Multiple Testing for Pattern Identification, With Applications to Microarray Time-Course Experiments

Wenguang SUN and Zhi WEI

In time-course experiments, it is often desirable to identify genes that exhibit a specific pattern of differential expression over time and thus gain insights into the mechanisms of the underlying biological processes. Two challenging issues in the pattern identification problem are: (i) how to combine the simultaneous inferences across multiple time points and (ii) how to control the multiplicity while accounting for the strong dependence. We formulate a compound decision-theoretic framework for set-wise multiple testing and propose a data-driven procedure that aims to minimize the missed set rate subject to a constraint on the false set rate. The hidden Markov model proposed in Yuan and Kendziorski (2006) is generalized to capture the temporal correlation in the gene expression data. Both theoretical and numerical results are presented to show that our data-driven procedure controls the multiplicity, provides an optimal way of combining simultaneous inferences across multiple time points, and greatly improves the conventional combined p-value methods. In particular, we demonstrate our method in an application to a study of systemic inflammation in humans for detecting early and late response genes.

KEY WORDS: Compound decision problem; Conjunction and partial conjunction tests; False discovery rate; Hidden Markov models; Microarray time-course data; Simultaneous set-wise testing.

1. INTRODUCTION

Microarray time-course (MTC) experiments are capable of capturing the dynamic changes of genes over time and have been widely applied for studying many biological processes, such as the regulation of development (Arbeitman et al. 2002), immune response (Calvano et al. 2005), and tissue inflammation program (Tian, Nowak, and Brasier 2005). The MTC experiments are usually conducted under one or multiple biological conditions. Of interest for one-condition experiment is to identify genes whose expression levels change over time in some specific way. A well-known example is the MTC experiment conducted by Spellman et al. (1998) for studying cell cycles. Recently two-condition MTC experiments have become popular. Like in common case-control studies, genes exhibiting different temporal patterns across two biological conditions are good candidates for further study because the biological process motivating the experiments may be driven by their differential expressions. Multiple-condition MTC experiments are rare due to the complication in design and analysis. In this article we focus on the problems under two biological conditions.

In two-condition MTC experiments, each gene at each time point has two possible states: equally expressed (EE) or differentially expressed (DE). A temporal pattern is a prespecified class of sequences of DE and EE states over time. The temporal patterns of dynamic changes often provide insights into the underlying biological mechanisms—genes that exhibit specific temporal patterns of DE can be informative in deciphering the underlying regulatory programs that govern the dynamic biological process of interest.

We first consider a motivating example. In studying the biological process of how the cytokine tumor necrosis factor (TNF) initiates tissue inflammation, Tian, Nowak, and Brasier (2005) conducted a time-course microarray experiment to profile gene activities before the inhibition of the NF-kB transcription factor, a mediator of the process, and at one, three, and six hours after the inhibition. The investigators were interested in identifying genes with the following distinct temporal patterns:

(1) "Early response genes" that were differentially expressed (DE) less than one hour after the NF-kB inhibition
(2) "Middle response genes" that were DE at three hours but no response prior to three hours
(3) "Late response genes" that were DE at six hours but no response prior to six hours
(4) "Biphasic genes" that were DE at both one hour and six hours, but not at three hours.

Such delicate expression patterns can be used to distinguish upstream regulatory genes from downstream regulated genes and help to form biological hypotheses for further experimental validations. The goal of this article is to develop powerful multiple testing procedures to identify genes that exhibit temporal patterns of interest.

Conventional gene selection procedures for analysis of time-course data were only developed for identifying genes that are DE at *one single time point*, and cannot be used for pattern identification. Recently a few approaches were proposed to select genes that show *overall* difference in their expression profiles. Some approaches view the time-course gene expression data as vectors of correlated observations. Under this formulation, statistical methods in multivariate analysis and longitudinal data analysis were proposed to test the equivalence of two vector values; such methods include the F-statistic derived from an ANOVA analysis (Park et al. 2003; Ma, Zhong, and Liu 2009), the robust Wald statistic derived from a generalized estimating equation analysis (Guo et al. 2003), the moderated likelihood ratio statistic and Hotelling T^2-statistic derived from a multivariate empirical Bayes model (Tai and Speed 2006), and the maximum ratio statistic derived from a hierarchical Bayes

Wenguang Sun is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27606-8203 (E-mail: *sun@stat.ncsu.edu*). Zhi Wei is Assistant Professor, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102. This work was supported in part by National Science Foundation grant DMS-10-07675. We thank the Associate Editor and two referees for detailed and constructive comments which lead to a much improved article.

Journal of the American Statistical Association
March 2011, Vol. 106, No. 493, Applications and Case Studies
DOI: 10.1198/jasa.2011.ap09587

model (Chi et al. 2007). Alternately, the time-course data can be viewed as samples from an underlying continuous gene expression trajectory. Under this formulation, statistical methods in functional data analysis were developed to test the equivalence of two expression curves; such methods include the hierarchical modeling and basis expansion approaches considered by Luan and Li (2004), Storey et al. (2005), Hong and Li (2006), and Telesca et al. (2009). However, it is not clear how to combine the simultaneous inferences from single time point analyses for a joint set analysis. In addition, the overall profile analysis (global null test) is too coarse to handle the subtle patterns like (1)–(4).

A useful framework is to conceptualize the gene selection problem as a set-wise multiple testing problem, where each time point of a particular gene gives rise to a hypothesis for testing DE versus EE and therefore each DE pattern corresponds to a set of hypotheses formed by combining the tests across all time points. Then a gene is claimed to be significant if a specific combination of the null hypotheses is rejected at the set level. However, several complicated issues arise in this testing framework: (i) optimality, that is, how to optimally combine testing results of multiple time points; (ii) multiplicity, that is, how to control testing errors (such as the false discovery rate) at the set level when thousands of gene are considered simultaneously; and (iii) dependency, that is, how to account for and exploit the high temporal correlation in the time-course data. Next we shall compare and review related works and then discuss a unified approach that addresses all three issues simultaneously.

Combining simultaneous tests in sets not only improves statistical power but also provides new scientific insights; successful applications include large epidemiological surveys (Zaykin et al. 2002), meta-analysis of microarray experiments (Pyne, Futcher, and Skiena 2006) and brain imaging studies (Benjamini and Hochberg 1995; Heller et al. 2006). The multiplicity issue in simultaneous set-wise tests was formally addressed in Benjamini and Heller (2008), where a threshold for Simes combined p-value is suggested based on the well-known BH procedure (Benjamini and Hochberg 1995). Although Benjamini–Heller's procedure was suitable for many large-scale studies, the applicability of their approach is quite limited in time-course experiments. First, the combined p-value method can only test how many DE time points are in a set but cannot distinguish subtler differences among DE patterns (e.g., early response versus late response). Second, even in situations where the Benjamini–Heller procedure is applicable, it can be further improved by exploiting the temporal correlation.

One important feature of the time-course experiments is that if a gene is differentially expressed at one time point, it is very likely to remain differentially expressed at the next time point. This local dependency structure can be approximated by a hidden Markov model (HMM). Specifically, an HMM assumes that the temporal sequence of the underlying states (DE or EE) of a particular gene form a Markov chain and the observed gene expression data are independent conditioning on the hidden states. The HMM has been shown to be an effective tool for analyzing biological sequences and processes; see Churchill (1992), Krogh et al. (1994), MacDonald and Zucchini (1997), and Durbin et al. (1999), among others. Successful applications of HMM to MTC data include the clustering analysis of gene expression profiles (Schliep, Schonhuth, and Steinhoff 2003) and the significance analysis of differential expression under multiple biological conditions (Yuan and Kendziorski 2006). By utilizing the temporal dependency, the HMM approach leads to results with both increased statistical power and better scientific interpretations.

Sun and Cai (2009) proposed a data-driven procedure for testing HMM-dependent hypotheses and showed that the procedure is asymptotically optimal. However, their approach cannot be applied in time-course experiments. First, the optimality of multiple testing was only addressed for single parameter analysis, but a temporal pattern cannot be identified by simply combining the results from single time points. It is important to note that the set-wise error rate can be extremely high even if the pointwise error rate is low. Second, Sun and Cai (2009) considered a homogeneous HMM with a normal mixture as the observation distribution, whereas the time-course experiment requires different models and assumptions. Specifically, we need to consider thousands of short sequences instead of one long sequence, and it is desirable to use an inhomogeneous HMM to account for the changing transition probabilities over time. Moreover, a Gamma–Gamma hierarchical model is more appropriate than a normal mixture model for MTC data based on the works of Newton et al. (2001), Kendziorski et al. (2003), and Yuan and Kendziorski (2006). We shall develop a new testing procedure that overcomes these limitations.

A key issue in studying set-wise testing problems is in the choice of the optimality criterion. Sun and Cai (2009) considered a criterion that maximizes the expected number of correct individual states given a constraint on the FDR, whereas in pattern identification problem it is more desirable to define the optimality criterion at the sequence level. The problem of uncovering an "optimal" state sequence has been studied for an HMM; the most well-known procedure is the Viterbi algorithm (Viterbi 1967; Rabiner 1989). Yuan and Kendziorski (2006) considered this method in time-course experiments for estimating the most likely DE sequence configuration and then used the results for grouping similar genes. However, the estimated DE states are not suitable for gene selection problems because the Viterbi algorithm does not address the multiplicity issue in simultaneous inferences. Essentially a desirable procedure should rank different genes according to a score on how likely it obeys a specific DE pattern, so that a thresholding rule can be applied along the rankings to yield a gene subset for future investigation. However, across-gene comparison is not provided by the Viterbi algorithm since it only selects the most likely DE pattern for a single gene. Consequently, the resultant Type I error rate is independent of the prespecified test level and can be over conservative or severely inflated. The numerical results in Section 3 will illustrate this point.

The goal of this article is to develop a unified framework to address the optimality, multiplicity, and dependency issues for set-wise tests simultaneously. Our methodological developments involve two steps. First, we derive an oracle procedure in a compound decision theoretic framework to test multiple dependent sets of hypotheses under a new optimality criterion. Our oracle procedure aims to minimize the missed set rate (MSR) subject to a constraint on the false set rate (FSR). The second step is to mimic the oracle and develop a data-driven procedure (GLIS) that is suitable for analysis of MTC

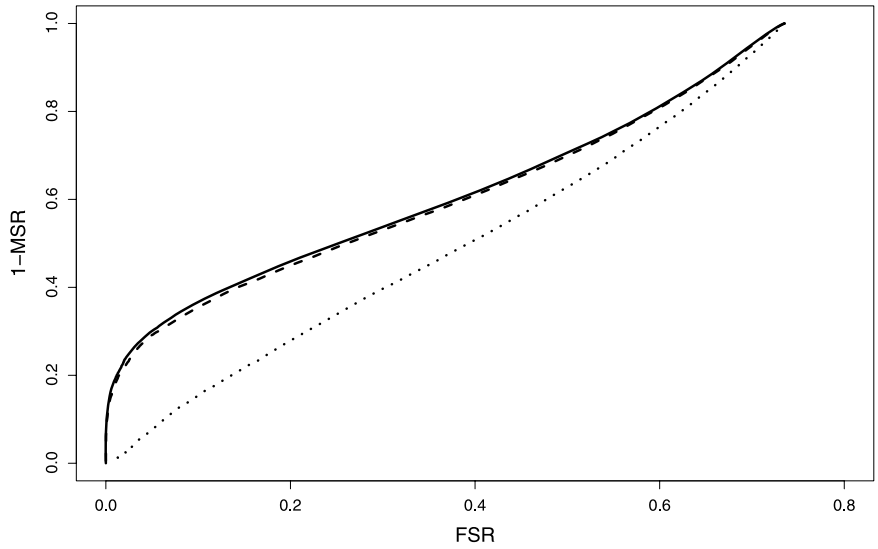

Figure 1. A comparison of the Benjamini–Heller (dotted), oracle (solid), and GLIS (dashed) procedures. Sensitivity $= 1 -$ MSR (a measure of testing power). We can see that the performance of the oracle procedure is achieved by the GLIS procedure asymptotically. At the same FSR level, both the oracle and GLIS procedures identify more true nonnulls than the Benjamini–Heller's procedure.

experiments. Specifically, we considered the more appropriate inhomogeneous HMMs and the hierarchical Gamma–Gamma model. Our data-driven procedure, which simultaneously addresses all three issues, controls the multiplicity and is asymptotically optimal. The GLIS procedure is capable of testing delicate DE patterns that cannot be handled by conventional FDR procedures. Both theoretical and numerical results are presented to show that the GLIS procedure leads to (i) asymptotically valid error rate control, (ii) improved statistical power, and (iii) more informative scientific findings.

A comparison of the Benjamini–Heller procedure, the oracle procedure and the GLIS procedure for testing against the global null (i.e., no DE at all time points) is shown in Figure 1. We can see that both the oracle and GLIS procedures have much higher sensitivity than the Benjamini–Heller procedure at the same FSR level. In addition, the performance of the oracle procedure is achieved by the GLIS procedure asymptotically. The real data analysis in Section 4 also shows substantial advantages of our method over conventional methods.

The article is organized as follows. The data structure, model assumption and method are discussed in Section 2. Simulation studies are carried out in Section 3 to compare our procedure versus conventional approaches. The methods are illustrated in Section 4 in a study of the systemic inflammation in humans. Technical details on methodological developments and proofs of theorems are given in Section 5 and the Appendix, respectively.

2. SET–WISE TESTING IN TIME–COURSE EXPERIMENTS

In time-course experiments, the gene expression levels are measured longitudinally with two possible states at each time point: equally expressed (EE) and differentially expressed (DE). Consider m sets of hypotheses: $\{(H_{i1}, \ldots, H_{iK}) : i = 1, \ldots, m\}$ for testing EE versus DE, where m is the number of genes and K is the number of time points. The problem of identifying genes with specific patterns involves the simultaneous testing of hypotheses at the set level. Typical questions considered by previous research include: (i) Are all K hypotheses in the set true? (ii) Are all K hypotheses in the set false? (iii) Are at least u out of K hypotheses in the set false? In Benjamini and Heller (2008), these questions are referred to as conjunction test, disjunction test, and partial conjunction test, respectively. The conjunction and disjunction tests can be viewed as special cases of the partial conjunction test. A big limitation of the above framework is that partial conjunction test can only handle patterns defined in terms of the total number of nonnulls in a set. However more complicated patterns that involve temporal ordering, such as "early response" and "late response," cannot be distinguished from each other. Next we introduce a more general testing framework that effectively describes various temporal patterns by defining appropriate null parameter spaces.

2.1 Characterizing Patterns in a Multiple Testing Framework

Consider the unknown state sequence $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})$. The null space for testing against the conjunction of null hypotheses (i.e., a gene is EE at all time points) is $\Theta_0^{\text{conj}} = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \in \{0,1\}^K : \sum_{k=1}^K \eta_k = 0\}$. For a partial conjunction test that at least u hypotheses in a set are false, the null parameter space is $\Theta_0^{\text{pconj}} = \{\boldsymbol{\eta} \in \{0,1\}^K : \sum_{k=1}^K \eta_k < u\}$. When temporal ordering is involved in a pattern, we need to go beyond the framework of partial conjunction test. Suppose the pattern of interest is "late response," which means that there are responses but no response prior to time k. It is more convenient to state the testing problem in terms of the nonnull parameter space Θ_1. Specifically, we define $\Theta_1^{\text{late},k} = \{\boldsymbol{\eta} \in \{0,1\}^K : \sum_{i=1}^{k-1} \eta_i = 0$ and $\sum_{i=k}^K \eta_i \geq 1\}$ as the nonnull parameter space for identifying late response genes. It is easy to see that the partial conjunction test is incapable of describing this pattern since it only has the information on the total number of nonnulls in a set, therefore the

combined p-value method (Benjamini and Heller 2008) cannot be used for detecting the "late response" pattern.

The gene selection problem involves the simultaneous testing of m sets of hypotheses; the inflation of Type I errors is a big issue. Similar to the ideas of false discovery rate (FDR, Benjamini and Hochberg 1995) and false nondiscovery rate (FNR, Genovese and Wasserman 2002), we define the false set rate (FSR) and missed set rate (MSR) to combine the errors in set-wise testing. The FSR and MSR will serve as the target for control and measure of power, respectively. Define a binary vector $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_m) \in \{0, 1\}^m$, where

$$\vartheta_i = 1 \quad \text{if } \boldsymbol{\theta}_i \in \Theta_0 \qquad \text{and} \qquad \vartheta_i = 1 \quad \text{otherwise.} \tag{2.1}$$

We are interested in testing m new hypotheses: $\mathcal{H}_{i0} : \boldsymbol{\theta}_i \in \Theta_0$ versus $\mathcal{H}_{i1} : \boldsymbol{\theta}_i \in \Theta_1$, $i = 1, \ldots, m$, where gene i is selected if the null hypothesis $\mathcal{H}_{i0}$ is rejected at the set level. The FSR and MSR are then defined as

$$\begin{aligned} \mathrm{FSR} &= E\left\{ \frac{\sum_{i=1}^m (1 - \vartheta_i)\delta_i}{(\sum_{i=1}^m \delta_i) \vee 1} \right\} \qquad \text{and} \\ \mathrm{MSR} &= E\left\{ \frac{\sum_{i=1}^m \vartheta_i (1 - \delta_i)}{(\sum_{i=1}^m \vartheta_i) \vee 1} \right\}, \end{aligned} \tag{2.2}$$

respectively. The FSR is the expected proportion of falsely rejected sets among all rejections and the MSR is the expected proportion of nonnull sets that are missed. Our ultimate goal is to find the "optimal" subset of genes for future experimental investigations. Under the multiple testing framework, *optimality* means that our procedure for subset construction controls the FSR at level α with the smallest MSR.

An efficient procedure should be developed based on a statistical model that well describes the data. The issues regarding the data structure, model and assumptions in time-course experiments will be discussed in the next section. An asymptotically optimal data-driven procedure for pattern identification is then introduced in Section 2.3.

2.2 HMM and Gamma–Gamma Model for Time-Course Data

Hidden Markov models have been successfully applied for analysis of time-course data (Schliep, Schonhuth, and Steinhoff 2003; Yuan and Kendziorski 2006). Let $\mathbf{e}_{ik} = (\mathbf{x}_{ik}, \mathbf{y}_{ik})$ denote the observed expression levels of gene i at time k, $i = 1, \ldots, m$; $k = 1, \ldots, K$. Here $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikn_1})$ and $\mathbf{y}_{ik} = (y_{ik1}, \ldots, y_{ikn_2})$ are expression data for n_1 and n_2 replicated measurements under two biological conditions X and Y, respectively. The state sequence of gene i over time is a binary vector $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})$, where $\theta_{ik} = 1$ indicates that gene i at time k is DE and $\theta_{ik} = 0$ otherwise. Assume that

$$\boldsymbol{\theta}_i \text{ is distributed as a Markov chain} \qquad \text{and} \tag{2.3}$$

$$\boldsymbol{\theta}_i \text{ and } \boldsymbol{\theta}_j,\ i \neq j, \text{ are independent.} \tag{2.4}$$

Let $\{\pi_j^i = P(\theta_{i1} = j) : i = 1, \ldots, m; j = 0, 1\}$ be the initial probabilities. Our real data analysis suggests that the HMMs are inhomogeneous. Therefore we allow the transition probabilities $a_{jl}^{ik} = P(\theta_{i,k} = l | \theta_{i,k-1} = j)$, $j, l \in \{0, 1\}$, to depend on time k, but assume that π_j^i and a_{jl}^{ik} are the same for all genes (hence i will be suppressed for transition probabilities).

A gene under one specific condition has a latent mean expression level μ^x or μ^y. To take advantage of information sharing, it is assumed that the latent mean expression levels follow a common genome-wide distribution. Following Yuan and Kendziorski (2006), we assume that at time k, the observed expression levels of gene i under condition X are n_1 independent samples from the conditional distribution $h_k(\cdot | \mu_{ik}^x)$, where μ_{ik}^x is the mean expression level of gene i, and $\{\mu_{ik}^x : i = 1, \ldots, m\}$ follow a common distribution $G_k(\cdot)$. Similarly, the expression data under condition Y are n_2 independent samples from $h_k(\cdot | \mu_{ik}^y)$, and μ_{ik}^y follows the same genome-wide distribution $G_k(\cdot)$. When it is EE ($\theta_{ik} = 0$), we have $\mu_{ik}^x = \mu_{ik}^y = \mu_{ik}$; hence the conditional density of $\mathbf{e}_{ik} = (\mathbf{x}_{ik}, \mathbf{y}_{ik})$ is

$$f(\mathbf{e}_{ik} | \mu_{ik}, \theta_{ik} = 0) = \prod_{j=1}^{n_1+n_2} h_k(e_{ikj} | \mu_{ik}). \tag{2.5}$$

Alternately, when it is DE ($\theta_{ik} = 1$) we have $\mu_{ik}^x \neq \mu_{ik}^y$. We sample μ_{ik}^x and μ_{ik}^y independently from $G_k(\cdot)$, then we have

$$f(\mathbf{e}_{ik} | \mu_{ik}^x, \mu_{ik}^y, \theta_{ik} = 1) = \prod_{j=1}^{n_1} h_k(x_{ikj} | \mu_{ik}^x) \prod_{l=1}^{n_2} h_k(y_{ikl} | \mu_{ik}^y). \tag{2.6}$$

Let p_k be the proportion of DE genes at time k, then the marginal density of $\mathbf{e}_{ik}$ is $f_k(\mathbf{e}_{ik}) = (1 - p_k) f_{k0}(\mathbf{e}_{ik}) + p_k f_{k1}(\mathbf{e}_{ik})$, where $f_{0k}(\mathbf{e}_{ik}) = \int \prod_{j=1}^{n_1+n_2} h_t(e_{ikj} | \mu_{ik})\, dG_t(\mu_{ik})$ and $f_{1k}(\mathbf{e}_{ik}) = \int \prod_{j=1}^{n_1} h_k(x_{ikj} | \mu_{ik}^x)\, dG_k(\mu_{ik}^x) \cdot \int \prod_{l=1}^{n_2} h_k(y_{ikl} | \mu_{ik}^y)\, dG_k(\mu_{ik}^y)$ are the marginal densities under EE and DE, respectively. We further assume that $\mathbf{e}_{ik}$'s are conditionally independent:

$$f(\mathbf{e}_{i1}, \ldots, \mathbf{e}_{iK} | \boldsymbol{\theta}_i) = \prod_{k=1}^{K} f(\mathbf{e}_{ik} | \theta_{ik}). \tag{2.7}$$

We call (2.3)–(2.7) a hidden Markov model (HMM) for time-course data. Denote by $\mathcal{A}_k = \{a_{jl}^k : j, l = 0, 1\}$ the transition matrix at time k, $k = 1, \ldots, K - 1$, $\boldsymbol{\pi} = \{\pi_0, \pi_1\}$ the initial probability distribution, $\mathcal{F}_k = \{h_k, G_k\}$ the observation distributions, and

$$\Psi = (\boldsymbol{\pi}, \mathcal{A}_1, \ldots, \mathcal{A}_{K-1}, \mathcal{F}_1, \ldots, \mathcal{F}_K)$$

the collection of all HMM parameters.

Choosing appropriate observation distribution $h_k(\cdot)$ and genome-wide distribution $G_k(\cdot)$ is another important issue. The Gamma–Gamma (GG) model has been shown to be a useful model that maintains most inherent characteristics of the microarray data (Newton et al. 2001; Kendziorski et al. 2003). Specifically in time-course experiments, a GG model assumes that $h_k(\cdot | \mu_{ik})$ (μ_{ik} may be taken as μ_{ik}^x or μ_{ik}^y) is a Gamma density function with shape parameter $\alpha_k > 0$ and rate parameter $\lambda_{ik} = \alpha_k / \mu_{ik}$. Fixing α_k, λ_{ik} is assumed to follow another Gamma distribution $G_k(\alpha_{0k}, \nu_k)$, where α_{0k} and ν_k are the shape parameter and rate parameter, respectively. Then after integrating out μ_{ik}, we can obtain explicit forms for $f_k(\cdot)$ with unknown parameters $\Gamma_k^0 = (\alpha_k, \alpha_{0k}, \nu_k)$. The EM algorithm can be used to estimate Γ_k. The GG model will be used in both our simulation studies and data analysis. See Newton et al. (2001) and Kendziorski et al. (2003) for more details about this model.

2.3 The Proposed Procedure for Pattern Identification

In pattern identification problems, the goal is to find a subset of genes which exhibit the temporal pattern of interest. After an HMM-GG model is fitted, the key issue is how to draw inferences based on the fitted model. We solve the problem in two steps: first rank all genes based on the likelihood of obeying a specific temporal pattern, then choose a threshold along the rankings. In a multiple testing framework, the goal in the ranking step is to derive a test statistic T to reflect the relative importance of each gene, and the goal in the thresholding step is to decide a critical value for T to control the false set rate.

Denote by $\hat{\Psi}$ the collection of all estimated parameters. In Section 5, we show in a compound decision-theoretic framework that a desirable statistic for ranking is the generalized local index of significance (GLIS):

$$\begin{aligned}\mathrm{GLIS}_i &= P_{\hat{\Psi}}(\vartheta_i = 0|\mathbf{e}_i)\\ &= \sum_{\mathbf{s}=\{s_1,\ldots,s_K\}\in\Theta_0} P_{\hat{\Psi}}(\boldsymbol{\theta}_i = \mathbf{s}|\mathbf{e}_i),\end{aligned} \tag{2.8}$$

which is asymptotically optimal for set-wise multiple testing when $\hat{\Psi}$ is a consistent estimate. Here ϑ_i is a binary random variable indicating whether a gene is interesting, $\mathbf{s}$ is a K-dimensional binary vector and Θ_0 is the null parameter space. The GLIS statistic can be interpreted as the probability of a gene being null (i.e., not exhibiting a pattern of interest) given the observed expression data. Compared to the combined p-value which can only be used for testing the total number of nonnulls in a set, the GLIS statistic can be used to screen for various complicated patterns. In addition, the GLIS procedure, combined with the HMM-GG model, can exploit the temporal correlation structure and borrow information across genes; hence it produces more accurate and interpretable results than conventional methods.

In Sun and Cai (2009), it was shown that the local index of significance (LIS) is the optimal statistic for testing hypotheses arising from an HMM. However, the LIS statistic can only be used for testing a single sequence of correlated hypotheses. Our problem is different since we have thousands of sequences, and we are interested in set-wise analysis instead of point-wise analysis. The optimality criteria of the two problems are also different in the sense that LIS maximizes the number of true individual point discoveries whereas GLIS maximizes the number of true set discoveries. In addition, the LIS procedure, which is designed for controlling the FDR, in general leads to an inflated FSR level.

To control the FSR, we determine a data-driven threshold for GLIS using the following step-up procedure. First we rank the GLIS values ascendingly as $\mathrm{GLIS}_{(1)}, \ldots, \mathrm{GLIS}_{(m)}$, and denote by $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses, then we choose the largest i such that $(1/i)\sum_{j=1}^{i} \mathrm{GLIS}_{(j)} \le \alpha$. In Theorem 4 we show that the data-driven procedure controls the FSR level at the nominal level and is asymptotically optimal. The key difference between GLIS procedure and conventional methods is in the gene rankings. By optimally combining testing results from individual time points and exploiting the HMM dependency, the GLIS rankings are more efficient than the rankings by combined p-values; this is illustrated in the simulation studies conducted in Section 3.

The implementation of GLIS procedure is simple and fast. The maximum likelihood estimate (MLE) can be used to estimate Ψ. Under certain regularity conditions, the MLE is strongly consistent and asymptotically normal (Leroux 1992; Bickel, Ritov, and Rydén 1998). The MLE can be obtained using the EM algorithm or other numerical optimization schemes. The EM algorithms for Gamma–Gamma model and normal mixture model were provided, for example, in Yuan and Kendziorski (2006) and Sun and Cai (2009), respectively. Given the HMM parameters $\hat{\Psi}$, the GLIS statistic can be calculated as

$$\widehat{\mathrm{GLIS}}_i = \frac{\sum_{\mathbf{s}\in\Theta_0} \pi_{s_1} f(\mathbf{e}_{i1}|s_1) \prod_{j=2}^{K} a_{s_{j-1}s_j} f(\mathbf{e}_{ij}|s_j)}{\alpha_{i,K}(0) + \alpha_{i,K}(1)},$$

where $\alpha_{i,k}(j) = P_\Psi[(\mathbf{e}_{it})_{t=1}^{k}, \theta_{ik} = j]$ is the forward variable that can be calculated using the *forward–backward procedure* (Baum et al. 1970; Rabiner 1989). Specifically, let $\alpha_{i,1}(j) = \pi_j f_{j1}(\mathbf{e}_{i1})$, then by induction we have $\alpha_{i,k+1}(j) = \{\sum_{l=0}^{1} \alpha_{i,k}(l) a_{lj}^{k}\} f_{j,k+1}(\mathbf{e}_{i,k+1})$.

2.4 Mixed Directional Errors

In this section, we discuss an extension of our GLIS procedure to deal with the *mixed directional errors*. The issue was previously studied by Finner (1999) and Benjamini and Yekutieli (2005), and recently raised by Guo, Sarkar, and Peddada (2010) in the context of MTC experiments. Guo, Sarkar, and Peddada (2010) considered the identification of specific expression patterns over ordered categories under one biological condition. The concern is that, in addition to the usual Type I errors, directional or sign errors can occur for rejected hypotheses. To resolve the issue, Guo, Sarkar, and Peddada (2010) extended the BH step-up procedure (Benjamini and Hochberg 1995) to control the so-called *mixed directional FDR* (mdFDR) of ordered patterns.

Similar issues remain for MTC experiments under two biological conditions. Define $\Delta = \mu^x - \mu^y$ to be the difference of gene expression levels. It is generally accepted that DE genes with time-condition interactions (i.e., both $\mu^x < \mu^y$ and $\mu^x > \mu^y$ occur for the same gene over the experimental period) are more biologically meaningful. In contrast, the patterns where the expression levels of one condition are dominant at all time points are of less interest because they are likely to be caused by the baseline differences between the case and control groups (Ma, Zhong, and Liu 2009). The identification of these patterns involves the inference about the sign of Δ; hence additional directional errors can occur.

The pattern identification problem is essentially a set-wise testing problem. From a biological point of view, the FSR is a more appealing concept than the mdFDR in Guo et al. (2010) for inference of patterns since scientific discoveries are usually made at the gene (or set) level, and specific patterns are more informative than the individual states to reveal the underlying biological mechanisms. It is natural to consider multiple testing procedures that aim to control the number of misspecified patterns. The concept of mdFDR is only useful for point-wise analysis. In particular, the FSR can be very high even if the mdFDR is controlled at the nominal level. Consider a situation where we make 20 errors out of 40 selected genes over 5 time points. In a point-wise analysis, the proportion of false positives

is 0.10; however in a set-wise analysis, the proportion of false sets can be as high as 0.5 if the 20 errors occur at exactly 20 different genes.

Dealing with mixed directional errors in conventional testing framework is complicated. When a set-wise analysis is of primary interest, we may circumvent the complication by considering a three-state HMM. The key observation is that the *mixed directional errors* are the consequence of using a binary decision rule to deal with a trichotomous situation. Therefore it is natural to define a multinomial variable

$$\theta_{ik} = \begin{cases} -1, & \text{if } \mu_{ik}^x < \mu_{ik}^y \text{ or } \Delta_{ik} < 0 \\ 0, & \text{if } \mu_{ik}^x = \mu_{ik}^y \text{ or } \Delta_{ik} = 0 \\ 1, & \text{if } \mu_{ik}^x > \mu_{ik}^y \text{ or } \Delta_{ik} > 0 \end{cases} \tag{2.9}$$

for our analysis. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK}) \in \{-1, 0, 1\}^K$ be the sequence of unknown states over time. Then for a pattern of interest, we may use a new binary variable to combine the individual states in a set as before:

$$\vartheta_i = 1 \quad \text{if } \boldsymbol{\theta}_i \in \Theta_0 \qquad \text{and} \qquad \vartheta_i = 1 \quad \text{otherwise.} \tag{2.10}$$

To identify genes that interact with time, we may define the null parameter space $\Theta_0^{\text{interact}} = \{\boldsymbol{\eta} \in \{-1, 0, 1\}^K : |\sum_i \eta_i| = \sum_i |\eta_i|\}$. The expression of GLIS shall remain the same as in (2.8) for appropriately defined θ_{ik}'s and Θ_0, and the same procedure can be applied for determining an appropriate cutoff for GLIS to control the FSR. By extending our two-state HMM to a three-state HMM, the issue of mixed directional errors no longer exists.

The forward–backward procedure and EM algorithm are easy to implement for a three-state HMM. However, a challenging issue is to specify an appropriate hierarchical model for the three-state HMM to replace (2.5)–(2.6). Specifically, the extension of the GG model to reflect both the sign and magnitude of the differences in expression levels is non-trivial. Finally, our testing framework is designed for set-wise inference, and it is interesting to extend it to consider mixed directional errors for pointwise testing as formulated by Guo, Sarkar, and Peddada (2010). Much research is needed under the formulation that involves mixed directional errors; in particular the dependency and optimality issues seem to be complicated. We leave these important topics for future research.

3. SIMULATION STUDIES

In this section, we first introduce some alternative approaches for pattern identification and then compare their performances with that of our method.

3.1 Alternative Methods

The first alternative method is the well-known Viterbi algorithm, which aims to estimate the most probable state sequence of a gene given the observed data. It seems natural to select the genes whose most probable configuration suggested by the Viterbi algorithm obeys the pattern of interest. Specifically, we calculate the most probable state sequence $\hat{\boldsymbol{\theta}}_i$ for gene i, $i = 1, \ldots, m$; then set $\delta_i = 1$ if $\hat{\boldsymbol{\theta}}_i \in \Theta_1$, and $\delta_i = 0$ otherwise. However, the Viterbi algorithm does not address the multiplicity issue in simultaneous inferences because it was only developed for selecting a pattern for a single gene, whereas in gene selection problems, across gene comparison is needed. For example, it is not clear how to select the "top 10" genes based on the results obtained from the Viterbi algorithm. We shall see that the Type I error rate of Viterbi algorithm is independent of any prespecified test level and hence can be either too conservative or too liberal.

The second approach is based on thresholding the combined p-values (Benjamini and Heller 2008). This procedure addresses the multiplicity issue in simultaneous set-wise inferences but is limited in its applicability and power. Let $p_{(1)}^i, \ldots, p_{(K)}^i$ be the ordered p-values from the ith set. Denote by $H_{u/K}^i$ the partial conjunction test that at least u out of K hypotheses in set i are false. The Simes' p-value $p_{u/K}^i = \min\{(\frac{K-u+1}{j})p_{(u-1+j)}^i : j = 1, \ldots, K-u+1\}$ can be used to summarize the K p-values from set i into a single index. The FDR procedure in Benjamini and Hochberg (1995) was then applied to the ordered Simes' p-values:

$$\text{let } l = \max\left\{i : p_{u/K}^{(i)} \le \frac{i}{m}\alpha\right\},$$
$$\text{then reject } H_{u/K}^{(i)}, i = 1, \ldots, l. \tag{3.1}$$

Benjamini and Heller (2008) showed that the procedure (3.1), referred to as the BH–Simes procedure, controls the FSR at the nominal level α. In addition, it was shown that the BH–Simes procedure is still valid under different dependency assumptions in the sense that it controls the FSR at level α. When a specific temporal pattern can be described as conjunction or partial conjunction tests, the BH–Simes procedure may be applied. Specifically, let t_{ik} be the two sample t-statistic of gene i at time k for comparing two biological conditions, then t_{ik} can be converted to a p-value using transformation $p_{ik} = 2F(-|t_{ik}|)$, where F is the cdf of the t-variable. We can first obtain the Simes' p-value $p_{u/K}^i$ for gene i, then apply the BH procedure. However, the BH–Simes procedure is only applicable for partial conjunction test but incapable of dealing with patterns that involves temporal ordering such as "early or late response." Moreover, the rankings by combined p-values can be much improved by the rankings of GLIS statistics which exploit the dependency structure.

In the next section, we design and conduct three simulation studies to compare the numerical performances of Viterbi, BH–Simes and GLIS. In Section 5, we also derive an "oracle" procedure that assumes all parameters are known. The oracle procedure, which is included in the comparison as well, provides a benchmark for defining optimality and comparing different procedures. We will see that the performance of the oracle procedure is asymptotically achieved by the GLIS procedure. We comment here that the simulation results show that GLIS is more powerful than BH–Simes and Viterbi. Another advantage is that GLIS is more flexible and can be used to test a variety of complicated patterns.

3.2 Simulation Results

Simulation Studies 1 and 2 consider conjunction and partial conjunction tests, respectively. To provide insights into the superiority of our procedure, we investigate the ranking efficiencies of GLIS versus BH–Simes in Simulation Study 3. To the

best of our knowledge, no multiple testing procedures can be used to screen for the temporal patterns (1)–(4) discussed in Section 1. Therefore we only illustrate how to implement the GLIS procedure for identifying such delicate patterns in Section 4 without doing comparisons.

In all simulations, the numbers of subjects under both conditions are $n_1 = n_2 = 10$, the number of genes is $m = 2000$, the number of time points measured for each gene is $K = 6$ and the number of replications is 200. The nominal FSR level is 0.1. For a given gene i, a Markov chain $\boldsymbol{\theta}_i = (\theta_{ik})_{k=1}^K$ is first generated with initial state distribution $\boldsymbol{\pi} = (0.95, 0.05)$ and transition matrix $\mathcal{A}_k = (a_{00}, 1 - a_{00}; 1 - a_{11}, a_{11})$, $k = 1, \ldots, 5$. The observed data $\mathbf{e}_{ik}$ are then generated conditionally on $(\theta_{ik})_{k=1}^K$ with GG parameters $\Gamma_k = (\alpha_k, \alpha_{0k}, \nu_k)$. We shall specify $\mathcal{A}_k$ and Γ_k later in each simulation study.

Simulation Study 1. This simulation study compares the efficiency of Viterbi, BH–Simes, and GLIS for conjunction tests, where a rejection is made at the set level if any hypothesis in a set is false. The simulation results are summarized in Figure 2. In the top row, we choose $\mathcal{A}_k = (0.95, 0.05; 1 - a_{11}, a_{11})$, $(\alpha_k, \alpha_{0k}, \nu_k) = (10, 1, 0.5)$, $k = 1, \ldots, 5$, and then apply the Viterbi algorithm, BH–Simes procedure, the oracle procedure, and the data-driven GLIS procedure to the simulated data. The FSR and MSR levels are plotted as functions of a_{11}. In the bottom row, we choose $\mathcal{A}_k = (0.95, 0.05; 0.2, 0.8)$ and $(\alpha_k, \alpha_{0k}, \nu_k) = (\alpha, 1, 0.5)$, $k = 1, \ldots, 6$. The FSR and MSR levels are plotted as functions of α.

Simulation Study 2. This simulation study compares the efficiencies of different FSR procedures for partial conjunction tests, where a rejection is made at the set level if the number of false hypotheses in a set is greater than or equal to 3 (the total number of time points is 6). The simulation results are summarized in Figure 3. In the top row, we choose $\mathcal{A}_k = (0.95, 0.05; 1 - a_{11}, a_{11})$ and $(\alpha_k, \alpha_{0k}, \nu_k) = (10, 1, 0.5)$, $k = 1, \ldots, 6$. In the bottom row, we choose $\mathcal{A}_k = (0.95, 0.05; 0.2, 0.8)$ and $(\alpha_k, \alpha_{0k}, \nu_k) = (\alpha, 1, 0.5)$, $k = 1, \ldots, 6$. The FSR and MSR levels are plotted as functions of α.

Performance of Viterbi-based decision rule. In Simulation Study 1 for conjunction test (Figure 2), Viterbi algorithm results in the most conservative FSR levels. The MSR of Viterbi algorithm is much higher than that of GLIS but comparable with that of BH–Simes. In Simulation Study 2 for partial conjunction test (Figure 3), both of the FSR and MSR of the Viterbi algorithm are between those of $\widehat{\text{GLIS}}$ and BH–Simes. The Viterbi-based decision rule does not take the prespecified test level into account; hence the decisions would be fixed regardless of the choice of the test level. Here we choose the FSR level to be 0.1 and observe that the Viterbi algorithm is conservative. Imagine if the FSR level is 0.01, the Viterbi algorithm will be too liberal. In contrast, the GLIS procedure is adaptive to our choice of test level and controls the FSR precisely.

$\widehat{\text{GLIS}}$ (GLIS) versus BH–Simes. From Figures 2 and 3, we can see that (i) the oracle and GLIS procedures can achieve the

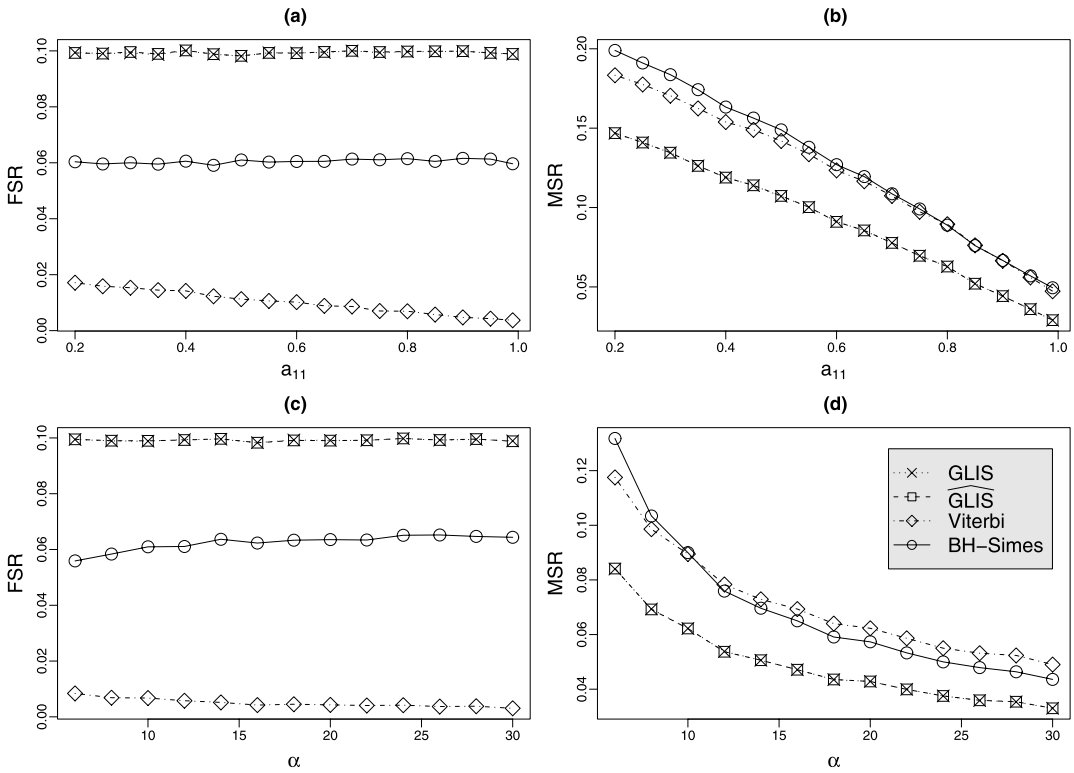

Figure 2. Comparison for conjunction tests: GLIS, the oracle procedure; $\widehat{\text{GLIS}}$: the data-driven procedure; Viterbi: Viterbi decision rule; BH–Simes: the combined p-value procedure. FSR of Viterbi is independent of the nominal level. BH–Simes is conservative and missed a large proportion of true signals. $\widehat{\text{GLIS}}$ controls the FSR precisely and has greater power. It also achieves the performance of the oracle procedure asymptotically.

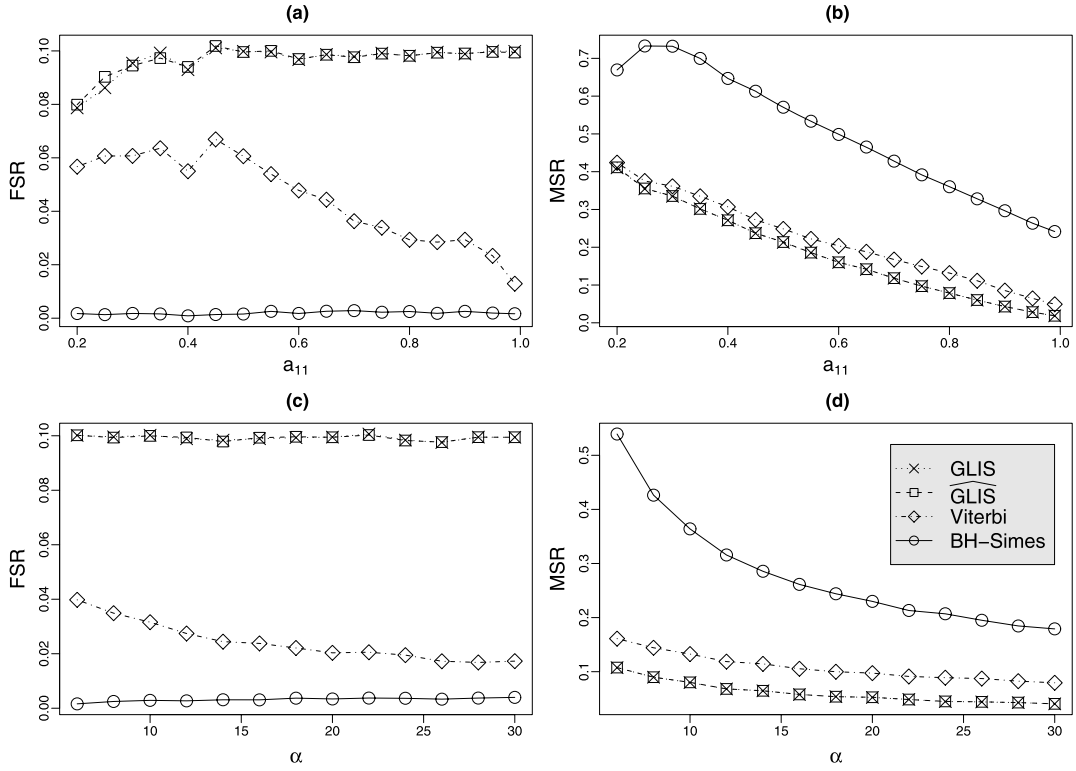

Figure 3. Comparison for partial conjunction tests: GLIS, the oracle procedure; $\widehat{\text{GLIS}}$: the data-driven procedure; Viterbi: Viterbi decision rule; BH–Simes: the combined p-value procedure. GLIS and $\widehat{\text{GLIS}}$ can achieve the nominal FSR levels approximately. In contrast, BH–Simes is very conservative and misses a big proportion of nonnulls. The MSR of BH–Simes is significantly higher than GLIS. The performance of Viterbi is in between.

nominal FSR levels accurately, whereas the BH–Simes procedure is very conservative; (ii) the two lines of the oracle procedure and data-driven GLIS procedure are almost overlapped, indicating that the performance of the oracle procedure is attained by the data-driven procedure asymptotically; (iii) the MSR levels of the data-driven procedure are lower than that of the BH–Simes procedure in all settings, even when the dependence is weak; (iv) the difference in MSR levels is larger as α increases; (v) the gain in efficiency is larger for partial conjunction tests. Note that α controls the signal strength, the results imply that our procedure is especially useful when the signal is weak.

It is important to note that the larger power of GLIS is not gained at the price of a higher FSR level. To illustrate this point, we conduct the following simulation study to compare the ranking efficiencies of GLIS statistics versus combined p-values.

Simulation Study 3 (Ranking efficiency). In this simulation study, we compare the MSR levels of the GLIS procedure versus the BH–Simes procedure at the same FDR level. Because the FSR levels of the Viterbi algorithm is fixed, we do not include Viterbi in the comparison. The simulation results are summarized in Figure 4. The top row considers conjunction tests and the bottom row considers partial conjunction tests, where a rejection is made at the set level if the number of false hypotheses is greater than or equal to 3. We choose $\mathcal{A}_k = (0.95, 0.05; 0.2, 0.8)$ and $(\alpha_k, \alpha_{0k}, \nu_k) = (\alpha, 32, 4)$, $k = 1, \ldots, 6$. In Panels (a)–(f), α is 10, 16, 30, 10, 16, and 30, respectively. We can see that (i) at the same FSR level, our proposed procedures have much lower MSR levels than the BH–Simes procedure, indicating that the rankings produced by GLIS are more efficient than those produced by the combined p-values; (ii) the efficiency gain of GLIS is large when the signals are weak; (iii) compared to conjunction tests, the difference in MSR levels is larger for partial conjunction tests.

4. APPLICATION

In this section we apply our method to a well-known MTC dataset collected by Calvano et al. (2005) for studying the systemic inflammation in human. The dataset contains eight study subjects which are randomly assigned to case and control groups and then administered with endotoxin and placebo, respectively. Affimetrix U133A chips were used to profile the expression levels of 22,283 genes in human leukocytes measured before infusion (0 hour) and at 2, 4, 6, 9, and 24 hours afterwards.

The goal of this MTC experiment is to extract information from genome-wide expression data to help identify functional networks responsible for the systemic activation. Inflammatory responses exhibit a quick, transient, and self-limiting nature. In the early activation of innate immunity, many secreted proinflammatory factors, including cytokines and chemokines, are quickly activated in response to exterior intrusion. The activated proinflammatory factors subsequently trigger the expression of several transcription factors to initiate the innate immune response. In the late period, the expression levels of a number of transcription factors limiting the innate immune response are

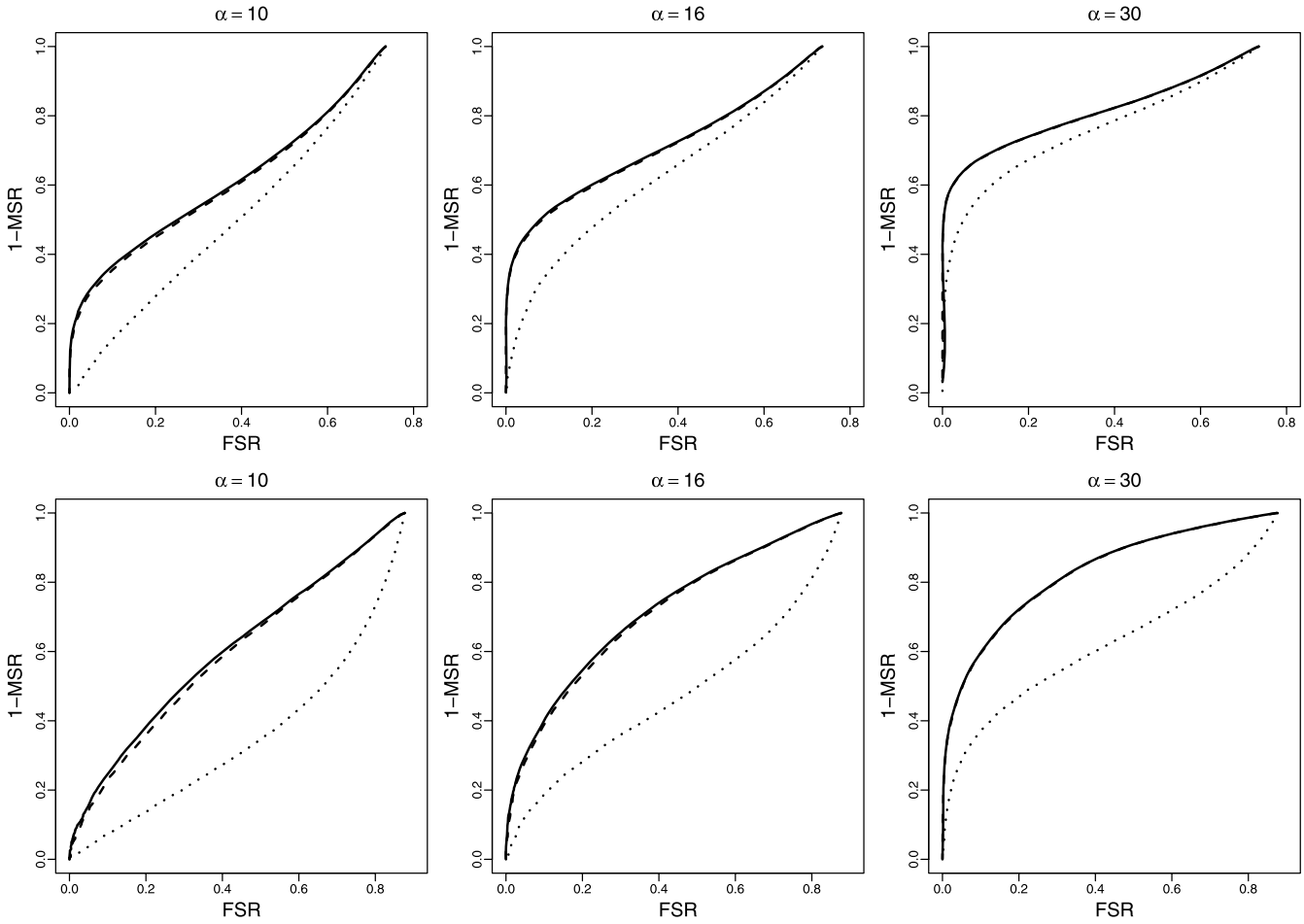

Figure 4. Ranking efficiency: BH–Simes (dotted, combined p-value), GLIS (solid, oracle), $\widehat{\text{GLIS}}$ (dashed, data-driven). GLIS and $\widehat{\text{GLIS}}$ have lower MSR/higher sensitivity than BH–Simes at the same FSR level. Sensitivity $= 1 -$ MSR.

increased. Finally the whole system concludes with full recovery and a normal phenotype. Although the overall activated-then-self-limiting pattern is known for the innate immune response process, the underlying regulatory programs are yet to be deciphered. Next we illustrate how to build an HMM and apply the GLIS procedure to analyze this MTC experiment.

Step 1 (Data preprocessing). Data preprocessing is a necessary and critical procedure to separate biologically meaningful signals from background hybridization noises and other confounding signals in microarray studies. For Affymetrix Genechip one-channel array, data preprocessing involves three steps: (a) background adjustment, (b) normalization, and (c) summarization (Gautier et al. 2004). Some popular software provides one-stop solutions for all three steps; representative ones include RMA (Irizarry et al. 2003), dChip (Li and Wong 2001), and MAS5 (Hubbell, Liu, and Mei 2002). For Illumina microarray (BeadArray), users can use lumi (Du, Kibbe, and Lin 2008), and two-color spotted cDNA arrays, Limma (Smyth and Speed 2003). We used RMA here to preprocess the raw Affimetrix array data collected by Calvano et al. (2005). It is noted that the normalized gene expression intensities may be returned in the log-transformed format as default by many software including RMA. We need to convert the transformed data back to their original scale, which is required by the GG model. After the data preprocessing step, we obtain the normalized expression intensities in original scale of 22,283 gene with four replicates for each of the two biological conditions (endotoxin and placebo) over 6 time points.

Step 2 (HMM model fitting). After the preprocessing step, we apply the EM algorithm (Yuan and Kendziorski 2006) to fit an HMM for the time-course data, with the Gamma–Gamma model as the underlying sampling distribution. The choice of a suitable model is crucial in the model building process. In particular, we expect that very few genes are perturbed in the first time point because there is no infusion of endotoxin or placebo. However, at later time points the immune system begins to respond to the two different treatments and the transition probabilities and gene expression intensities will vary over time. Hence inhomogeneous Markov chains are recommended in modeling the dynamic process, and different GG distributions are used at different time points. The fitted HMM parameters are summarized in Table 1. We observe considerably different (α, α_0, ν) and (a_{00}, a_{11}) across the 6 time points, which confirms our previous model assumptions. In situations where the estimated model parameters are very similar for all time points, it is more appropriate to use an homogeneous HMM to refit the data to increase the estimation precision.

Table 1. HMM parameter estimation: a_{00} and a_{11} denote the initial state distribution at 0 hour and the transition probabilities afterwards

HMM parameters	0 h	2 h	4 h	6 h	9 h	24 h
α	34.26	20.91	38.76	39.07	41.22	31.47
α_0	0.95	0.97	0.94	0.93	0.93	0.96
ν	4.13	7.66	4.06	3.94	3.74	4.67
a_{00} (π_0)	0.9595	0.7996	0.9196	0.9976	0.9982	0.9998
a_{11} (π_1)	0.0405	0.9258	0.9813	0.9349	0.9735	0.2034

Step 3 (Pattern design). Specifying a temporal pattern is an important step in the analysis of MTC experiments; this involves forming an appropriate hypothesis to be tested, and like in most randomized clinical trials, we recommend this step to be completed prior to the experiment. In particular, the pattern of interest should be application specific, independent of the observed data and incorporate prior information and related domain knowledge. Here we discuss the design of several commonly used patterns that meet the needs of most biological applications.

1. *Genes perturbed in response to treatments.* In many MTC experiments, the first time point usually serves as a control point and we expect no much difference in gene expression levels across the two treatments. Therefore any difference at 0 hour can be viewed as nonimmune responsive perturbations, which may be attributed to baseline difference. Hence it would be of interest to select perturbed genes in response to treatment which have EE before the inhibition of endotoxin and DE afterwards. If we further hypothesize that the immune responsive genes should conclude with full recovery and a normal phenotype after 24 hours due to the self-limiting nature of immune response, then the genes perturbed by treatment should also have EE at the end. In summary, we aim to select genes which are (i) EE at 0 or 24 hours, and (ii) DE at one or more time points from 2, 4, 6, and 9 hours. Therefore the nonnull parameter space for identifying genes perturbed in response to treatment is

$$\Theta_1^{\text{Resp}} = \left\{ \boldsymbol{\eta} \in \{0,1\}^6 : \eta_1 = \eta_6 = 0, \sum_{k=2}^{5} \eta_k \geq 1 \right\}.$$

More generally, we may want to exclude genes that are DE at all time points since this type of perturbation is often caused by the baseline difference. Hence the nonnull parameter space for identifying genes that have both EE and DE states during the experimental period is

$$\Theta_1^{\text{Mix}} = \left\{ \boldsymbol{\eta} \in \{0,1\}^6 : 1 \leq \sum_{k=1}^{6} \eta_k \leq 5 \right\}.$$

2. *Sequentially perturbed genes*. Sequentially activated genes, which are ordered temporally, may reveal meaningful activation sequence that governs the immune responses. For example, Calvano et al. (2005) identified point-wise DE genes and then clustered them into different groups based on the first time point when they became DE. Define the nonnull parameter space for identifying genes that are not DE prior to time point t but are DE afterwards:

$$\Theta_1^{t\text{-Resp}} = \left\{ \boldsymbol{\eta} \in \{0,1\}^6 : \sum_{k=1}^{t-1} \eta_k = 0, \sum_{k=t}^{6} \eta_k \geq 1 \right\}.$$

Such patterns would be particularly helpful for studying biological processes such as cell cycle and development, during which definite and specialized genes are expected to be activating sequentially at each step.

3. *Early and late response genes*. The magnitude of the difference in expression levels often varies for different genes. As a result, among genes which are perturbed at the same time point, some may be detected early while some cannot be detected until the change reaches its peak. If the studied system does not show definite change at every time point, and there is no time point of particular interest in a cell cycle and development, it would be sufficient to roughly separate the genes that responded to treatment early from those that responded late. For example, we define the following patterns, "DE within 4 h" and "not DE until 4 h or later," for early and late immune response genes, respectively. Therefore the nonnull parameter spaces for identifying early and late response genes are

$$\Theta_1^{\text{Early}} = \left\{ \boldsymbol{\eta} \in \{0,1\}^6 : \sum_{k=1}^{4} \eta_k \geq 1 \right\} \quad \text{and}$$

$$\Theta_1^{\text{Late}} = \left\{ \boldsymbol{\eta} \in \{0,1\}^6 : \sum_{k=1}^{4} \eta_k = 0, \sum_{k=5}^{6} \eta_k \geq 1 \right\},$$

respectively. The potential regulatory relationships between early and late response genes can still provide insights on how signals are transferred during the cell cycle; and the identified genes can serve as good candidates for further experimental verifications.

We just give a few examples of patterns that may be commonly used. Actually we have in total 2^K atomic patterns and each pattern represents a state sequence over the K time points. Those atomic patterns can flexibly be partitioned into any two complementary sets Θ_0 and Θ_1, by which Θ_1 represents the collection of the patterns of our interest. The number of possible partitions allowed by our testing framework is as large as $2^{2^K} - 1$, which is flexible enough to meet various biological needs.

Step 4 (Application of GLIS). With the pattern of interest in mind, the application of our GLIS procedure is relatively straightforward. As an example, we applied the GLIS procedure to identify "early" and "late" immune response genes, which are respectively defined as the genes that are DE within four hours after endotoxin injection, and the genes that begin to be DE at four hours or later. The formal definitions of Θ_1's were given in the example in Step 3. Since Θ_1 and Θ_0 are complementary, we can calculate GLIS$_i$ as either $\sum_{\mathbf{s}\in\Theta_0} P_{\hat{\Psi}}(\boldsymbol{\theta} = \mathbf{s}|\mathbf{e}_i)$ or $1 - \sum_{\mathbf{s}\in\Theta_1} P_{\hat{\Psi}}(\boldsymbol{\theta} = \mathbf{s}|\mathbf{e}_i)$, depending on which set has a smaller cardinality, to save computational time.

The HMM parameters, which were estimated at Step 2, are used to calculate the values of GLIS statistics. The genes are then ranked based on GLIS and a cutoff is chosen along the rankings for a given FSR level using the step-up procedure. At

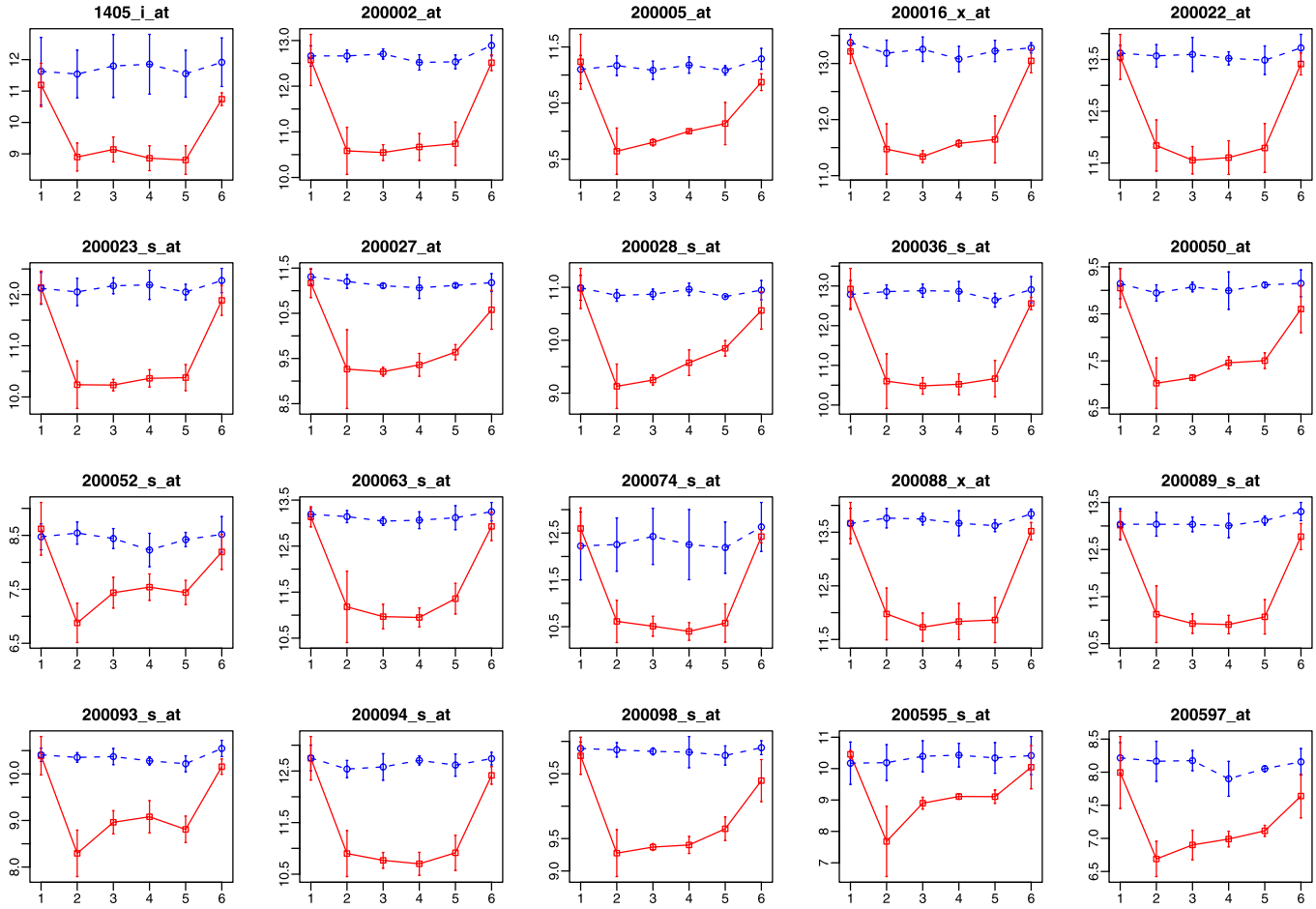

Figure 5. Expression profiles of top 20 "early" genes with 95% confidence bands. Vertical and horizontal axes are the log-transformed gene expression level and time (0, 2, 4, 6, 9, 24), respectively. Solid line: cases; dashed line: controls. The online version of this figure is in color.

the FSR level of 0.05, 4385 genes are identified to be "early" response genes and 56 are identified to be "late" response genes. The expression profiles of the top 20 "early" and "late" response genes are shown in Figures 5 and 6, respectively.

Interpretation and discussion of results. A number of genes, many of which are known to take part in initiating the innate immune response and implement many functions of leukocytes, such as cellular movement, migration and proliferation, are identified as "early" response genes by GLIS (Table 2). In addition, it is interesting to find that Calcineurin (PPP3CA) is among the "late" immune response genes. Specifically, its expression profile shows a significant down-regulation at six hours followed by a gradual recovery. Calcineurin induces transcription factors for the transcription of IL-2 genes, the amount of which is believed to have significant influence on the extent of the immune response. Similar down-regulation of the LTB, an inducer of the inflammatory response system, was also found among the "late" immune response genes. Meanwhile, immune response repressors were observed to be up-regulated, including the BCL6 (a corepressor of the transcription of START-dependent IL-4 responses of B cells), and LILRA6 and LILRB3 (leukocyte immunoglobulin-like receptors that bind to MHC class I molecules on antigen-presenting cells to inhibit the stimulation of an immune response). These up/down regulations together indicate a concluding signal, which is consistent with the self-limiting nature of the innate immune response. The application of our GLIS procedure to time-course data provides more insights in deciphering the cell regulatory mechanism. The possible regulatory relationships between these "early" and "late" response genes are worthy of further biological investigations.

5. TECHNICAL DETAILS AND DERIVATIONS OF THE PROPOSED PROCEDURE

In this section, we develop optimal procedures in a compound decision-theoretic framework for testing sets of hypotheses arising from the HMM defined in (2.3)–(2.7). We first show that the multiple testing and weighted classification problems are "equivalent" under mild conditions, then derive an oracle procedure, under an ideal setting, that minimizes the MSR subject to a constraint on the FSR. Finally we derive a data-driven procedure that mimics the oracle procedure. The Appendix shows that the data-driven procedure is valid and optimal asymptotically.

5.1 Compound Decision Theory

Consider ϑ_i defined in (2.1). We are interested in inference of the unknown ϑ_i's based on the observed data and need to solve m component problems simultaneously. This is referred to as a *compound decision problem* (Robbins 1951). A so-

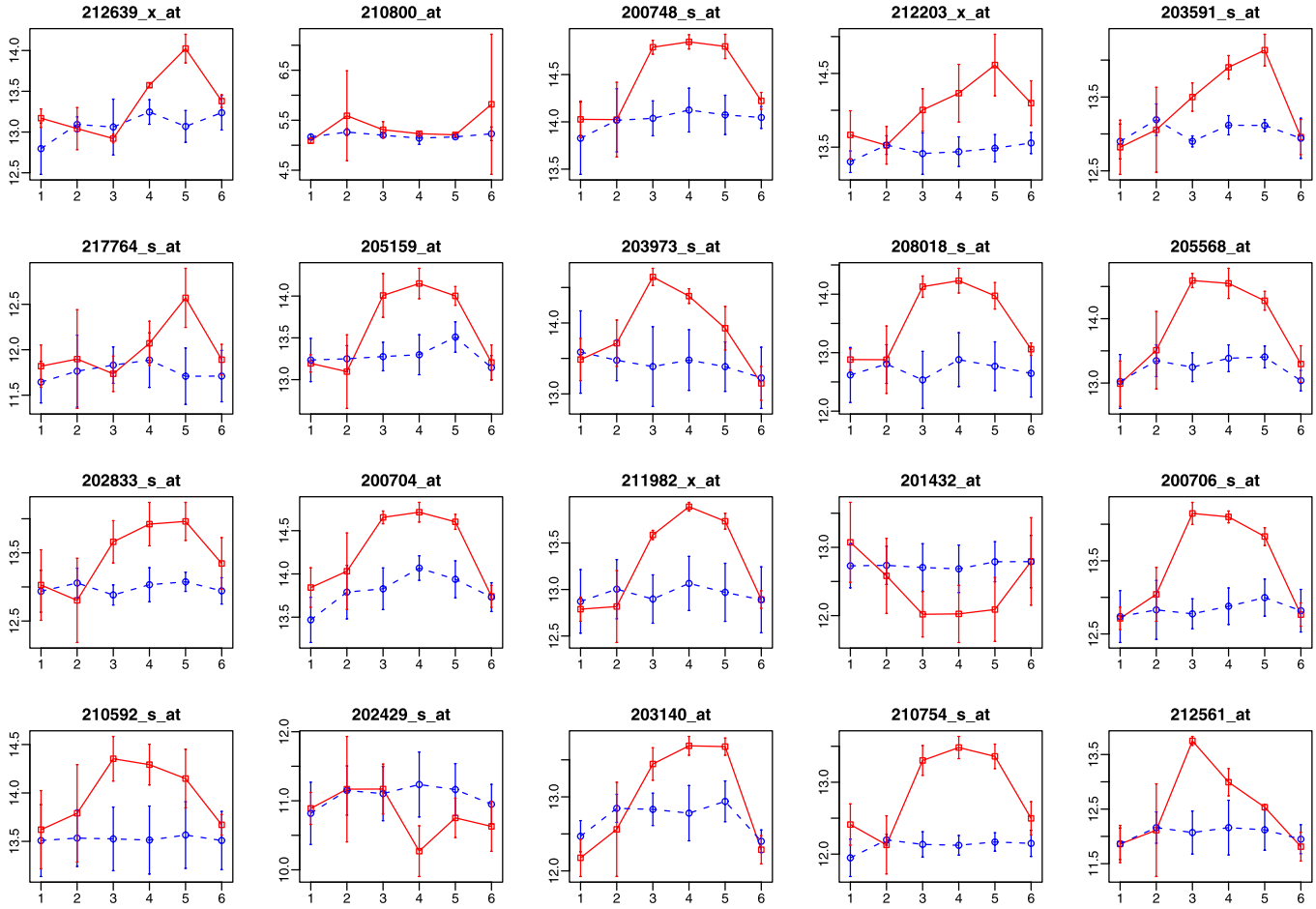

Figure 6. Expression profiles of top 20 late response genes with 95% confidence bands. Vertical and horizontal axes are the log-transformed gene expression level and time (0, 2, 4, 6, 9, 24), respectively. Solid line: cases; dashed line: controls. The online version of this figure is in color.

lution to this problem can be represented by a decision rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \{0, 1\}^m$, where $\delta_i = 1$ if we claim that $\mathcal{H}_{i0}$ is false and $\delta_i = 0$ otherwise.

In Yuan and Kendziorski (2006), the problem of separating DE time points from EE time points was formulated as a classification problem. Specifically, let $\hat{\theta}_{ik}$ be an estimate of the unknown state, with $\hat{\theta}_{ik} = 1$ indicating that we classify kth time point of gene i as DE and $\hat{\theta}_{ik} = 0$ otherwise. Yuan and Kendziorski (2006) proposed a classification rule based on the maximum *a posterior* (MAP) estimates $\hat{\theta}_{ik} = \text{argmax}_{s\in\{0,1\}} P(\theta_{ik} = s|\mathbf{e}_i)$, for $i = 1, \ldots, m$ and $k = 1, \ldots, K$. Let $\mathbf{a} = (a_{ik} : i = 1, \ldots, m; k = 1, \ldots, K) \in \{0, 1\}^{mK}$ denote a general decision rule. It can be shown that the MAP estimate is the optimal solution to a classification problem with the follow-

Table 2. The known "early" innate immune response genes

Category	Members
Cytokines, chemokines	TNFSF10, TNFSF14, TNF, IL1A, IL1B, IL8, IL15, IL32, ILF2, ILF3, CXCL1, CXCL2, CCL3, CCL20, CCL3L1, CCL3L3, CCL4, CCL5, XCL1, and XCL2
Membrane receptors of cytokines, chemokines	CCR1, CCR2, CCR3, CCR5, CCR6, CCRL2, CXCR3, CXCR4, CX3CR1, IL11RA, IL13RA1, IL18R1, IL2RB, IL2RG, IL1RAP, IL1R1, IL1R2, IL27RA
Toll-like receptors	TLR1, TLR4, TLR5, TLR7, and TLR8
Fc receptors	FCGR1A, FCGR1B, FCGR1C, FCGR2A, FCGR3A, FCGR3B
IFN receptors	IFNAR2, IFNGR2
kappa/relA family	NFKB2, RELA, RELB
Tyrosine kinase	JAK2, JAK3
STAT family	STAT1, STAT2, STAT4, STAT5B, STAT6

ing loss function

$$L(\boldsymbol{\theta}, \mathbf{a}) = (mK)^{-1} \sum_{i=1}^{m} \sum_{k=1}^{K} \{(1-\theta_{ik})a_{ik} + \theta_{ik}(1-a_{ik})\}. \quad (5.1)$$

In loss function (5.1), it is assumed that a false positive and a false negative have the same cost. However, in a gene selection problem, a false positive is often considered to be more serious than a false negative. Therefore it is desirable treat the two types of errors differently. Moreover, the analysis needs to be conducted at the set level. Hence we construct the loss function using the set indicator ϑ_i, where $\vartheta_i = 1$ if a gene has the temporal pattern of interest and $\vartheta_i = 0$ otherwise. Also, let λ denote the relative cost of a false positive to a false negative, and $\boldsymbol{\delta} = (\delta_i : i = 1, \ldots, m) \in \{0, 1\}^m$ denote a general decision rule, where $\delta_i = 1$ indicates that we claim that a gene has a specific temporal pattern of DE and $\delta_i = 0$ otherwise. Consider a *weighted classification problem* with loss function

$$L_\lambda(\boldsymbol{\vartheta}, \boldsymbol{\delta}) = m^{-1} \sum_{i=1}^{m} \{\lambda(1-\vartheta_i)\delta_i + \vartheta_i(1-\delta_i)\}. \quad (5.2)$$

The goal of this weighted classification problem is to find a decision rule $\boldsymbol{\delta}^\lambda$ that minimizes the classification risk $R_\lambda = E\{L_\lambda(\boldsymbol{\vartheta}, \boldsymbol{\delta})\}$.

In MTC studies, we expect that only a small number of genes are differentially expressed at time k. The formulation of a weighted classification problem is not tractable in practical applications since the relative cost λ is usually unknown. Alternately, a common strategy in practice is to identify a list of genes that minimizes the "missed findings," while incurring a relative low proportion of false positive results. This naturally gives rise to a *multiple testing problem*, where the optimal testing procedure $\boldsymbol{\delta}^\alpha$ minimizes the MSR subject to the constraint FSR $\leq \alpha$. The multiple testing and weighted classification problems are closely connected, but the former is more difficult to deal with theoretically. Next we show that under mild conditions the two problems are "equivalent" and then derive the optimal multiple testing procedure by solving an equivalent weighted classification problem.

5.2 Multiple Testing and Weighted Classification

The goal of both multiple testing and weighted classification problems is to separate the nonnull cases from the null cases, and the solution to both problems can be represented by a decision rule of the form

$$\boldsymbol{\delta}(\mathbf{T}, c\mathbf{1}) = \{I(T_i < c) : i = 1, \ldots, m\}, \quad (5.3)$$

where $\mathbf{T}$ is a classifier or a test statistic and c is a cutoff. The most popular choice of $\mathbf{T}$ is the vector of p-values. In the multiple testing literature, the following assumption has been used (e.g., Storey 2003; Genovese and Wasserman 2004):

the FDR (FNR) level yielded by $\boldsymbol{\delta}(\mathbf{T}, c\mathbf{1})$ is increasing (decreasing) in c. (5.4)

Assumption (5.4) is desirable for developing FDR procedures since it implies that in order to minimize the FNR, we should choose the largest cutoff c that satisfies FDR$(c) \leq \alpha$. For setwise FDR analysis, we would similarly require that

the FSR (MSR) level yielded by $\boldsymbol{\delta}(\mathbf{T}, c\mathbf{1})$ is increasing (decreasing) in c. (5.5)

Next we develop a general condition that guarantees (5.5). Denote by $G_j(t) = P(T_i < t | \vartheta_i = j), j = 1, 2$, the conditional cumulative distribution function (cdf) of T_i and $g_j(t) = (d/dt)G_j(t)$ the conditional probability distribution function (pdf). Assumes that

$$g_1(t)/g_0(t) \text{ is monotonically decreasing in } t. \quad (5.6)$$

The above condition is referred to as the monotone ration condition (MRC). The next lemma shows that the MRC is a desirable condition for multiple testing.

Lemma 1. The MRC (5.6) implies that (5.5) holds asymptotically.

The class of MRC statistics, denoted by $\mathcal{T}$, includes most commonly used test statistics. For example, let $\mathbf{p}$ be a vector of independent p-values, then $\mathbf{p} \in \mathcal{T}$ if $G_1(p)$, the p-value distribution under the alternative, is concave. The concavity of p-value distribution is a desirable condition that guarantees (5.4) and has been assumed in Genovese and Wasserman (2004) and Storey (2003). In addition, both the local false discovery rate (Lfdr, Efron et al. 2001) statistic and the LIS (Sun and Cai 2009) statistic belong to $\mathcal{T}$. See Sun and Cai (2007) for more discussions of the MRC.

The next theorem, which underlies our theoretical development, states that the multiple testing and weighted classification problems are "equivalent" when the MRC holds (Sun and Cai 2007). Therefore the more complicated multiple testing problem can be solved by studying an equivalent weighted classification problem. The following theorem can be proved similarly as in Sun and Cai (2007).

Theorem 1. Suppose that the classification risk with the loss function defined in (5.2) is minimized by $\delta^\lambda\{\boldsymbol{\Lambda}, c(\lambda)\}$, so that $\boldsymbol{\Lambda}$ is optimal in the weighted classification problem. If $\boldsymbol{\Lambda} \in \mathcal{T}$, then $\boldsymbol{\Lambda}$ is also optimal in the multiple testing problem, in the sense that for each FSR level α, there exists a unique $\lambda(\alpha)$, and hence $c\{\lambda(\alpha)\} = c(\alpha)$, such that $\delta^{\lambda(\alpha)}\{\boldsymbol{\Lambda}, c(\alpha)\}$ controls the FSR at level α with the smallest MSR level among all testing rules in $\mathcal{D}_\alpha$, where $\mathcal{D}_\alpha$ is the collection of all α-level FSR rules of the form $\boldsymbol{\delta} = I(\boldsymbol{\Lambda} < c\mathbf{1})$.

5.3 The Oracle Procedure

Now we derive the optimal test statistic for FSR control. The next theorem considers, in an ideal situation, the optimal classification rule in the HMM defined by (2.3)–(2.5).

Theorem 2. Consider the HMM defined by (2.3)–(2.7). Suppose that the HMM parameters Ψ are known. Then the classification risk with loss function (5.2) is minimized by $\boldsymbol{\delta}\{\boldsymbol{\Lambda}, (1/\lambda)\mathbf{1}\} = (\delta_{ik} : i = 1, \ldots, m)$, where $\Lambda_i = \frac{P_\Psi(\vartheta_i = 0|\mathbf{e}_i)}{P_\Psi(\vartheta_i = 1|\mathbf{e}_i)}$ and $\delta_i = I(\Lambda_i < 1/\lambda)$.

Theorem 2 and the equivalence between multiple testing and weighted classification imply that $\boldsymbol{\Lambda}$ is also the optimal test statistic for FSR control. Define $\text{GLIS}_i = P_\Psi(\vartheta_i = 0|\mathbf{e}_i)$. Note that

$\text{GLIS}_i = \Lambda_i/(1+\Lambda_i)$ is strictly increasing in Λ_i, the (oracle) optimal multiple testing procedure must be of the form

$$\boldsymbol{\delta}(\mathbf{GLIS}, c_{\text{OR}}\mathbf{1}) = [I(\text{GLIS}_i < c_{\text{OR}}) : i = 1, \ldots, m], \qquad (5.7)$$

where the oracle cutoff $c_{\text{OR}} = \sup\{c \in (0,1) : \text{FSR}(\mathbf{GLIS}, c\mathbf{1}) \le \alpha\}$, due to (5.5).

5.4 A Data-Driven Procedure

The optimal testing procedure (5.7) is difficult to implement because it is difficult to determine the optimal cutoff c_{OR} directly. We propose the following procedure that is asymptotically equivalent to (5.7). The derivation of this procedure is given in the Appendix. Let $\text{GLIS}_{(1)}, \ldots, \text{GLIS}_{(m)}$ be the ranked GLIS statistics and $H_{(1)}, \ldots, H_{(m)}$ be corresponding hypotheses.

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \text{GLIS}_{(j)} \le \alpha \right\},$$
$$\text{then reject all } H_{(i)}, i = 1, \ldots, k. \qquad (5.8)$$

The next theorem shows that the GLIS procedure (5.8) is asymptotically equivalent to the oracle procedure (5.7); in particular, it is valid for FSR control.

Theorem 3. Consider the HMM defined by (2.3)–(2.7). Let $\text{GLIS}_i = P_\Psi(\vartheta_i = 0|\mathbf{e}_i)$. Denote by $\text{GLIS}_{(1)}, \ldots, \text{GLIS}_{(m)}$ the ranked GLIS values, and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. Then the GLIS procedure (5.8) controls the FSR at level α. In addition, let MSR_{OR} and MSR_{GLIS} be the MSR levels of the oracle procedure (5.7) and the GLIS procedure (5.8), respectively, then $\text{MSR}_{\text{GLIS}} = \text{MSR}_{\text{OR}} + o(1)$.

Denote by $\widehat{\text{GLIS}}_{(1)}, \ldots, \widehat{\text{GLIS}}_{(m)}$ the ranked estimates and $H_{(1)}, \ldots, H_{(m)}$ the corresponding hypotheses. In light of the GLIS procedure (5.8), we propose the following data-driven procedure:

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\text{GLIS}}_{(j)} \le \alpha \right\},$$
$$\text{then reject all } H_{(i)}, i = 1, \ldots, k. \qquad (5.9)$$

The next theorem shows that the data-driven procedure (5.9) attains the performance of the oracle procedure (5.7) asymptotically.

Theorem 4. Consider the HMM defined by (2.3)–(2.7). Let $\hat{\Psi}$ be an estimate of the HMM parameters Ψ such that $\hat{\Psi} \xrightarrow{p} \Psi$. Let FSR_{OR} and FSR_{DD} be the FSR levels of the oracle procedure (5.7) and data-driven procedure (5.9), respectively, and MSR_{OR} and MSR_{DD} the corresponding MSR levels. Then $\text{FSR}_{\text{DD}} = \alpha + o(1)$ and $\text{MSR}_{\text{DD}} = \text{MSR}_{\text{OR}} + o(1)$.

APPENDIX: PROOFS AND DERIVATIONS OF TECHNICAL RESULTS

Proof of Lemma 1

Let $\rho = P(\vartheta_i = 0)$ be the proportion of nonnull sets. The marginal cdf of T_i is $G = (1-\rho)G_0 + \rho G_1$, where G_0 and G_1 are conditional cdf's defined in Section 5.2. Define the marginal FSR (mFSR) and marginal MSR (mMSR) as

$$\text{mFSR} = \frac{E\{\sum_{i=1}^m (1-\vartheta_i)\delta_i\}}{E(\sum_{i=1}^m \delta_i)} \quad \text{and}$$
$$\text{mMSR} = \frac{E\{\sum_{i=1}^m \vartheta_i(1-\delta_i)\}}{E(\sum_{i=1}^m \vartheta_i)}.$$

Then mFSR and mMSR are asymptotically equivalent measures to the FSR and MSR in the sense that, under mild conditions, $\text{mFSR} = \text{FSR} + O(m^{-1/2})$ and $\text{mMSR} = \text{MSR} + O(m^{-1/2})$. It is easy to show that $\text{mFSR} = (1-\rho)G_0(t)/G(t)$ and $\text{mMSR} = 1 - G_1(t)$. Obviously the mMSR is decreasing in t. Next note that the MRC implies that $g_0 G_1 > g_1 G_0$, we have

$$(d/dt)\,\text{mFSR}(t) = \{\rho(1-\rho)(g_0 G_1 - g_1 G_0)\}/G^2(t) > 0.$$

Hence the mFSR is increasing in t. The results follow by noting that $\text{mFSR} = \text{FSR} + o(m^{-1/2})$ and $\text{mMSR} = \text{MSR} + o(m^{-1/2})$.

Proof of Theorem 2

The posterior distribution of $\boldsymbol{\vartheta}$ given $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_m)$ is

$$P_{\boldsymbol{\vartheta}|\mathbf{e}}(\boldsymbol{\vartheta}|\mathbf{e}) = \prod_{i=1}^m \{(1-\vartheta_i)P(\vartheta_i = 0|\mathbf{e}_i) + \vartheta_i P(\vartheta_i = 1|\mathbf{e}_i)\}.$$

Hence the posterior risk is

$$\begin{aligned} &E_{\boldsymbol{\vartheta}|\mathbf{e}}\{L_\lambda(\boldsymbol{\vartheta}, \boldsymbol{\delta})\} \\ &= m^{-1}\sum_i E_{\vartheta_i|\mathbf{e}_i}\{\lambda(1-\vartheta_i)\delta_i + \vartheta_i(1-\delta_i)\} \\ &= m^{-1}\sum_i \{\lambda\delta_i P(\vartheta_i = 0|\mathbf{e}_i) + (1-\delta_i)P(\vartheta_i = 1|\mathbf{e}_i)\} \\ &= m^{-1}\sum_i P(\vartheta_i = 1|\mathbf{e}_i) \\ &\quad + m^{-1}\sum_i \{\lambda P(\vartheta_i = 0|\mathbf{e}_i) - P(\vartheta_i = 1|\mathbf{e}_i)\}\delta_i. \end{aligned}$$

Therefore the classification risk is minimized by $\delta_i = I\{\lambda P(\vartheta_i = 0|\mathbf{e}_i) < P(\vartheta_i = 1|\mathbf{e}_i)\}$.

Proof of Theorem 3

(i) *Validity*. First according to the definition of GLIS, we have

$$E\{(1-\vartheta_i)\delta_i|\mathbf{e}_i\} = \delta_i P(\vartheta_i = 0|\mathbf{e}_i) = \delta_i \text{GLIS}_i.$$

Suppose the total number of rejections at FSR level α is R_α, then $R_\alpha = \sum_i \delta_i$. The actual FSR level by the GLIS procedure (5.8) is

$$\begin{aligned} \text{FSR}_{\text{GLIS}} &= E\left\{ \frac{\sum_i (1-\vartheta_i)\delta_i}{(\sum_i \delta_i) \vee 1} \right\} \\ &= E\left[\frac{1}{(\sum_i \delta_i) \vee 1} \sum_i E\{(1-\vartheta_i)\delta_i|\mathbf{e}_i\} \right] \\ &\le E\left(\frac{1}{R_\alpha} \sum_{i=1}^{R_\alpha} \text{GLIS}_{(i)} \right). \end{aligned}$$

The validity of the GLIS procedure follows by noting that (5.8) guarantees that, for all realizations of $\mathbf{e}$, $(1/R_\alpha)\sum_{i=1}^{R_\alpha} \text{GLIS}_{(i)} \le \alpha$.

(ii) *Asymptotic optimality*. Note that $\text{mMSR} = \text{MSR} + o(1)$, it is sufficient to show that $\text{mMSR}_{\text{GLIS}} = \text{mMSR}_{\text{OR}} + o(1)$. Let c_{OR} and $\hat{c}_{\text{OR}}$ be the thresholds of the oracle and PLIS procedures, respectively. Then $\text{mMSR}_{\text{OR}} = P(\text{GLIS} > c_{\text{OR}}|\vartheta_i = 1)$ and $\text{mMSR}_{\text{GLIS}} = P(\text{GLIS} > \hat{c}_{\text{OR}}|\vartheta_i = 1)$. The continuous mapping theorem implies that we only need to show that $\hat{c}_{\text{OR}} \xrightarrow{p} c_{\text{OR}}$. Denote by G_0 and G_1 the

null and nonnull cdf's of GLIS_i. The mFSR level of the oracle procedure for a given cutoff t is $\text{mFSR}_{\text{OR}}(t) = (1-\rho)G_0(t)/G(t)$, where $G = (1-\rho)G_0 + \rho G_1$. Next define

$$\hat{Q}(t) = \left\{\sum_i I(\text{GLIS}_i < t)\text{GLIS}_i\right\} \Big/ \left\{\sum_i I(\text{GLIS}_i < t)\right\}.$$

According to the law of large numbers,

$$m^{-1}\sum_i I(\text{GLIS}_i < t)\text{GLIS}_i \xrightarrow{P} E\{I(\text{GLIS}_i < t)\text{GLIS}_i\} = (1-\rho)G_0(t).$$

Similarly we can show that $m^{-1}\{\sum_i I(\text{GLIS}_i < t)\} \xrightarrow{P} G(t)$. Therefore $\hat{Q}(t) \xrightarrow{p} \text{mFSR}(t)$. Observe that $\hat{Q}(t)$ is a constant in the interval $\text{GLIS}_{(i)} \le t < \text{GLIS}_{(i+1)}$, we have

$$\begin{aligned}\hat{c}_{\text{OR}} &= \max_{i=1,\ldots,m}\left\{\text{GLIS}_{(i)} : \frac{1}{i}\sum_{j=1}^{i}\text{GLIS}_{(j)} \le \alpha\right\}\\ &= \max_{i=1,\ldots,m}\left\{\text{GLIS}_{(i)} : \hat{Q}\big(\text{GLIS}_{(i)}\big) \le \alpha\right\}\\ &= \sup\{c \in (0,1) : \hat{Q}(c) \le \alpha\}\\ &\equiv \hat{Q}^{-1}(\alpha).\end{aligned}$$

We have shown that $\hat{Q}(t) \xrightarrow{p} \text{mFSR}(t)$. It follows from functional delta method that $\hat{c}_{\text{OR}} \xrightarrow{P} c_{\text{OR}}$. Therefore the GLIS procedure attains the MSR level of the oracle procedure asymptotically.

Proof of Theorem 4

Define

$$\hat{Q}_{\text{DD}}(t) = \left\{\sum_i I(\widehat{\text{GLIS}}_i < t)\widehat{\text{GLIS}}_i\right\} \Big/ \left\{\sum_i I(\widehat{\text{GLIS}}_i < t)\right\}.$$

The α-level FSR cutoff of the data-driven procedure is $\hat{c}_{\text{DD}} = \sup\{t \in (0,1) : \hat{Q}_{\text{DD}}(t) \le \alpha\}$. Note that $\hat{\Psi} \xrightarrow{p} \Psi$. Then by continuous mapping theorem, we have $\widehat{\text{GLIS}}_i \xrightarrow{p} \text{GLIS}_i$. Applying the weak law of large numbers for triangular arrays we can show that $\hat{Q}_{\text{DD}}(t) \xrightarrow{p} \text{mFSR}_{\text{OR}}(t)$. By using similar arguments in Theorem 2, we can show that $\hat{c}_{\text{DD}} \xrightarrow{p} c_{\text{OR}}$. Therefore we have

$$\begin{aligned}\text{FSR}_{\text{DD}} &= \text{mFSR}_{\text{DD}} + o(1)\\ &= \frac{(1-\rho)P(\widehat{GLIS}_i < \hat{c}_{\text{DD}} | \vartheta_i = 0)}{P(\widehat{\text{GLIS}}_i < \hat{c}_{\text{DD}})} + o(1)\\ &\xrightarrow{p} \frac{(1-\rho)P(\text{GLIS}_i < c_{\text{OR}} | \vartheta_i = 0)}{P(\text{GLIS}_i < c_{\text{OR}})}\\ &= \text{mFSR}_{\text{OR}}(c_{\text{OR}}) = \alpha.\end{aligned}$$

Similarly we can show that $\text{MSR}_{\text{DD}} = \text{MSR}_{\text{OR}} + o(1)$.

Derivation of the GLIS procedure (5.8)

Note that FSR and mFSR are asymptotically equivalent measures, we shall derive the GLIS procedure using mFSR. Let $\text{GLIS}_{(1)}, \ldots, \text{GLIS}_{(m)}$ be the ranked test statistics and $H_{(1)}, \ldots, H_{(m)}$ be corresponding hypotheses. Let ρ be the proportion of the nonnull sets, G_0 and G_1 be the conditional cdf of GLIS under the null and nonnull hypotheses, respectively. The mFSR level of $\boldsymbol{\delta}(\mathbf{GLIS}, t\mathbf{1})$ is

$$\begin{aligned}\text{mFSR} &= \frac{(1-\rho)G_0(t)}{G(t)}\\ &= \frac{E\{(1-\vartheta_i)I(\text{GLIS}_i < t)\}}{E\{I(\text{GLIS}_i < t)\}}\\ &= \frac{E\{\text{GLIS}_i I(\text{GLIS}_i < t)\}}{E\{I(\text{GLIS}_i < t)\}}\\ &= \frac{\sum_{i=1}^m \text{GLIS}_i I(\text{GLIS}_i < t)}{\sum_{i=1}^m I(\text{GLIS}_i < t)} + o(1).\end{aligned}$$

Let $\text{GLIS}_{(k)}$ be the largest test statistic that is less than t, then the mFSR level for a given cutoff t can be approximated by $\hat{Q}(k)(1/k) \times \sum_{i=1}^k \text{GLIS}_{(i)}$. Also note that $\hat{Q}(k)$ is increasing in k, we shall choose the largest k such that the mFSR is controlled at level α. Then by noting assumption (5.5), the procedure (5.8) follows.

[Received September 2009. Revised July 2010.]

REFERENCES

Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002), "Gene Expression During the Life Cycle of Drosophila Melanogaster," *Science*, 297 (5590), 2270–2275. [73]

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970), "A Maximization Technique Occuring in the Statistical Analysis of Probablistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171. [77]

Benjamini, Y., and Heller, R. (2008), "Screening for Partial Conjunction Hypotheses," *Biometrics*, 64 (4), 1215–1222. [74-76,78]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57 (1), 289–300. [74,76-78]

Benjamini, Y., and Yekutieli, D. (2005), "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters," *Journal of the American Statistical Society*, 100, 71–93. [77]

Bickel, P. J., Ritov, Y., and Rydén, T. (1998), "Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models," *The Annals of Statistics*, 26 (4), 1614–1635. [77]

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F., and Inflamm and Host Respanse to Injury Large Scale Collab. Res. Program (2005), "A Network-Based Analysis of Systemic Inflammation in Humans," *Nature*, 437 (7061), 1032–1037. [73,80-82]

Chi, Y.-Y., Ibrahim, J. G., Bissahoyo, A., and Threadgill, D. W. (2007), "Bayesian Hierarchical Modeling for Time Course Microarray Experiments," *Biometrics*, 63 (2), 496–504. [74]

Churchill, G. A. (1992), "Hidden Markov Chains and the Analysis of Genome Structure," *Computers & Chemistry* 16, 107–115. [74]

Du, P., Kibbe, W. A., and Lin, S. M. (2008), "lumi: A Pipeline for Processing Illumina Microarray," *Bioinformatics*, 24 (13), 1547–1548. [81]

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge, U.K.: Cambridge University Press. [74]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Society*, 96 (456), 1151–1160. [85]

Finner, H. (1999), "Stepwise Multiple Test Procedures and Control of Directional Errors," *The Annals of Statistics*, 27, 274–289. [77]

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004), "affy—Analysis of Affymetrix Genechip Data at the Probe Level," *Bioinformatics*, 20 (3), 307–315. [81]

Genovese, C., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, Ser. B*, 64 (3), 499–517. [76]

——— (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32 (3), 1035–1061. [85]

Guo, W., Sarkar, S., and Peddada, S. (2010), "Controlling False Discoveries in Multidimensional Directional Decisions, With Applications to Gene Expression Data on Ordered Categories," *Biometrics*, 66 (2), 485–492. [77,78]

Guo, X., Qi, H., Verfaillie, C. M., and Pan, W. (2003), "Statistical Significance Analysis of Longitudinal Gene Expression Data," *Bioinformatics*, 19 (13), 1628–1635. [73]

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006), "Cluster-Based Analysis of fmri Data," *Neuroimage*, 33 (2), 599–608. [74]

Hong, F., and Li, H. (2006), "Functional Hierarchical Models for Identifying Genes With Different Time-Course Expression Profiles, *Biometrics*, 62 (2), 534–544. [74]

Hubbell, E., Liu, W.-M., and Mei, R. (2002), "Robust Estimators for Expression Analysis," *Bioinformatics*, 18 (12), 1585–1592. [81]

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), "Exploration, Normalization, and Sum-

maries of High Density Oligonucleotide Array Probe Level Data," *Biostatistics*, 4 (2), 249–264. [81]

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003), "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles," *Statistics in Medicine*, 22 (24), 3899–3914. [74,76]

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994), "Hidden Markov Models in Computational Biology. Applications to Protein Modeling," *Journal of Molecular Biology*, 235 (5), 1501–1531. [74]

Leroux, B. G. (1992), "Maximum-Likelihood Estimation for Hidden Markov Models," *Stochastic Processes and Their Applications*, 40 (1), 127–143. [77]

Li, C., and Wong, W. H. (2001), "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proceedings of the National Academy of Sciences of the USA*, 98 (1), 31–36. [81]

Luan, Y., and Li, H. (2004), "Model-Based Methods for Identifying Periodically Expressed Genes Based on Time Course Microarray Gene Expression Data," *Bioinformatics*, 20 (3), 332–339. [74]

Ma, P., Zhong, W., and Liu, J. S. (2009), "Identifying Differentially Expressed Genes in Time Course Microarray Data," *Statistics in Biosciences*, 1, 144–159. [73,77]

MacDonald, I. L., and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, New York: Chapman & Hall. [74]

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8 (6), 37–52. [74,76]

Park, T., Yi, S.-G., Lee, S., Lee, S. Y., Yoo, D.-H., Ahn, J.-I., and Lee, Y.-S. (2003), "Statistical Tests for Identifying Differentially Expressed Genes in Time-Course Microarray Experiments," *Bioinformatics*, 19 (6), 694–703. [73]

Pyne, S., Futcher, B., and Skiena, S. (2006), "Meta-Analysis Based on Control of False Discovery Rate: Combining Yeast Chip-Chip Datasets," *Bioinformatics*, 22 (20), 2516–2522. [74]

Rabiner, L. R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–286. [74,77]

Robbins, H. (1951), "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, Berkeley and Los Angeles: University of California Press, pp. 131–148. [83]

Schliep, A., Schonhuth, A., and Steinhoff, C. (2003), "Using Hidden Markov Models to Analyze Gene Expression Time Course Data," *Bioinformatics*, 19 (Suppl 1), i255–i263. [74,76]

Smyth, G. K., and Speed, T. (2003), "Normalization of cDNA Microarray Data," *Methods*, 31 (4), 265–273. [81]

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, 9 (12), 3273–3297. [73]

Storey, J. D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 31 (6), 2013–2035. [85]

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005), "Significance Analysis of Time Course Microarray Experiments," *Proceedings of the National Academy of Sciences of the USA*, 102 (36), 12837–12842. [74]

Sun, W., and Cai, T. T. (2007), "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of the American Statistical Society*, 102 (479), 901–912. [85]

——— (2009), "Large-Scale Multiple Testing Under Dependence," *Journal of the Royal Statistical Society, Ser. B*, 71 (2), 393–424. [74,77,85]

Tai, Y. C., and Speed, T. P. (2006), "A Multivariate Empirical Bayes Statistic for Replicated Microarray Time Course Data," *The Annals of Statistics*, 34 (5), 2387–2412. [73]

Telesca, D., Inoue, L. Y. T., Neira, M., Etzioni, R., Gleave, M., and Nelson, C. (2009), "Differential Expression and Network Inferences Through Functional Data Modeling," *Biometrics*, 65 (3), 793–804. [74]

Tian, B., Nowak, D. E., and Brasier, A. R. (2005), "A TNF-Induced Gene Expression Program Under Oscillatory NF-Kappab Control," *BMC Genomics*, 6, 137. [73]

Viterbi, A. J. (1967), "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," *IEEE Transactions on Information Theory*, 13, 268–278. [74]

Yuan, M., and Kendziorski, C. (2006), "Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions," *Journal of the American Statistical Society*, 101 (476), 1323–1332. [73,74,76,77,81,84]

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002), "Truncated Product Method for Combining p-Values," *Genetic Epidemiology*, 22 (2), 170–185. [74]