$ for library l of type t , where q_{tli} represents the relevant re-scaled conditional probability of observing X_{tli} for a unique tag i under library l of type t . Depending on

whether the unique tag i belongs to an abundant set or a scarce set, $\mathbf{q}_{tl}$ can be partitioned into $\mathbf{q}_{A_{tl}} = \{q_{A_{tli}} = p_{tli} / p_{tl}^* : i \in A_t\}$ or $\mathbf{q}_{S_{tl}} = \{q_{S_{tli}} = p_{tli} / (1 - p_{tl}^*) : i \in S_t\}$. This implies that when a unique tag i belongs to the abundant set, $p_{tli} = q_{A_{tli}} p_{tl}^*$, and when a unique tag i belongs to the scarce set, $p_{tli} = q_{S_{tli}} (1 - p_{tl}^*)$. Note that when all tags are abundant, $\mathbf{q}_{S_{tl}}$'s are not defined, and when all unique tags are scarce, $\mathbf{q}_{A_{tl}}$'s are not defined. Based on the relationships

between p_{tli} , $q_{A_{tli}}$, and $q_{S_{tli}}$, we have $p_{tli} = (q_{A_{tli}} p_{tl}^*)^{\lambda_{ti}} \{q_{S_{tli}} (1 - p_{tl}^*)\}^{(1 - \lambda_{ti})}$. Let

$$\mathbf{p}_t^* = (p_{t1}^*, \dots, p_{tL_t}^*)', \mathbf{p}^* = ((\mathbf{p}_1^*)', \dots, (\mathbf{p}_T^*)')', \mathbf{q}_{At} = (q_{At1}', \dots, q_{AtL_t}')',$$

$$\mathbf{q}_{St} = (q_{St1}', \dots, q_{StL_t}')', \mathbf{q}_A = (q_{A1}', \dots, q_{AT}')', \mathbf{q}_S = (q_{S1}', \dots, q_{ST}')', \text{ and } \boldsymbol{\lambda} = (\lambda_1', \dots, \lambda_T')'.$$

Given $\boldsymbol{\lambda}$ and the observed data D , the conditional likelihood can be rewritten as

$$L(\mathbf{p}^*, \mathbf{q}_A, \mathbf{q}_S | \boldsymbol{\lambda}, D) = \prod_{t=1}^T \prod_{l=1}^{L_t} (X_{tl1}, \dots, X_{tLG}) \prod_{i=1}^G (q_{Atli} p_{tl}^*)^{\lambda_{ti} X_{tli}} \{q_{Stli} (1 - p_{tl}^*)\}^{(1 - \lambda_{ti}) X_{tli}}$$

. If the population is composed of very abundant species, very rare species, and species in between, then we can generalize the above model to a mixture of three Dirichlet distributions, each conditioning on the total mass of a gene in each category.

3. Prior Elicitation

We assume that $\mathbf{q}_{Atl}$ and $\mathbf{q}_{Stl}$ and p_{tl}^* are conditionally independent given λ_t . We note that this assumption holds if a Dirichlet prior is assumed for $\mathbf{p}_{tl}$. For $\mathbf{q}_{Atl}$ and $\mathbf{q}_{Stl}$, we specify the symmetric Dirichlet distributions with $q_{Atl} \sim \mathcal{D}_{k_{At}}(\theta_{Atl})$ and $q_{Stl} \sim \mathcal{D}_{k_{St}}(\theta_{Stl})$, where

$\mathcal{D}_{k_{At}}(\theta_{Atl}) \left(\mathcal{D}_{k_{St}}(\theta_{Stl}) \right)$ denotes a symmetric $k_{At} - 1$ ($k_{St} - 1$) variate Dirichlet distribution with all parameters equal to θ_{Atl} (θ_{Stl}). Based on the definition of p_{tl}^* , we observe that (i) p_{tl}^* is highly dependent on the value of λ_t ; (ii) the value of p_{tl}^* increases when the size of the abundance set k_{At} increases; and (iii) when $k_{At} = G$, $p_{tl}^* = 1$, and when $k_{At} = 0$, $p_{tl}^* = 0$ in probability 1. To account for the dependence of p_{tl}^* on λ_t , we specify the conditional distribution for p_{tl}^* as a mixture distribution composed of a beta distribution depending on k_{At} and two degenerate distributions at 0 (for all genes to be scarce) and 1 (for all genes to be abundant), where the beta distribution is given by

$$\pi(p_{tl}^* | \lambda_t, \alpha_{tl}) = \frac{\Gamma(G \alpha_{tl})}{\Gamma(k_{At} \alpha_{tl}) \Gamma(k_{St} \alpha_{tl})} (p_{tl}^*)^{k_{At} \alpha_{tl} - 1} (1 - p_{tl}^*)^{k_{St} \alpha_{tl} - 1} \text{ for } 0 < \sum_{i=1}^G \lambda_{ti} < G.$$

The hyperparameter α_{tl} controls the degree of concentration of p_{tl}^* around k_{At}/G . The larger it is, the more concentrated the p_{tl}^* is about its mean value (the proportion of abundant genes of type t). This size dependent prior takes advantages of information provided by λ_t as it is not only more natural in defining the total mass of abundant genes, but also reduces the two hyperparameters (mean and variance) specification as in Morris et al. (2003) into only one parameter (the variance). This will streamline our hierarchical construction and interpretation for borrowing information across libraries.

Priors at the Library Level

For hyperparameters α_{tl} , θ_{Atl} and θ_{Stl} , we assume α_{tl} is independent of θ_{Atl} and θ_{Stl} *a priori*.

We specify $\alpha_{tl} \stackrel{iid}{\sim} \mathcal{E}(\alpha_t)$ for $l = 1, \dots, L_t$, where $\mathcal{E}(\alpha_t)$ denotes an exponential distribution with mean $1/\alpha_t$. As the frequencies of the scarce genes have smaller variances than those of the abundant genes, we impose the constraint $\theta_{Atl} < \theta_{Stl}$ to ensure identifiability. To ease the computational burden, we consider an de-constrained transformation: $\theta_{Stl}^* = \theta_{Stl} - \theta_{Atl}$. Then, we assume that θ_{Atl} and θ_{Stl}^* are independent and take $\theta_{Atl} \stackrel{iid}{\sim} \mathcal{E}(\theta_{At})$ and $\theta_{Stl}^* \stackrel{iid}{\sim} \mathcal{E}(\theta_{St}^*)$. Note

that the prior mean of α_{tl} is $E(\alpha_{tl}|\alpha_t) = \frac{1}{\alpha_t}$, which measures the average value of α_{tl} over the L_t libraries of tissue type t . Similarly, $1/\theta_{At}$ and $1/\theta_{At} + 1/\theta_{St}^*$ measure the average values of θ_{Atl} and θ_{Stl} over the L_t libraries of tissue type t .

Priors at the Tissue Type Level

To borrow information across different types of tissues, we take $\alpha_t \stackrel{iid}{\sim} \mathcal{E}(\alpha)$, $\theta_{At} \stackrel{iid}{\sim} \mathcal{E}(\theta_A)$ and $\theta_{St}^* \stackrel{iid}{\sim} \mathcal{E}(\theta_S^*)$ for $t = 1, \dots, T$, where α_t , θ_{At} , and θ_{St}^* are assumed to be independent. We further specify independent gamma priors for α , θ_A , and θ_S^* as $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$, $\theta_A \sim \text{Gamma}(a_A, b_A)$ and $\theta_S^* \sim \text{Gamma}(a_S, b_S)$, where the shape and scale parameters a_α , b_α , a_A , b_A , a_S , and b_S are pre-specified. In Sections 6 and 7, we use $a_\alpha = a_A = a_S = 1$, and $b_\alpha = b_A = b_S = 0.001$ (Section 6) or 0.1 (Section 7). Finally, we specify independent *beta* priors for ϕ_t (the probability of abundant genes for tissue type t) given by

$\phi_t \stackrel{iid}{\sim} \text{beta}\left\{\frac{1-C_0-\phi}{C_0}\phi, \frac{1-C_0-\phi}{C_0}(1-\phi)\right\}$, where $0 < C_0 < 1$ is a pre-specified hyperparameter. We further specify a uniform prior, $\phi \sim U(0, \phi_0)$, for ϕ , where $0 < \phi_0 < 1 - C_0$ is pre-specified. It is easy to show that the prior for ϕ_t has the mean and variance $E(\phi_t|\phi) = \phi$ and $\text{Var}(\phi_t|\phi) = C_0\phi$. Thus, a large value of C_0 reflects a vague prior belief in the prior mean ϕ and a small value yields a strong prior belief in ϕ . In Sections 6 and 7, we consider $(\phi_0, C_0) = (0.3, 0.5)$, $(0.3, 0.1)$, and $(0.6, 0.3)$ for carrying out sensitivity analysis of the posterior estimates of ϕ_t 's with respect to ϕ_0 and C_0 in our simulation studies and real data analysis. Morris et al. (2003) considered only one SAGE experiment with only one tissue type. As in their case there is only a single ϕ , ϕ is not estimable. Thus, they set ϕ to be fixed. As mentioned in Morris et al. (2003), their analysis was very sensitive to the choice of ϕ . On the contrary, as empirically shown in Sections 6 and 7, our prior specification for ϕ_t 's and ϕ results in more robust posterior estimates of ϕ_t 's due to the availability of multiple tissue types in our EST data. In order to promote a better understanding of the hierarchical model, a graphical display of the model structure and its prior specifications for $T = 2$ is constructed in Figure 2.

4. Posterior Distribution and Computational Development

We develop an efficient MCMC sampling algorithm to circumvent the problem of no analytical solution for the posterior distribution due to the complex hierarchical model. First, we integrate out parameters α , θ_A and θ_S^* , and sample from several conditional posterior distributions in a hierarchical fashion. We then use collapsed Gibbs sampler as in Liu (1994) to sample λ , $\alpha^{(1)}$, $\theta_A^{(1)}$, and $\theta_S^{*(1)}$ after collapsing out p^* , q_A and q_S , where $\alpha^{(1)} = (\alpha_{tl}, l = 1, \dots, L_t, t = 1, \dots, T)'$, $\theta_A^{(1)} = (\theta_{Atl}, l = 1, \dots, L_t, t = 1, \dots, T)'$, and $\theta_S^{*(1)} = (\theta_{Stl}^*, l = 1, \dots, L_t, t = 1, \dots, T)'$. We note that this is the key step of the MCMC sampling algorithm as q_A and q_S are extremely high-dimensional. Moreover, instead of sampling α_{tl} , we sample its log value using a localized Metropolis algorithm with a normal

distribution proposal. To sample ϕ_t and ϕ , we use the collapsed Gibbs sampler again by first sampling ϕ given λ and D (data) and then sampling the vector of ϕ_t given (ϕ, λ, D) .

It is worthy of mentioning there are many gamma functions involved in the conditional posterior distributions. Computing these gamma functions is time-consuming and may encounter a potential over-floating problem. However, all gamma functions involved in the conditional posterior distributions can be paired in a form of fraction with integer difference in the arguments. Combining this manipulation with Stirling's formula, we can efficiently sample those variates from their posterior distributions and avoid over-floating problems. In our MCMC algorithm, the generation of λ shares the same feature as in Morris et al. (2003). Otherwise, it is quite different from their algorithm, due to the hierarchical structure developed in this paper. The details for the derivation and computing of the posterior distribution are given in Web Appendix A.

5. Gene Selection Criteria

One of our major objectives is to detect DE genes between two different tissue types. The probability p_{tli} represents the relative abundance of the gene i under library l of tissue type t . To assess whether a gene is DE between two different tissue types, we construct a weighted average of the p_{tli} to obtain the type-specific gene level summary measure as

$p_{wti} = \sum_{l=1}^{L_t} w_{tl} p_{tli}$, where $\sum_{l=1}^{L_t} w_{tl} = 1$ for the tissue type t . Following Baggerly et al. (2003), we choose the weight $w_{tl} \propto n_{tl}$. We declare the gene i to be DE if the difference between the weighted averages p_{wti} and $p_{wt'i}$ for types t and t' is large enough with respect to its standard deviation. Specifically, we compute the posterior probability

$\gamma_i = P_r \left\{ \frac{|p_{wti} - p_{wt'i}|}{\sqrt{\text{Var}(p_{wti} - p_{wt'i} | D)}} \geq 2 | D \right\}$, where $\text{Var}(p_{wti} - p_{wt'i} | D)$ is the posterior variance of $p_{wti} - p_{wt'i}$. Then we declare that a gene i is DE if $\gamma_i > \gamma_0$, where γ_0 is a predetermined cut-off value (for example 0.5, 0.6, or 0.7).

This method called 2-criterion is an extension of the 1-criterion algorithm proposed in Ibrahim et al. (2002), where a gene is declared to be DE when its posterior probability of at least one standardized unit of changes in the weighted averages is larger than γ_0 . Compared to the 1-criterion, our new criterion is better calibrated and has a better interpretation. Assuming $T = 2$, and $p_{w1i} - p_{w2i}$ is normally distributed, after some algebra, we can show that the 2-criterion ensures the posterior probability, $p_{0i} = \max\{Pr(p_{w1i} - p_{w2i} > 0 | D), Pr(p_{w1i} - p_{w2i} < 0 | D)\}$, to be at least $\Phi[2 - \Phi^{-1}\{1 + \Phi(-2) - \gamma_0\}]$ asymptotically for any $\gamma_0 > 0$, where Φ denotes the standard normal cumulative distribution function. Furthermore, we can also show that asymptotically the 2-criterion ensures that $p_{0i} > 97.7\%$ when $\gamma_0 = 0.5$ and

$p_{0i} > 99.4\%$ when $\gamma_0 = 0.7$. Let $\sigma = \sqrt{\text{Var}(p_{w1i} - p_{w2i} | D)}$, which is the posterior standard deviation of $p_{w1i} - p_{w2i}$. Let us also observe that when $\gamma_0 = 0.5$, the “center” of the posterior distribution of $p_{w1i} - p_{w2i}$ is approximately 2.0σ away from 0; when $\gamma_0 = 0.7$, the “center” of the posterior distribution of $p_{w1i} - p_{w2i}$ is approximately 2.5σ away from 0. These results imply that the 2-criterion ensures that a gene is declared to be DE if the difference between p_{w1i} and p_{w2i} is larger than approximately two standard deviations when $\gamma_0 = 0.5$ and two

and half standard deviations when $\gamma_0 = 0.7$. Finally, we note that fewer DE genes are selected if a larger value of γ_0 is chosen.

6. Simulation Studies

We conducted two simulation studies each with 100 data sets simulated from the NLMD model and a real NCBI dataset, respectively. We evaluated the performance of the proposed method (called two-class) in comparison to three other methods: 2×2 table with a χ^2 test (Romualdi et al., 2001), a multinomial Dirichlet (called one-class) model and the multinomial NLMD (called two-class (non-hier)) model (Morris et al., 2003, 2006). Under the χ^2 test, a 2×2 cross classification table is constructed for each gene where the rows correspond to gene i and “all other genes” and the columns correspond to two different tissue types. A gene is declared to be DE if the resulting p-value for the homogeneous test is less than a pre-specified cut-off value, say 0.01 or 0.05. Under the one-class model that does not account for scarce and abundant genes, we fit the expression relative probabilities $\mathbf{p}_{tl} = \{p_{tli}, i = 1, \dots, G\}$ with a symmetric Dirichlet hierarchical prior distribution $\mathcal{D}_{G-1}(\theta_{tl})$ for each t and l ; $\theta_{tl} | \theta_t \text{ iid } \mathcal{E}(\theta_t), \forall l$ for each fixed t ; and $\theta_t \text{ iid } \mathcal{E}(\theta), \forall t$. Under the two-class (non-hier) model, we modified the prior for the total mass of abundant genes p_{tl}^* to be size dependent as defined in Section 3. As in Morris et al. (2003), we prespecify the values of the hyperparameters $\{\phi_t, \alpha_{tl}, \theta_{Atl}, \theta_{Stl}\}$. Specifically, we set $\phi_t = 0.2$ in Simulation I and 0.05 in Simulation II, $\alpha_{tl} = 0.001$, $\theta_{Atl} = 0.5$, and $\theta_{Stl} = 1.0, \forall t$ and l . Note that the two-class (non-hier) model fits the same NLMD model for each library independently without any hierarchical structures. For the two-class method, we considered (0.3, 0.5), (0.3, 0.1) and (0.6, 0.3) for (ϕ_0, C_0) . All one-class, two-class (non-hier) and two-class model use the 2-criterion based on 5000 MCMC samples with $\gamma_0 = 0.5, 0.6, 0.7$ to detect DE genes.

To evaluate the performance of these four methods, we use four error rates: false negative (FNR), false positive (FPR), false discovery rate (FDR), and false non-discovery rate (FNDR). FNR is the proportion of DE genes failed to be detected as DE. The FPR is the proportion of EE genes wrongly declared to be DE. The FDR is the realized rate of false detections in the detected genes and FNDR is the realized rate of false non-detections in the non-detected genes. A method is considered to be better than the others, if all error rates based on it are smaller than those of the others. In addition to these error rates, we also report the number of genes claimed to be DE (CDE), the number of genes correctly claimed as DE (CCDE), the number of genes correctly claimed as EE (CCEE) averaged over 100 data sets with their standard deviation (in parentheses in Table 1).

6.1 Simulation Study I

Data were generated for 6 libraries with equal library size for each tissue type for each data set. The observed counts under each library were generated from a multinomial distribution $\mathcal{M}(n_{tl}, (p_{tl1}, \dots, p_{tlG}))$ with $G = n_{tl} = 5000$. The parameters in the multinomial distribution $(p_{tl1}, \dots, p_{tlG})$ were generated as follows. We first set $A_1 = \{1, \dots, 1000\}$, $S_1 = \{1001, \dots, 5000\}$, $A_2 = \{1, \dots, 900, 1001, \dots, 1100\}$, $S_2 = \{901, \dots, 1000, 1101, \dots, 5000\}$, and $p_{tl}^* = 0.9$. Then, we generated parameters $(q_{tli}, i \in A_t) \sim \text{symmetric Dirichlet } \mathcal{D}(11)$ and $(q_{tli}, i \in S_t) \sim \mathcal{D}(15)$ under library l and tissue type t . Finally, we set $p_{tli} = q_{tli} p_{tl}^*$ if $i \in A_t$ and

$p_{tli} = q_{tli}(1 - p_{tli}^*)$ if $i \in S_t$. This simulation setting implies that genes 901 to 1100 are DE while all other genes are EE. Specifically, the first (second) half of the 200 DE genes are abundant (scarce) under type 1 and scarce (abundant) under type 2.

The results under Simulation I are summarized in Table 1. From the top panel of Table 1, we see that the two-class method clearly outperforms the χ^2 test, the one-class method and the two-class (non-hier) method based on FNR, FPR, FDR and FNDR. The one-class method and the χ^2 test are comparable in terms of magnitudes of error rates. The error rates from one-class are bounded between that obtained from the χ^2 test with cutoff values of 0.05 and 0.01. The number of DE genes selected by the two-class (non-hier) method tends to fall between the one-class and the two-class methods. Compared to the one-class method, the two-class (non-hier) method performed better by correctly selecting more genes, hence providing smaller error rates for all four. Compared to the two-class method, the two-class (non-hier) method does not perform well for its high value of FDR. The top panel of Table 1 shows that the two-class method is least sensitive to the different choices of cut-off value γ_0 . Looking at the column of CDE from χ^2 down to two-class models, the number of DE genes selected by each method with different cutoff values averaged over 100 simulated data sets decreases also with smaller variances as well. Looking at the error rates, it also demonstrates that the proposed gene selection 2-criterion works well as it leads to satisfactorily low FNR and FPR for all the latter three methods.

We further compared the results obtained from the same methods with different library sizes. The results with $G = 5000$ and $n_{tl} = 2500$ are shown in Table 4 of Web Appendix B. As the library size increases, we have observed all four error rates (FNR, FPR, FDR and FNDR) of the two-class method decreases. For other methods, improvements were only seen on some of the error rates. In particular, the FPR and FDR obtained from the χ^2 test and the one-class method increase when the library size increases; while only the FDR increases for the two-class (non-hier) method.

6.2 Simulation Study II

In this study we simulated more realistic data. We started with a real EST data set with 2 brain tumor tissue libraries (type 1) and 2 normal tissue libraries (type 2). We first obtained the MLEs of the relative probabilities $\hat{p}_{tg}$ for all genes at the tissue type t . Let

$\hat{p}_g = (\hat{p}_{1g} + \hat{p}_{2g})/2$. Then, the top $G = 5000$ genes with the largest $\hat{p}_g$ were used in the study, given genes with small frequency counts for both types are not interesting in detecting DE genes. Among them, 500 genes with largest differences in the MLEs from the two tissue types were set to be DE. Other genes are EE. We applied a logit transformation to the original relative frequency count (probability) for each gene in a library, and computed their means and variances for each gene over the libraries. If a gene is EE, we simulated normal variates for each library with the same mean (averaged over two types) and pooled variance for each gene. If a gene is DE, we simulated normal variates separately for each library to match individual means and variances for each type. Then we applied an inverse logit transformation to take the normal variates back to the original scale. A re-scaling within each library was applied to ensure that their sum is one. The counts for each library were generated independently from a multinomial distribution with library size $n_{tl} = 5000$ and

these relative probabilities for a total of 9 libraries (6 were assigned to type 1 and 3 to type 2), to mimic the real data set. One hundred datasets were simulated.

The results under Simulation II are also summarized in Table 1. From the bottom half of Table 1, we see that the two-class method performs better than the one-class method with smaller or comparable error rates. The χ^2 test with cutoff for p-value set to be 0.05 provides the lowest FNR of 0.275 among all methods, but with a huge FDR of 0.505. The χ^2 test with a different cutoff value of 0.01 provides a much smaller FDR of 0.176, which is still much higher than the other methods. The results also show that the χ^2 test method with the cutoff value of 0.01 perform worse than the two-class method and the one-class method for $\gamma_0 = 0.5$ or 0.6 for its high values for all four error rates. The two-class (non-hier) method provides small error rates for FPR and FDR, but its FPR is much larger with a range of 0.470 to 0.547 compared to those from the two-class method (0.318 to 0.368), the one-class method (0.351 to 0.501), and the χ^2 test (0.275 to 0.445). In addition, the two-class (non-hier) method has the largest FNDR's among all methods. Overall, the two-class method performs better than all other three methods as it successfully identifies a big proportion of DE genes yet with a good control of FDR to be lower than 0.05.

From the two-class model, we have obtained the posterior means for ϕ_1 and ϕ_2 range from 4.99% to 5.71%, and 5.13% to 6.08%, respectively. Both of them are slightly above the prespecified value of 5% for ϕ_t under the two-class (non-hier) method. We also checked the posterior mean for θ_{Atl} and θ_{Stl} for all libraries under both tissue types. Under the first (second) tissue type, the posterior means for the θ_{Atl} range from 0.985 to 1.685 (0.62 to 0.85) and for the θ_{Stl} range from 3.43 to 10.99 (3.85 to 20.88). The differences between the true value of the hyperparameters and the setting specified in the two-class (non-hier) method might explain the poor performance of the two-class (non-hier) method. Similar to the results of Simulation I, we do not observe any significant changes among the results under the three choices of (θ_0, C_0) .

7. Real Data Analysis

An in-house Perl program was used to evaluate 291 most reliable libraries from the NCBI unigene database (<http://www.ncbi.nlm.nih.gov/unigene> and <http://www.ncbi.nlm.nih.gov/dbEST/>). The program outputs EST sequences, their corresponding gene name, UniGene cluster ID, the cDNA libraries they were derived from, as well as the size of the cDNA libraries and their numbers of detected sequences in each library. The samples were extracted from a variety of tissues including lung, muscle, bone, and liver. We focus only on the samples from lymphoreticular or lymph node tissue. We are primarily interested in identifying the genes (unique tags) that are DE between the normal tissue and tumor cell line. There were in total 18,324 unique tags recorded under 6 cDNA libraries of the normal tissue and 3 cDNA libraries of the tumor cell line type. After removing those genes with zero counts under both tissue types, there were 8190 unique tags with at least one non-zero count under one or more cDNA libraries, which forms the subset of the data used in our analysis.

The same four methods were used to analyze the real data with settings similar to the simulation study II. The results are summarized in Table 2. The χ^2 test with a cut-off value of 0.05 selects 2192 DE genes, the largest number of genes being selected among the four methods. The other three methods in order select 1322, 1013, and 1059 DE genes with $\gamma_0 = 0.5$. The last column in Table 2 reports the proportion of genes that are also selected by the χ^2 test (with cutoff value of 0.05) relative to the DE genes selected by the listed methods. They are all essentially above 90%, which is not surprising given the large number of genes selected by the χ^2 test. The two-class methods have slightly larger overlap with the χ^2 test than the two-class (non-hier) method. The results among the two-class methods with the three choices of (φ_0, C_0) are similar, which further confirms that the proposed method is not too sensitive to the choice of (φ_0, C_0) . To evaluate the performance of the proposed method, we applied a permutation test (modified from Storey and Tibshirani, 2003) proposed by Jiao and Zhang (2008) with 1000 permutations to estimate the false discovery rate. We obtained an estimate of 0.10–0.12 for the false discovery rate with γ_0 chosen between 0.5 to 0.7 in the 2-criterion algorithm.

We further examined the sensitivity of the posterior estimates to the choice of (φ_0, C_0) . Specifically, we compared the posterior means, the posterior standard deviations (SDs), and the 95% HPD intervals of all parameters under the different choices of (φ_0, C_0) . The results with $(\varphi_0, C_0) = (0.3, 0.5)$ and $(0.3, 0.1)$ are summarized in Table 3. We see that these two sets of the posterior estimates for all parameters except for φ are very similar. Note that the decision on selecting DE genes is primarily based upon the parameters at the library and tissue levels including φ_t , α_{tl} , θ_{Atl} and θ_{Stl} for all t and l . The results with $(\varphi_0, C_0) = (0.6, 0.3)$ were not shown in Table 3. However, the posterior estimates are similar to those with the other two choices of (φ_0, C_0) . For example, the posterior means, SDs, and 95% HPD intervals under $(\varphi_0, C_0) = (0.6, 0.3)$ are 0.1987, 0.0126, and (0.1745, 0.2238) for φ_1 ; and 0.1387, 0.0008, and (0.1235, 0.1535) for φ_2 ; Therefore, the proposed method is fairly robust to the choice of (φ_0, C_0) in terms of posterior estimates.

Finally, we note that forty-nine DE genes were selected by the two-class method with $C_0 = 0.5$ but not by the χ^2 test or the one-class methods. Sixteen of them were also not selected by the two-class (non-hier) method. Among them, many are important and of great biological interest. For example, cyclin-dependent kinase 4 (CDK4, Hs.95577) was found to be up-regulated in tumor cell lines but not in normal lymph node. CDK4 is a catalytic subunit of the protein kinase complex and plays an important role in cell cycle G1 phase progression. Mutations in this gene as well as in its related proteins were found to be associated with tumorigenesis of a variety of cancers (Molenaar et al., 2008). Another gene found to be up-regulated is heat shock 90kDa protein 1, alpha (HSPCA, Hs.525600), which was shown to be overexpressed in poor-prognosis acute myeloid leukemia cells and plays a role in cell survival and resistance to chemotherapy (Flandrin et al., 2008). Other genes found to be up-regulated code for proteins are involved in pre-mRNA splicing, protein translation, and cellular metabolism. Among the downregulated genes, there are several known tumor suppressors: the forkhead box O transcription factor (FOXO1, Hs.370666), which functions as a tumor suppressor by regulating expression of genes involved in apoptosis, cell cycle arrest and oxidative detoxification (Liu et al., 2008); BRCA1 (Hs. 194143), which is known

to suppress tumor growth; and runt-related transcription factor 3 (RUNX3, Hs.170019), which functions as a tumor suppressor and is frequently deleted or transcriptionally silenced in cancer cells (Kim et al., 2008).

8. Discussion

In this paper, we have developed a hierarchical multinomial NLMD model for robust estimation of the EST data by borrowing information from multiple libraries and multiple tissue types. Due to the availability of multiple libraries and multiple tissue types, the hierarchical modeling allows the parameters at the library level and the tissue level (the second and third levels in the hierarchy) to be estimable and identifiable, which yields the observed advantages as seen in the simulation studies in Section 6. In the proposed model, the hierarchical structure is built on the parameters of the NLMD distributions across different libraries. A natural alternative is to build a hierarchical model directly on the overall distributions of the p_{tli} 's by assuming

$\mathbf{p}_{tli} \sim \mathcal{D}_G \left\{ \frac{1 - \xi_t}{\xi_t} p_{t1}, \dots, \frac{1 - \xi_t}{\xi_t} p_{t,G-1}, \frac{1 - \xi_t}{\xi_t} \left(1 - \sum_{i=1}^{G-1} p_{ti} \right) \right\}$ and an NLMD prior for $(p_{t1}, \dots, p_{tG})'$, where p_{ti} is the tissue and gene specific probability and $\xi_t > 0$. Under this formulation, the prior mean $E(p_{tli}|p_{ti}) = p_{ti}$, ξ_t controls the degree of concentration of p_{tli} around p_{ti} , and posterior inference can be carried out directly on the p_{ti} rather than the average of the p_{tli} 's across l . However, this alternative approach greatly increases the computational difficulty and prevents carrying out posterior sampling. In addition, since p_{ti} equals the mean of the p_{tli} 's, a simple average of p_{tli} 's across l may serve as a reasonable approximation of p_{ti} and, hence, posterior inference on the average of the p_{tli} 's across l for detecting DE genes may be similar to the one based on the p_{ti} . Finally, to detect DE genes between two tissue types, we have proposed a new gene selection criterion, called the 2-criterion. This gene selection criterion is easy to implement and also has a nice statistical interpretation. Although the current version of the gene selection algorithm is written for comparing two types of tissues, it can be extended to three or more tissue types. The further theoretical properties of the 2-criterion and its extension are currently under investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the editor, the associate editor, and the two referees for their helpful comments and suggestions, which have led to a considerable improvement of this article. Dr. Chen's research was partially supported by NIH grants #GM 70335 and #CA 74015 and Dr. Kuo's research was partially supported by NIH grant #GM 5764-01.

References

- Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC. Sequence identification of 2,375 human brain genes. *Nature*. 1992; 355:632–634. [PubMed: 1538749]
- Audic S, Claverie JM. The significance of digital gene expression profiles. *Genome Research*. 1997; 7:986–995. [PubMed: 9331369]

- Baggerly KA, Deng L, Morris JS, Aldaz CM. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*. 2003; 19(12):1477–1483. [PubMed: 12912827]
- Baggerly KA, Deng L, Morris JS, Aldaz CM. Overdispersed logistic regression for SAGE: Modelling multiple groups and covariates. *BMC Bioinformatics*. 2004; 5:144–159. [PubMed: 15469612]
- Claverie JM. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics*. 1999; 8(10):1821–1832. [PubMed: 10469833]
- Flandrin P, Guyotat D, Duval A, Cornillon J, Tavernier E, Nadal N, Campos L. Significance of heat-shock protein (HSP) 90 expression in acute myeloid leukemia cells. *Cell Stress Chaperones*. 2008; 13(3):357–364. [PubMed: 18386162]
- Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*. 2002; 97:88–99.
- Jiao S, Zhang S. On correcting the overestimation of the permutation-based false discovery rate estimator. *Bioinformatics*. 2008; 24(15):1655–1661. [PubMed: 18573796]
- Kim EJ, Kim YJ, Jeong P, Ha YS, Bae SC, Kim WJ. Methylation of the RUNX3 promoter as a potential prognostic marker for bladder tumor. *Journal of Urology*. 2008; 180(3):1141–1145. [PubMed: 18639281]
- Kuznetsov VA. Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP Journal on Applied Signal Processing*. 2001; 4:285–296.
- Liu JS. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*. 1994; 89:958–966.
- Liu P, Kao TP, Huang H. CDK1 promotes cell proliferation and survival via phosphorylation and inhibition of FOXO1 transcription factor. *Oncogene*. 2008; 27(34):4733–4744. [PubMed: 18408765]
- Lu J, Tomfohr JK, Kepler TB. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*. 2005; 6:165–178. [PubMed: 15987513]
- Molenaar JJ, Ebus ME, Koster J, van Sluis P, van Noesel CJ, Versteeg R, Caron HN. Cyclin D1 and CDK4 activity contribute to the undifferentiated phenotype in neuroblastoma. *Cancer Research*. 2008; 68(8):2599–2609. [PubMed: 18413728]
- Morris JS, Baggerly KA, Coombes KR. Bayesian shrinkage estimation of the relative abundance of mRNA transcripts using SAGE. *Biometrics*. 2003; 59:476–486. [PubMed: 14601748]
- Morris, JS.; Baggerly, KA.; Coombes, KR. Shrinkage estimation for SAGE data using a mixture Dirichlet prior. In: Do, KA.; Müller, P.; Vannucci, M., editors. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press; New York: 2006. p. 254-268.
- Romualdi C, Bortoluzzi S, Danieli GA. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. *Human Molecular Genetics*. 2001; 10(19):2133–2141. [PubMed: 11590130]
- Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B, Rosenthal A. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumor tissues. *Nucleic Acids Research*. 1999; 27(21):4251–4260. [PubMed: 10518618]
- Stekel DJ, Git Y, Falciani F. The comparison of gene expression from multiple cDNA libraries. *Genome Research*. 2000; 10:2055–2061. [PubMed: 11116099]
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*. 2003; 100:9440–9445. [PubMed: 12883005]
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995; 270:484–487. [PubMed: 7570003]

Figure 1.
Frequency of observed counts

Figure 2.

Graphical Display of the Hierarchical Structure. The structure on the left (right) is for libraries in tissue type 1 (2). Multiple libraries are stacked up with only l^{th} library in each tissue type plotted. The information sharing across libraries is done by second rows top and bottom, and information sharing across tissues is done by first row top and bottom.

Table 1

Method comparison based on two simulation studies (G =5000)

Simulation	Method	CutOff	CDE	CCDE	CCEE	FNR	FPR	FDR	FNDR
I (#DE=200)	χ^2	0.05	313.2(8.0)	199.9(0.3)	4686.7(7.9)	0.001	0.024	0.361	0.000
		0.01	224.8(5.0)	199.8(0.4)	4775.0(5.0)	0.001	0.005	0.111	0.005
	one-class	0.5	271.2(6.4)	199.9(0.3)	4728.7(6.3)	0.001	0.015	0.263	0.000
		0.6	243.8(6.2)	199.9(0.3)	4756.1(6.0)	0.001	0.009	0.180	0.000
		0.7	225.0(5.2)	199.8(0.4)	4774.8(5.3)	0.001	0.005	0.112	0.000
	two-class (non-hier)	0.5	265.9(8.5)	200.0(0.0)	4734.1(8.6)	0.000	0.014	0.247	0.000
		0.6	239.6(6.1)	200.0(0.1)	4760.4(6.2)	0.000	0.008	0.165	0.000
		0.7	221.6(4.8)	200.0(0.2)	4778.4(4.8)	0.000	0.005	0.097	0.000
	two-class* (0.3, 0.5)	0.5	201.8(1.6)	200.0(0.0)	4798.2(1.4)	0.000	0.000	0.009	0.000
		0.6	200.5(0.7)	200.0(0.0)	4799.5(0.0)	0.000	0.000	0.003	0.000
		0.7	200.3(0.5)	200.0(0.0)	4799.7(0.0)	0.000	0.000	0.002	0.000
	two-class (0.3, 0.1)	0.5	201.7(1.6)	200.0(0.0)	4798.3(1.4)	0.000	0.000	0.008	0.000
		0.6	200.5(0.7)	200.0(0.0)	4799.5(0.0)	0.000	0.000	0.003	0.000
		0.7	200.3(0.5)	200.0(0.0)	4799.7(0.0)	0.000	0.000	0.002	0.000
	two-class (0.6, 0.3)	0.5	201.4(1.3)	200.0(0.1)	4798.6(1.2)	0.000	0.000	0.007	0.000
		0.6	200.3(0.6)	200.0(0.1)	4799.7(0.6)	0.000	0.000	0.002	0.000
		0.7	200.1(0.3)	200.0(0.1)	4799.9(0.3)	0.000	0.000	0.001	0.000
II (#DE=500)	χ^2	0.05	732.8(20.7)	362.4(8.8)	4129.6(17.6)	0.275	0.082	0.505	0.032
		0.01	336.6(10.4)	277.4(8.6)	4440.8(7.1)	0.445	0.013	0.176	0.048
	one-class	0.5	364.2(11.5)	324.4(9.1)	4460.2(7.0)	0.351	0.009	0.109	0.038
		0.6	303.7(9.6)	287.5(8.6)	4483.8(4.0)	0.425	0.004	0.053	0.045
		0.7	253.7(8.8)	249.4(8.5)	4495.7(2.1)	0.501	0.001	0.017	0.053
	two-class (non-hier)	0.5	267.1(8.6)	264.8(8.5)	4497.7(1.6)	0.470	0.000	0.009	0.050
		0.6	246.6(8.5)	245.1(8.3)	4498.5(1.2)	0.510	0.000	0.006	0.054
		0.7	227.6(8.1)	226.6(8.0)	4499.0(1.0)	0.547	0.000	0.004	0.057
	two-class (0.3, 0.5)	0.5	353.3(10.3)	341.2(8.9)	4487.9(3.7)	0.318	0.003	0.034	0.034
		0.6	337.6(9.0)	328.3(8.2)	4490.7(3.2)	0.343	0.002	0.028	0.037
		0.7	323.0(8.4)	315.9(7.9)	4492.9(2.9)	0.368	0.002	0.022	0.039
	two-class (0.3, 0.1)	0.5	353.6(10.3)	341.4(9.0)	4487.8(3.7)	0.317	0.003	0.034	0.034
		0.6	337.8(8.8)	328.4(8.0)	4490.6(3.2)	0.343	0.002	0.028	0.037
		0.7	322.8(8.4)	315.7(7.9)	4492.9(3.0)	0.369	0.002	0.022	0.039
	two-class (0.6, 0.3)	0.5	353.4(10.1)	341.3(9.0)	4487.9(3.2)	0.317	0.003	0.034	0.034
		0.6	337.7(8.8)	328.3(8.1)	4490.7(2.8)	0.343	0.002	0.028	0.037
		0.7	322.9(8.4)	315.8(8.0)	4492.9(2.6)	0.368	0.002	0.022	0.039

Note: Under the method column, “two-class (non-hier)” denotes the two-class method without hierarchical structure with $\phi_1 = \phi_2 = 0.2$ under stimulation study I and $\phi_1 = \phi_2 = 0.05$ under stimulation study II.

* The two numeric values inside the parentheses under “two-class” correspond to the values of ϕ_0 and C_0 respectively.

Table 2

Real data analysis based on samples from lymphoreticular or lymph node tissues ($G = 8190$).

Method (Cut-off)	CDE	DE shared by χ^2 (0.05)	% DE shared by χ^2 (0.05)
χ^2 (0.05)	2192	2192	100.0
(0.01)	1073	1073	100.0
one-class (0.5)	1322	1213	91.8
(0.6)	1056	1035	98.0
(0.7)	828	825	99.6
two-class (0.5)	1013	909	89.7
(non-hier) (0.6)	931	854	91.7
(0.7)	837	786	93.9
two-class* (0.5)	1059	996	94.1
(0.3, 0.5) (0.6)	897	859	95.8
(0.7)	772	751	97.3
two-class (0.5)	1059	994	93.9
(0.3, 0.1) (0.6)	900	861	95.7
(0.7)	773	753	97.4
two-class (0.5)	1059	997	94.2
(0.6, 0.3) (0.6)	895	858	95.9
(0.7)	778	757	97.3

Note: Under the method column, “two-class (non-hier)” denotes the two-class method without hierarchical structure with setting of $\phi_1 = \phi_2 = 0.05$.

* The two numeric values within the parentheses under “two-class” correspond to the values of ϕ_0 and C_0 , respectively.

Table 3

Posterior estimates of the parameters from real data analysis.

	$C_0 = 0.5$				$C_0 = 0.1$		
Parameter	Mean	SD	95% HPD Interval	Mean	SD	95% HPD Interval	
φ	0.1432	0.0585	(0.0536, 0.2630)	0.2305	0.0548	(0.1223, 0.2999)	
φ_1	0.1991	0.0116	(0.1782, 0.2235)	0.1977	0.0115	(0.1748, 0.2199)	
φ_2	0.1387	0.0078	(0.1236, 0.1537)	0.1386	0.0078	(0.1248, 0.1549)	
α_1	0.0003	0.0002	(0.0000, 0.0006)	0.0003	0.0002	(0.0000, 0.0006)	
θ_{A1}	0.2826	0.1485	(0.1055, 0.5689)	0.2829	0.1390	(0.0948, 0.5614)	
θ_{S1}^*	0.1158	0.0732	(0.0260, 0.2541)	0.1162	0.0722	(0.0276, 0.2564)	
α_2	0.0003	0.0003	(0.0000, 0.0008)	0.0003	0.0003	(0.0000, 0.0007)	
θ_{A2}	0.4961	0.3580	(0.1028, 1.1493)	0.5049	0.3793	(0.1158, 1.1894)	
θ_{S2}^*	0.5760	0.5002	(0.1350, 1.3476)	0.5587	0.4401	(0.1226, 1.2384)	
α_{11}	0.0003	0.0002	(0.0000, 0.0007)	0.0003	0.0003	(0.0000, 0.0008)	
θ_{A11}	0.0963	0.0083	(0.0804, 0.1129)	0.0958	0.0085	(0.0801, 0.1133)	
θ_{S11}^*	0.1445	0.0200	(0.1073, 0.1849)	0.1446	0.0208	(0.1037, 0.1849)	
α_{12}	0.0003	0.0002	(0.0000, 0.0007)	0.0003	0.0002	(0.0000, 0.0007)	
θ_{A12}	0.0497	0.0047	(0.0412, 0.0594)	0.0497	0.0049	(0.0410, 0.0599)	
θ_{S12}^*	0.1081	0.0128	(0.0842, 0.1342)	0.1072	0.0131	(0.0802, 0.1316)	
α_{13}	0.0003	0.0002	(0.0000, 0.0008)	0.0003	0.0002	(0.0000, 0.0007)	
θ_{A13}	0.1636	0.0175	(0.1328, 0.1997)	0.1645	0.0182	(0.1295, 0.1995)	
θ_{S13}^*	0.0123	0.0124	(0.0005, 0.0364)	0.0115	0.0115	(0.0002, 0.0341)	
α_{14}	0.0002	0.0002	(0.0000, 0.0006)	0.0002	0.0002	(0.0000, 0.0006)	
θ_{A14}	0.4373	0.0257	(0.3870, 0.4873)	0.4386	0.0246	(0.3924, 0.4899)	
θ_{S14}^*	0.0603	0.0366	(0.0019, 0.1243)	0.0589	0.0340	(0.0018, 0.1218)	
α_{15}	0.0001	0.0001	(0.0000, 0.0002)	0.0001	0.0001	(0.0000, 0.0002)	
θ_{A15}	0.3180	0.0343	(0.2496, 0.3814)	0.3210	0.0325	(0.2595, 0.3859)	
θ_{S15}^*	0.1027	0.1492	(0.0017, 0.3583)	0.1028	0.1403	(0.0013, 0.3512)	
α_{16}	0.0001	0.0001	(0.0000, 0.0003)	0.0001	0.0001	(0.0000, 0.0004)	
θ_{A16}	0.1559	0.0217	(0.1140, 0.1983)	0.1554	0.0215	(0.1123, 0.1945)	
θ_{S16}^*	0.0072	0.0097	(0.0003, 0.0243)	0.0075	0.0095	(0.0001, 0.0233)	
α_{21}	0.0002	0.0001	(0.0000, 0.0004)	0.0002	0.0001	(0.0000, 0.0004)	
θ_{A21}	0.2559	0.0119	(0.2335, 0.2797)	0.2562	0.0121	(0.2335, 0.2798)	

	$C_0 = 0.5$				$C_0 = 0.1$		
Parameter	Mean	SD	95% HPD Interval	Mean	SD	95% HPD Interval	
θ_{S21}^*	0.0014	0.0014	(0.0001, 0.0041)	0.0014	0.0013	(0.0000, 0.0042)	
α_{22}	0.0002	0.0002	(0.0000, 0.0005)	0.0002	0.0002	(0.0000, 0.0006)	
θ_{A22}	0.4683	0.0260	(0.4201, 0.5205)	0.4686	0.0259	(0.4181, 0.5215)	
θ_{S22}^*	1.3839	0.1537	(1.0777, 1.6656)	1.3822	0.1519	(1.0934, 1.6765)	
α_{23}	0.0002	0.0001	(0.0000, 0.0004)	0.0002	0.0001	(0.0000, 0.0004)	
θ_{A23}	0.3005	0.0126	(0.2759, 0.3257)	0.3012	0.0131	(0.2767, 0.3265)	
θ_{S23}^*	0.0017	0.0017	(0.0001, 0.0053)	0.0018	0.0018	(0.0000, 0.0055)	