Inference on the limiting false discovery rate and the p-value threshold parameter assuming weak dependence between gene expression levels within subject

Glenn Heller[1] **and Jing Qin**[2]

[1]Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
New York, NY 10021, USA

[2]Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
Bethesda, Maryland 20892, USA

Corresponding author: Glenn Heller
email: hellerg@mskcc.org
phone: 646 735 8112
fax: 646 735 0010

Running Head: FDR analysis with dependent data

Summary. An objective of microarray data analysis is to identify gene expressions that are associated with a disease related outcome. For each gene, a test statistic is computed to determine if an association exists, and this statistic generates a marginal p-value. In an effort to pool this information across genes, a p-value density function is derived. The p-value density is modeled as a mixture of a uniform (0,1) density and a scaled ratio of normal densities derived from the asymptotic normality of the test statistic. The p-values are assumed to be weakly dependent and a quasi-likelihood is used to estimate the parameters in the mixture density. The quasi-likelihood and the weak dependence assumption enables estimation and asymptotic inference on the false discovery rate for a given rejection region, and its inverse, the p-value threshold parameter for a fixed false discovery rate. A false discovery rate analysis on a localized prostate cancer data set is used to illustrate the methodology. Simulations are performed to assess the performance of this methodology.

Keywords: Asymptotic normal test statistic, confidence interval, microarray, p-value mixture model, quasi-likelihood, weak dependence.

1 Introduction

Microarray analysis is used to identify gene expressions that are associated with a disease related outcome. It is typically exploratory, with no apriori hypothesis concerning the association between a specific gene expression and the outcome variable. The intent of the analysis is to generate hypotheses for further exploration, either in the laboratory or in the clinic. Microarray technology is currently applied in many areas including: clinical staging, cell line classification, distinguishing tumor type, and understanding the effect of a biological agent. Scientists believe that identification of informative genes will provide insight into a disease mechanism, a genetic pathway, or isolation of a therapeutic target.

Microarray analysis generates a vast amount of data. In a typical study, tens of thousands of gene expressions are recorded on the subjects under study. For each gene, a test of association of gene expression and outcome variable is performed. The statistical challenge is how to determine which genes are truly associated with outcome. Simply testing each gene individually, without adjustment for the number of genes examined, provides little confidence that true associations are identified and nonimportant genes are eliminated from further study. For example, if a test statistic is computed for each gene, and the genes with the highest test statistic are found discriminatory, the cut-off for the test statistic is still problematic. Due to the thousands of tests performed, use of a standard nominal significance level to determine this critical region will result in overstating the number of significant associations identified.

Traditionally, protection against multiple comparisons is undertaken by choosing the critical region of a test to satisfy a familywise error rate of α, where the familywise error rate is defined as the probability of rejecting at least one true hypothesis (Hochberg and Tamhane, 1987). While protecting against falsely rejecting tests, the familywise error rate is a conservative approach, resulting in a loss of power in each of the individual tests. As a result, using this approach to adjust for the multiple tests could potentially miss genes that are associated with outcome. To correct for this conservativeness, Benjamini and Hochberg (1995) developed the false discovery rate (FDR), which is defined as the expected proportion of false rejections of the null hypothesis. The FDR represents a compromise between the conservative familywise error rate and testing each gene at the nominal significance level.

Variations of the false discovery rate, termed the positive, conditional, and marginal false discovery rates, have been proposed in the literature (Benjamini and Hochberg 1995, Storey 2002, Tsai 2003). Assuming weak dependence and at least one test statistic is rejected, as the number of genes tested increases, these false discovery rates all converge to the probability a gene is not associated with the outcome conditional on the test statistic lying in the rejection region. As a result of the asymptotic equivalence of these false discovery rates, we call this conditional probability the limiting false discovery rate. For the analysis of gene array data, where tens of thousands of tests are carried out, this asymptotic evaluation of the FDR is reasonable.

The limiting FDR has been estimated in the literature by pooling information across genes and using a mixture model for the density of the test statistic or corresponding p-value. Pan et al. (2003) employ a normal mixture model for the density of the t-statistic. Parker and Rothenberg (1988), Allison et al. (2002), and Pounds and Morris (2003) use a Uniform-Beta mixture to model the p-value density. The accuracy of the resulting FDR estimates rely on the adequacy of the mixture density specification. To avoid this specification, nonparametric estimates of the mixture density have been proposed by Efron et al. (2001), Storey (2002), and Black (2004). In addition to estimation of the limiting FDR, estimates of the p-value threshold, adapted from the sequential p-value method of Benjamini and Hochberg (1995), have been developed by Benjamini and Hochberg (2000), Genovese and Wasserman (2004), and Storey et al. (2004). Although these FDR and p-value threshold estimates are commonly employed in the analysis of microarray data, their precision is typically ignored. One exception is Owen (2005), who computed the variance of the number of false discoveries when genes are dependent, but this calculation is conditional on the observed gene expression data and assumes all genes are unrelated to the outcome variable.

In this paper, we develop estimates of the limiting false discovery rate evaluated at a p-value threshold, and its inverse, the p-value threshold evaluated at a fixed FDR, from a quasi-likelihood derived from marginal p-value mixture densities. The asymptotic normality of the FDR and p-value threshold estimates stem from quasi-likelihood based results and a weak dependence assumption between gene expression values within subject. Estimation of the asymptotic variance for the limiting FDR estimate is derived and confidence intervals for the limiting FDR and p-value threshold parameters are developed, accounting for the potential dependence between genes. We believe these estimates of precision provide a unique perspective to error rate analysis of microarray data.

The methodology is demonstrated on a microarray gene expression data set obtained from 79 patients who underwent a radical prostatectomy for localized prostate cancer. The data were obtained from tissue samples taken at the time of surgery. In the analysis, patients were followed for at least seven years; 37 patients were classified with recurrent disease based on a rising PSA profile, whereas 42 patients classified with nonrecurrent disease, remained with an undetectable PSA seven years after surgery (Stephenson et al. 2005). Prostate specific antigen (PSA) is a biomarker that is commonly used to determine the existence of prostate tumor cells in the patient. The gene expression analysis was carried out using the Affymetrix U133A human gene array, which has 22,283 genes. Expression values on each array were preprocessed using Affymetrix MAS 5.0. This preprocessing algorithm includes a background adjustment of the expression values, and a within array scale transformation, producing a 2% trimmed mean within each array equal to 500. The choice of 500 is the default value for MAS 5.0.

2 P-value Mixture Model

To determine differential gene expression between the recurrent/nonrecurrent outcomes, a t-test was performed for each gene, and the accompanying p-value, based on the standard normal reference distribution, was computed to test the hypothesis

H_{0g}: no difference in gene g expression between outcome groups

H_{1g}: gene g expression is different between outcome groups ($g = 1, \ldots G$).

Each p-value is generated from one of these two classes (not different/different). A random variable D_g, indicates whether the observed p-value for gene g, denoted by p_g, was generated from the null class ($D_g = 0$) or the alternative class ($D_g = 1$). The marginal distribution of D_g is Bernoulli with parameter $\lambda = Pr(\text{not different})$ and the density of P given D is written as $f_D(p)$. Since the D_g are not observed, the marginal density of P is represented as a mixture of two density functions

$$f(p) = \lambda f_0(p) + (1 - \lambda) f_1(p).$$

In the null class, the distribution of P is uniform (0,1),

$$f_0(p) = 1 \qquad 0 < p < 1.$$

The alternative density is

$$f_1(p; \tau) = \frac{1}{2} \frac{\phi_{\tau,1}(\Phi^{-1}(1 - p/2))}{\phi_{0,1}(\Phi^{-1}(1 - p/2))},$$

where $\phi_{\mu,\sigma}(u)$ denotes a normal density function with location and scale parameters μ and σ, and Φ is the standard normal distribution function. This density is derived from the asymptotic normality of the test statistic T and the change of variable $P = 2(1 - \Phi(|T|))$ (Hung et al. 1997).

Estimation and inference in this work are based on the marginal p-values derived from the asymptotic normality of the test statistic. The accuracy of the inference derived from the proposed methodology is a function of the accuracy of the asymptotic normal approximation and thus improves as the sample size increases. Although Student's t-statistic is used to generate the p-values for the prostate cancer gene expression data, the application is wide ranging and can be applied to any k-sample comparison or test of association that is based on an asymptotic normal test statistic.

The mixture model used to represent the density of the p-values is

$$f(p; \lambda, \tau) = \lambda + (1 - \lambda) \frac{1}{2} \frac{\phi_{\tau,1}(\Phi^{-1}(1 - p/2))}{\phi_{0,1}(\Phi^{-1}(1 - p/2))} \qquad 0 < p \leq 1. \tag{1}$$

In this model, the parameter τ measures the strength of the differentially expressed genes, with a large value signaling that there are a group of genes with very small p-values. For the two-sample t-test, assuming a common variance σ^2, and n_1, n_2 subjects in the two groups,

$$\tau = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \left(\frac{|\mu_1 - \mu_2|}{\sigma}\right).$$

Estimation of the parameters $\boldsymbol{\beta} = (\lambda, \tau)$ is based on the loglikelihood

$$L_G(\lambda, \tau) \equiv \sum_{g=1}^{G} l_g(\boldsymbol{\beta}) = \sum_{g=1}^{G} \log \left\{ \lambda + (1-\lambda) \frac{1}{2} \frac{\phi_{\tau,1}(\Phi^{-1}(1-p_g/2))}{\phi_{0,1}(\Phi^{-1}(1-p_g/2))} \right\}. \quad (2)$$

In the context of microarray analysis, it is not plausible to treat the G gene derived p-values as independent. We therefore treat $L_G(\lambda, \tau)$ as a log quasi-likelihood. A form of weak dependence between the quasi-score components, described in conditions (D1)-(D3) below, is sufficient to satisfy the central limit theorem and the weak law of large numbers for the parameter estimates (Serfling 1968). As a result, the theorem stated following these conditions provides the asymptotic inferential structure for $\boldsymbol{\beta}$.

Denote the quasi-score component as

$$s_g(\boldsymbol{\beta}) = \frac{\partial l_g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

$\mathcal{M}^a$ as a σ-algebra generated by the quasi-score components $\{s_1(\boldsymbol{\beta}), \ldots, s_a(\boldsymbol{\beta})\}$, and the partial sum of the quasi-score as

$$T_a(\boldsymbol{\beta}) = G^{-1/2} \sum_{g=a+1}^{a+G} s_g(\boldsymbol{\beta}).$$

The following conditions are used to define weak dependence of the quasi-score components

$(D1)$ $E[s_g(\boldsymbol{\beta}_0)] = 0$

$(D2)$ $\lim_{G\to\infty} \mathrm{var}[T_a(\boldsymbol{\beta}_0)] = V(\boldsymbol{\beta}_0)$ uniformly in a

$(D3)$ $E[E^2(T_a(\boldsymbol{\beta}_0)|\mathcal{M}^a\ )] \leq b(G)$

$E|\mathrm{var}(T_a(\boldsymbol{\beta}_0)|\mathcal{M}^a) - \mathrm{var}(T_a(\boldsymbol{\beta}_0))| \leq b(G)$

where $b(G) = O(G^{-\theta});\ \theta > 0$.

THEOREM: Assume that the weak dependence conditions (D1-D3) are satisfied. Define $\hat{\boldsymbol{\beta}} = (\hat{\lambda}, \hat{\tau})$ as the maximum quasi-likelihood estimate derived from equation (2) and let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$. Let $\mathcal{N}(\boldsymbol{\beta}_0)$ denote a bounded neighborhood around $\boldsymbol{\beta}_0$ and assume $G^{-1}L_G(\boldsymbol{\beta})$ and its derivative are uniformly bounded in $\mathcal{N}(\boldsymbol{\beta}_0)$. Then as $n \to \infty$ and $G \to \infty$,

(1) $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$,

(2) $\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \Sigma), \qquad \Sigma = A^{-1}VA^{-1}$

where

$$\frac{1}{\sqrt{G}} \sum_{g=1}^{G} \frac{\partial l_g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \xrightarrow{D} N(0, V)$$

$$\frac{1}{G} \sum_{g=1}^{G} \frac{\partial^2 l_g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \xrightarrow{P} A \qquad\qquad \frac{1}{G} \sum\sum_{g,h \in \mathcal{C}} \frac{\partial l_g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \frac{\partial l_h(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \xrightarrow{P} V$$

and $\mathcal{C}$ is the set of quasi-score component pairs with nonzero correlation (Lumley and Heagerty 1999). The matrices A and V are consistently estimated by

$$A_n(\hat{\boldsymbol{\beta}}) = \frac{1}{G} \sum_{g=1}^{G} \left. \frac{\partial^2 l_g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \qquad V_n(\hat{\boldsymbol{\beta}}) = \frac{1}{G} \sum\sum_{g,h \in \mathcal{C}} \left. \frac{\partial l_g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial l_h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}$$

Estimation of V requires knowledge of the correlated quasi-score component pairs. The following argument is used to carry out this computation. Let

$$t(\mathbf{e}_g) = \Phi^{-1}(1 - p_g/2)$$

represent the observed value of the t-statistic for gene g as a function of the gene expression data for the n subjects, where $\mathbf{e}_g^T = (e_{1g}, \ldots, e_{ng})$ is the gene expression data vector. These n elements are derived from the two outcome groups (recurrence / no recurrence). It is assumed that the elements are generated independently, and up to a shift in location, identically distributed. The quasi-score for gene g is now written as

$$s(\boldsymbol{\beta}; \mathbf{e}_g) = \frac{\partial}{\partial \boldsymbol{\beta}} \log \left\{ \lambda + (1-\lambda) \frac{1}{2} \frac{\phi_{\tau,1}(t(\mathbf{e}_g))}{\phi_{0,1}(t(\mathbf{e}_g))} \right\}$$

where we have added the gene expression vector as an argument to the quasi-score component. Application of the mean value theorem produces the following covariance calculation for the quasi-score components corresponding to genes (g, h)

$$\begin{aligned} &\operatorname{cov}[s(\boldsymbol{\beta}; \mathbf{e}_g), s(\boldsymbol{\beta}; \mathbf{e}_h)] = \\ &\quad \operatorname{cov}\left[\{s(\boldsymbol{\beta}; \boldsymbol{\mu}_g) + W^T(\boldsymbol{\beta}; \boldsymbol{\mu}_g^*)(\mathbf{e}_g - \boldsymbol{\mu}_g)\}, \{s(\boldsymbol{\beta}; \boldsymbol{\mu}_h) + W^T(\boldsymbol{\beta}; \boldsymbol{\mu}_h^*)(\mathbf{e}_h - \boldsymbol{\mu}_h)\} \right], \end{aligned}$$

where $E(\mathbf{e}_g) = \boldsymbol{\mu}_g$, and $W(\boldsymbol{\beta}; \boldsymbol{\mu}_g^*) = \partial s(\boldsymbol{\beta}; \mathbf{e}_g)/\partial \mathbf{e}_g$ is an $n \times 2$ matrix evaluated at the point $\boldsymbol{\mu}_g^*$ which lies on a line between $\mathbf{e}_g$ and $\boldsymbol{\mu}_g$. Letting σ_{gh} denote the covariance of the gene expression data for genes g and h, it follows that

$$\operatorname{cov}[s(\boldsymbol{\beta}; \mathbf{e}_g), s(\boldsymbol{\beta}; \mathbf{e}_h)] = \sigma_{gh} W^T(\boldsymbol{\beta}; \boldsymbol{\mu}_g^*) W(\boldsymbol{\beta}; \boldsymbol{\mu}_h^*).$$

Thus, elimination of noncorrelated quasi-score component pairs can be accomplished by testing whether the corresponding gene expression data pairs are correlated.

For the prostate cancer data, the sample correlation matrix $R = (r)_{gh}$ was computed for the 22,283 genes and Fisher's z-test statistic was used to test whether each gene pair had correlation zero. There were $G(G-1)/2 \approx$ 250 million tests to determine correlated gene pairs, resulting in a further multiple comparison problem. Using the Benjamini and Hochberg (1995) procedure of controlling the false discovery rate at the 0.05 level, gene pairs were considered correlated if the test statistic produced a p-value less than 0.003. Seven percent of the gene pairs demonstrated a nonzero correlation for the expression data using this FDR criterion and were included in the summand for the estimate $V_n(\hat{\boldsymbol{\beta}})$.

Maximization of the quasi-likelihood was accomplished through the Nelder-Mead simplex algorithm, under the constrained parameter space $(0 < \lambda < 1,\ 0 < \tau)$. The parameter estimates from the prostate cancer data were, $\hat{\lambda} =$

0.75, $\hat{\tau} = 1.89$. To assess the adequacy of the mixture model (1), the observed p-values were compared to the model derived p-values. As shown in Figure 1, the mixture model along with the maximum quasi-likelihood parameter estimates provide an adequate fit to the data.

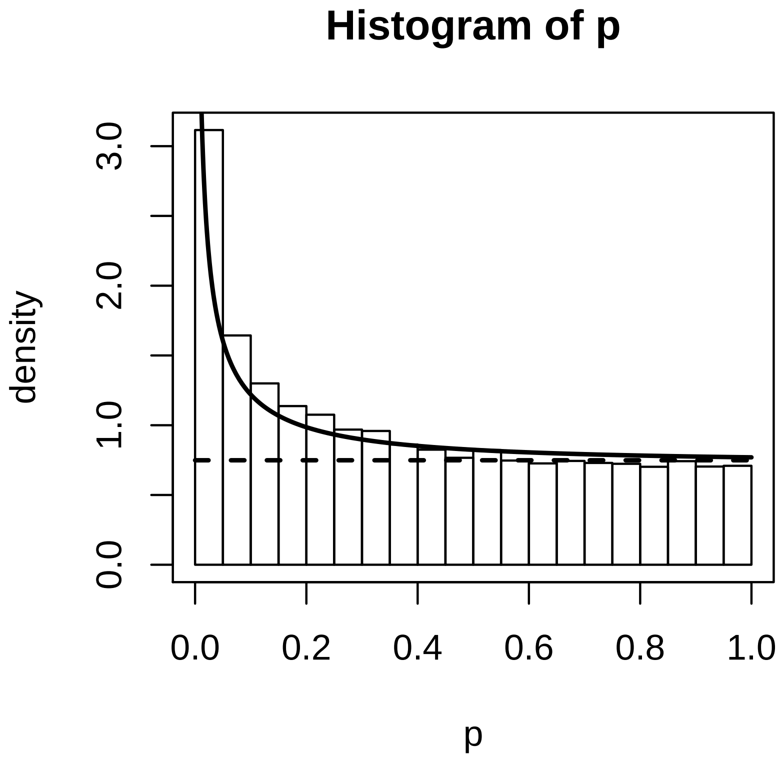

Figure 1. Histogram for the observed 22,283 gene p-values and the p-value mixture density estimate. The horizontal dotted line represents the quasi-likelihood estimate of λ.

The mixture parameter estimate indicates that 25% of the 22,283 genes analyzed are associated with recurrence. It is generally recognized that the level of confidence regarding membership into this alternative class is not equal for all genes. Gene membership into the alternative class becomes increasingly likely as the p-value decreases. Our strategy is to report those genes where there is a high level of confidence of this association.

3 Gene Selection based on the FDR and P-Value Threshold Parameters

The false discovery rate (FDR) is a popular measure of this confidence level. Assuming a common marginal distribution for each p-value, the limiting FDR is defined at a fixed rejection region γ_0 by

$$\pi(\gamma_0) = Pr(D = 0|P \leq \gamma_0);$$

the probability a gene belongs to the null class given its associated p-value is less than γ_0 (Storey 2004, Genovese and Wasserman 2004). Using the p-value mixture model framework, the limiting false discovery rate parameter at γ_0 is defined as

$$\pi(\gamma_0; \boldsymbol{\beta}) = \frac{\lambda\gamma_0}{\lambda\gamma_0 + (1-\lambda)(1 - \Phi[\Phi^{-1}(1-\gamma_0/2) - \tau])}. \tag{3}$$

The consistency of $\pi(\gamma_0; \hat{\boldsymbol{\beta}})$ results from the consistency of the quasilikelihood estimates. Alternatively, since the FDR defined in (3) is a monotone function of γ, it can be set to a sufficiently small value π_0, and a consistent estimate of the threshold p-value γ is found through the equation $\pi(\gamma; \hat{\boldsymbol{\beta}}) = \pi_0$. Functionally, this is accomplished by solving the equation

$$\gamma(\pi_0; \hat{\boldsymbol{\beta}}) = \frac{(1-\hat{\lambda})F_1(\gamma; \hat{\tau})\pi_0}{\hat{\lambda}(1-\pi_0)} \tag{4}$$

where $F_1(\gamma; \tau) = 1 - \Phi[\Phi^{-1}(1-\gamma/2) - \tau]$ is the distribution function of the p-values under the alternative hypothesis and evaluated at γ.

Although the FDR and p-value threshold estimates are consistent, their precision is a function of the level of dependence in the gene expression data. As this dependence increases, the confidence that the FDR and threshold estimates lie in a small neighborhood around their parameter values diminishes. To obtain a better understanding of these parameters, an asymptotic normal pivotal statistic is constructed to produce an asymptotic confidence interval for the FDR parameter $\pi(\gamma_0; \boldsymbol{\beta}_0)$ and its inverse function, the p-value threshold parameter $\gamma(\pi_0; \boldsymbol{\beta}_0)$.

The asymptotic normality of $\pi(\gamma_0; \hat{\boldsymbol{\beta}})$ stems from the asymptotic normality of the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$, derived in the theorem, and the continuity of π with respect to $\boldsymbol{\beta}$. The asymptotic variance of $\pi(\gamma;\hat{\boldsymbol{\beta}})$ follows directly from the asymptotic variance of the quasi-likelihood estimates and the delta method

$$\text{var}[\pi(\gamma_0;\hat{\boldsymbol{\beta}})] = \theta^T \Sigma \theta \qquad\qquad \text{where} \quad \theta^T = \left[\frac{\partial \pi}{\partial \lambda}, \frac{\partial \pi}{\partial \tau}\right].$$

The resulting symmetric $(1-\alpha)$ asymptotic confidence interval for the FDR at the p-value threshold level γ_0 is

$$\left[\pi(\gamma_0;\hat{\boldsymbol{\beta}}) - z_{1-\alpha/2}\sqrt{\text{var}\{\pi(\gamma_0;\hat{\boldsymbol{\beta}})\}}\ ,\ \pi(\gamma_0;\hat{\boldsymbol{\beta}}) + z_{1-\alpha/2}\sqrt{\text{var}\{\pi(\gamma_0;\hat{\boldsymbol{\beta}})\}}\ \right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ standard normal quantile. A confidence interval for the p-value threshold parameter at a given FDR level π_0 is constructed by applying the inverse transformation to the lower and upper confidence limits of $\pi(\gamma;\boldsymbol{\beta})$. Specifically, a $(1-\alpha)$ FDR confidence interval for any given γ may be written as

$$Pr[a_\gamma < \pi(\gamma;\boldsymbol{\beta}) < b_\gamma] = 1-\alpha.$$

By choosing γ to be the p-value threshold parameter for a FDR level π_0, i.e. $\pi(\gamma;\boldsymbol{\beta}) = \pi_0$, and applying the inverse transform, the $(1-\alpha)$ p-value threshold confidence interval is

$$Pr\left[\frac{(1-\lambda)F_1(\gamma;\tau)a_\gamma}{\lambda(1-a_\gamma)} < \gamma(\pi_0;\boldsymbol{\beta}) < \frac{(1-\lambda)F_1(\gamma;\tau)b_\gamma}{\lambda(1-b_\gamma)}\right] = 1-\alpha.$$

Evaluation of this asymptotic confidence interval for the p-value threshold parameter at the FDR level π_0 is obtained by substituting consistent estimates for $(\boldsymbol{\beta},\gamma)$ in the upper and lower confidence bounds. These estimates are obtained from equations (2) and (4).

Although either estimate, $\pi(\gamma_0;\hat{\boldsymbol{\beta}})$ or $\gamma(\pi_0;\hat{\boldsymbol{\beta}})$, may be used in differential gene expression analysis, for the purpose of gene selection, a direct approach is to fix the FDR and estimate the p-value threshold region; the genes that fall into the rejection region are chosen for further analysis. This is the approach carried out for the prostate cancer data set.

The proposed confidence interval can be employed in multiple ways for gene selection depending upon the objective of the analysis. If differential gene expression analysis is used to choose a small set of genes for validation by the laboratory scientist at the bench, then a $(1-\alpha)$ lower confidence bound for the p-value threshold parameter could be used for gene selection. Typically, RT-PCR, northern blots, or immunohistochemistry are used to validate the differentially expressed genes. Alternatively, if the goal is to use the error rate analysis as a screening tool to weed out uninteresting genes for subsequent classification or prediction analysis, then a liberal approach using a $(1-\alpha)$ upper confidence bound for $\gamma(\pi_0; \boldsymbol{\beta})$ would be suitable.

The original localized prostate cancer data analysis applied differential gene expression analysis as a filter to select candidate genes for model building and prediction (Stephenson et al. 2005). Using the p-value mixture model, the estimated p-value threshold is 2.1×10^{-3} for a FDR level equal to 0.05. Accounting for the variability and dependency in the gene expression data, the interval width from the 95% confidence interval for the p-value threshold parameter is 2.9×10^{-3}, which is substantial relative to the threshold estimate. The 95% upper confidence bound for the p-value threshold parameter is 3.8×10^{-3}. For the prostate cancer p-value gene list, if the threshold estimate was used as the filter, 367 genes would be selected for further analysis. In contrast, the 95% upper confidence bound filter would incorporate 575 genes for the model building component of the analysis. Thus, when using the FDR as a filter, accounting for the variability provides a more liberal, but rational, gene selection mechanism.

It is interesting to note the effect of the dependence assumption on the selection mechanism. Under the assumption that the gene expression values were independent, the interval width from the 95% confidence interval for the p-value threshold parameter is 4.6×10^{-4}, approximately one-eighth the interval width of the threshold parameter estimate under weak dependence. Not surprisingly, the independence assumption used in conjunction with 22,283 genes, provides a justification for the use of the threshold estimate as the filter, with little penalty for substituting the estimate for the parameter value. For this data set, however, the dependence between genes is an important aspect of the analysis. Thus, the uncertainty of the location of the FDR and p-value threshold parameters should be accounted for in the gene selection analysis.

Finally, an alternative approach is to infer the asymptotic FDR parameter for a given p-value threshold parameter. Gene selection based on the p-value threshold equal to 0.05, would produce an asymptotic FDR estimate equal to 0.240 with an estimated standard error equal to 0.030. Thus, the probability a selected gene is not differentially expressed may be as high as 0.300.

4 Simulations

A series of simulations were performed to assess the adequacy of the point and interval estimates of the limiting FDR and the p-value threshold parameters. A two-sample t-test was used to compute the p-value for each of 10000 'genes'. The two-group comparison was based on either 20 or 40 subjects per group. Within each group, the expression data for each gene were generated independently and identically distributed from either a normal or log-Weibull family. For each gene in group 1, a vector of n_1 independent identically distributed mean zero and variance 1 random variables were generated. For each gene in group 2, a vector of n_2 independent identically distributed random variables were generated with either mean zero and variance 1 or mean 2 and variance 1. The probability that the n_2 vector components had mean 2 was set equal to $1 - \lambda$. The parameter λ represents the proportion of true null hypotheses and was chosen to equal $\{0.3, 0.6, 0.9\}$ for the simulations. Within each subject, a block dependence structure between genes was generated. The block size was 500, with equal correlation ρ between genes within a block. The values of ρ used in the simulations were $\{0, 0.3, 0.6\}$. This correlation represents a m-dependence structure and satisfies the weak dependence conditions. Five hundred replications were run for each simulation.

The results of the simulations are presented in Tables 1 and 2. In general, the level of correlation between genes did not influence the bias or coverage estimates. For both the FDR and p-value threshold estimates, the bias increased as the parameter moved away from zero. The log-Weibull simulations were less accurate than the normal simulations.

Table 1. The columns in the table represent: the proportion of true null hypotheses (λ), the limiting FDR and p-value threshold (PVT) parameters, the block correlation parameter (ρ), the bias ($\times 10^3$) of the estimates of these parameters, and the empirical 95% coverage probability of these parameters. The sample size is forty observations per group.

		$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.6$	
λ	parameter value	Bias $\times 10^3$	Coverage	Bias $\times 10^3$	Coverage	Bias $\times 10^3$	Coverage
				Normal data			
0.3							
	FDR = 0.021	-0.024	0.960	-0.014	0.952	0.013	0.920
	PVT = 0.123	0.200	0.962	0.142	0.946	-0.021	0.926
0.6							
	FDR = 0.047	-0.089	0.942	0.034	0.948	0.082	0.928
	PVT = 0.053	0.063	0.942	-0.004	0.946	-0.030	0.920
0.9							
	FDR = 0.310	-0.027	0.922	0.697	0.952	0.082	0.914
	PVT = 0.006	0.006	0.924	-0.015	0.950	0.003	0.918
				log-Weibull data			
0.3							
	FDR = 0.021	-0.005	0.958	-0.006	0.942	0.018	0.926
	PVT = 0.123	0.083	0.966	0.096	0.948	-0.054	0.918
0.6							
	FDR = 0.047	-0.044	0.940	0.052	0.932	0.094	0.928
	PVT = 0.053	0.039	0.942	-0.014	0.932	-0.037	0.924
0.9							
	FDR = 0.310	0.160	0.922	0.766	0.950	0.152	0.918
	PVT = 0.006	0.000	0.928	-0.017	0.950	0.001	0.914

Table 2. The columns in the table represent: the proportion of true null hypotheses (λ), the limiting FDR and p-value threshold (PVT) parameters, the block correlation parameter (ρ), the bias ($\times 10^3$) of the estimates of these parameters, and the empirical 95% coverage probability of these parameters. The sample size is twenty observations per group.

		$\rho = 0.0$		$\rho = 0.3$		$\rho = 0.6$	
λ	parameter value	Bias $\times 10^3$	Coverage	Bias $\times 10^3$	Coverage	Bias $\times 10^3$	Coverage
				Normal data			
0.3							
	FDR = 0.021	0.079	0.948	0.062	0.956	0.036	0.948
	PVT = 0.123	-0.414	0.954	-0.310	0.948	-0.159	0.944
0.6							
	FDR = 0.047	-0.016	0.954	0.184	0.948	-0.009	0.934
	PVT = 0.053	0.022	0.952	-0.084	0.948	0.020	0.936
0.9							
	FDR = 0.310	0.141	0.952	-1.102	0.950	-0.767	0.934
	PVT = 0.006	0.001	0.946	0.035	0.962	0.027	0.944
				log-Weibull data			
0.3							
	FDR = 0.021	0.504	0.806	0.281	0.906	0.210	0.934
	PVT = 0.123	-2.893	0.792	-1.600	0.886	-1.159	0.926
0.6							
	FDR = 0.047	0.958	0.894	0.717	0.932	0.440	0.924
	PVT = 0.053	-0.497	0.890	-0.369	0.914	0.219	0.922
0.9							
	FDR = 0.310	3.274	0.938	1.059	0.966	1.542	0.926
	PVT = 0.006	-0.083	0.928	-0.024	0.956	-0.035	0.920

In Table 1, with forty subjects per group, the bias remained small and the 95% empirical coverage was uniformly good for all the simulations examined. In Table 2, however, with twenty subjects per group, the bias sometimes became large and had a negative impact on the coverage estimate, particularly in the log-Weibull simulations.

Additional simulations were run to explore the impact of violating the weak dependence assumption. Within each subject, all 10000 genes were equally correlated. The correlations examined were $\rho = \{0.2, 0.4, 0.6\}$. For the normal simulations with 40 subjects per group, the simulations resulted in a significant percentage of negative variance estimates for the FDR estimates. The percentage of replicates within a simulation that resulted in a negative variance estimate ranged from 15% to 50%. Thus, the proposed methodology is not robust to a strong dependence structure.

5 General Comments

A mixture model is proposed to determine a subset of genes associated with an outcome variable. Since the observed p-values used in the mixture model are derived from the asymptotic normality of the test statistic, this method is not confined to a specific test statistic or outcome variable type. The proposed methodology can be applied to a test statistic based on a comparison between groups (Student's t-statistic, Wilcoxon rank sum statistic, the log rank statistic, or their k-sample analogs), a test of association between variables (Pearson's correlation coefficient or Kendall's Tau), regression analyses or a multilevel factorial analysis.

The accuracy of the proposed methodology is a function of the accuracy of the asymptotic normality of the test statistic and the weak dependence assumption. We believe, however, there is a growing recognition that the inability to validate many gene expression analyses is a function of the limited number of samples in these analyses. Thus, it is our expectation that future gene expression studies will be based on larger sample sizes, enabling the asymptotic normality assumption to be justified on a greater proportion of studies.

The effect of within subject gene expression dependence on FDR measures is a current subject of research. Qui et al. (2005) demonstrated that dependence can impact the variability of an FDR measure. Efron (2005), using a test statistic mixture density, demonstrated that strong dependence can produce a significant deviation between the empirical and theoretical null component of the mixture density. The resulting bias in the FDR estimate may be reduced by adapting the null density to the observed data. For the p-value mixture density, this could entail replacing the standard Uniform null density with a Beta (ξ, θ) null density in the quasi-likelihood. Note that the standard Uniform null density is the special case $\xi = \theta = 1$. Whether this generalization produces a less biased FDR and p-value threshold estimate under strong dependence will be the subject of future research.

Strong dependence relationships such as exchangeability (Qui et al. 2005) and positive regression dependence (Benjamini and Yekutieli 2001) appear to have an adverse effect on the FDR measure and do not hold for the methodology presented in this paper. In contrast, under weak dependence, our simulations demonstrate that the FDR and p-value threshold estimates are accurate. What remains unclear is whether weak dependence is congruent to the concept of genetic pathways and hence whether it is sufficient to approximate the gene expression correlation structure. If weak dependence is not sufficient it may be possible to transform the expression data in the preprocessing algorithm prior to performing the proposed FDR analysis.

Our measure of the FDR differs from the conventional measure proposed in Benjamini and Hochberg (1995). Their FDR measure is based on a fixed number of tests performed; we have modified the FDR to present its limiting value. When the marginal p-values are generated from a single mixture distribution, the asymptotic FDR for a given threshold γ, is defined as the probability a gene is not differentially expressed given its p-value is less than γ. A benefit of the asymptotic FDR is the creation of an asymptotic pivotal statistic that is used to create a confidence interval for either the asymptotic FDR parameter, or its inverse function, the p-value threshold parameter. The confidence intervals are used to provide control of the error rate with a high level of confidence or as a liberal gene filter for subsequent statistical analyses of the gene expression data.

An alternative error rate analysis is based on controlling the tail probability for the proportion of false rejections. The properties of this error rate estimate, also known as the proportion of false positives (PFP) or the false discovery proportion (FDP), have been studied for both independent and dependent gene expression values (Korn et al. 2004, Genovese and Wasserman 2004, van der Laan et al. 2004). As noted in Genovese and Wasserman (2004), the FDP can be written as a function of the chosen threshold γ_0

$$S(\gamma_0) = \frac{\sum_g I(P_g \leq \gamma_0) I(D_g = 0)}{\sum_g I(P_g \leq \gamma_0)},$$

where it is assumed that at least one p-value is below the threshold γ_0. A connection between the tail probability for the FDP

$$\Pr[S(\gamma_0) > c]$$

and a confidence bound for the asymptotic FDR can be obtained using the central limit theorem assumption, $G^{1/2}[S(\gamma_0) - \pi(\gamma_0)]$ converges in distribution to a mean zero normal random variable, with asymptotic variance denoted by $W(\gamma_0)$ and the asymptotic FDR represented as $\pi(\gamma_0)$. The relationship between the two error rate analyses is realized by substituting

$$\pi(\gamma_0) + \frac{z_q W^{1/2}(\gamma_0)}{G^{1/2}},$$

for c in the FDP tail probability, producing a $(1-q)$ upper confidence bound for the asymptotic FDR

$$S(\gamma_0) - \frac{z_q W^{1/2}(\gamma_0)}{G^{1/2}}.$$

Finally, in this paper an empirical criterion was used to group genes into the dependent sets for the asymptotic variance calculation. As our understanding of the gene environment continues to improve, dependent gene sets may be established from biological determinants, such as through linkage of their gene function, location in the cell, or involvement in the biological process. One source currently available to establish these connections is the gene ontology consortium website (www.geneontology.org). As knowledge of gene interactions increase, the dependency classification can be carried out using external databases.

Appendix: Derivations of the asymptotic properties of the quasi-likelihood estimates

1) $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$

The proof will use the following arguments.
The log quasi-likelihood is defined as the sum of log p-value mixture densities

$$L_G(\boldsymbol{\beta}) = \sum_{g=1}^{G} \log \left\{ \lambda + (1-\lambda) \frac{1}{2} \frac{\phi_{\tau,1}(\Phi^{-1}(1-p_g/2))}{\phi_{0,1}(\Phi^{-1}(1-p_g/2))}. \right\}$$

where $\boldsymbol{\beta} = (\lambda, \tau)$.

It follows that for $\boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$,

$$E\left[\frac{\partial}{\partial \boldsymbol{\beta}} G^{-1} L_G(\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0$$

$$E\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} G^{-1} L_G(\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \quad \text{is negative definite.}$$

Let $\mathcal{N}(\boldsymbol{\beta}_0)$ denote a bounded neighborhood around $\boldsymbol{\beta}_0$ and assume $G^{-1}L_G(\boldsymbol{\beta})$ and its derivative are uniformly bounded in $\mathcal{N}(\boldsymbol{\beta}_0)$.

Define $\rho(\boldsymbol{\beta}) = \lim_{G\to\infty} G^{-1} L_G(\boldsymbol{\beta})$.

Then for $\delta > 0$,

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|>\delta} \rho(\boldsymbol{\beta}) < \rho(\boldsymbol{\beta}_0) \tag{A.1}$$

Proof:

Since $\hat{\boldsymbol{\beta}}$ is the maximum quasi-likelihood estimate,

$$G^{-1}L_G(\hat{\boldsymbol{\beta}}) \;\; \geq \;\; G^{-1}L_G(\boldsymbol{\beta}_0)$$

By the assumptions above, $G^{-1}L_G(\boldsymbol{\beta})$ converges uniformly to $\rho(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{N}(\boldsymbol{\beta}_0)$. Thus for $\epsilon > 0$,

$$G^{-1}L_G(\hat{\boldsymbol{\beta}}) \;\; \geq \;\; \rho(\boldsymbol{\beta}_0) - \epsilon$$

Adding and subtracting $\rho(\hat{\boldsymbol{\beta}})$ to the left side of the inequality and again using the uniform convergence argument

$$\rho(\hat{\boldsymbol{\beta}}) \;\; \geq \;\; \rho(\boldsymbol{\beta}_0) - \epsilon$$

which by (A.1) cannot occur unless $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}_0$.

2) $\qquad \sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \Sigma)$

Proof:

Let $S_G(\boldsymbol{\beta}) = \partial L_G(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $A_G(\boldsymbol{\beta}) = \partial^2 L_G(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$. Under the assumption of weak dependence, the weak law of large numbers and a Taylor expansion are applied to produce

$$G^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = [G^{-1}A_G(\boldsymbol{\beta}_0)]^{-1}G^{-1/2}S_G(\boldsymbol{\beta}_0) + o_p(1).$$

It follows that

$$\text{var}[G^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)] = [G^{-1}A_G(\boldsymbol{\beta}_0)]^{-1}[G^{-1}\text{var}S_G(\boldsymbol{\beta}_0)][G^{-1}A_G(\boldsymbol{\beta}_0)]^{-T}$$

and therefore, using the central limit theorem for weakly dependent data,

$$\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, A^{-1}VA^{-T}).$$

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C. K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, *39*, 1-20.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society, Series B*, *66*, 297-304.

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, Series B*, *57*, 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60-83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165-1188.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*, 1151-1160.

Efron B. (2006). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, *102*, 93-103.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, *32*, 1035-1061.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures.* New York: Wiley.

Hung, H. M. J., O'Neill, R. T., Bauer, P., Köhne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*, 11-22.

Korn, E. L., Troendle, J. F., McShane, L. M., Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, *124*, 379-398.

Lumley, T. and Heagerty, P. J. (1999). Weighted empirical adaptive variance estimators. *Journal of the Royal Statistical Society, Series B*, *61*, 459-477.

Owen, A. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society, Series B*, *67*, 411-426.

Pan, W., Lin, J., Le, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics*, *3*, 117-124.

Parker, R. A. and Rothenberg R. B. (1988). Identifying important results from multiple statistical tests. *Statistics in Medicine* *7*, 1031-1043.

Pounds, S. and Morris, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, *19*, 1236-1242.

Qiu, X., Klebanov, L., Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article 34.

Serfling, R. J. (1968). Contributions to central limit theorem for dependent variables. *The Annals of Mathematical Statistics*, *39*, 1158-1175.

Stephenson, A. J., Smith, A., Kattan, M.W., Satagopan, J., Reuter, V. E., Scardino, P. T., Gerald, W. L. (2005). Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, *104*, 290-298.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, *64*, 479-498.

Storey, J. D., Taylor, J. E., Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, *66*, 187-205.

Tsai, C. A., Hsueh, H. M., Chen, J. J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics*, *59*, 1071-1081.

van der Laan, M. J., Dudoit, S., Pollard K. S. (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, *3*, Issue 1, Article 15.