A mixture model for estimating the local false discovery rate in DNA microarray analysis

J.G. Liao1,*, Yong Lin1, Zachariah E. Selvanayagam2 and Weichung Joe Shih1

1Biometrics Division, School of Public Health and The Cancer Institute of New Jersey, University of Medicine and Dentistry of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA and 2Department of Pediatrics, UMDNJ—Robert Wood Johnson Medical School, One Robert Wood Johnson Place, MEB 3rd Floor, P.O. Box 19, New Brunswick, NJ 08903-0019, USA

Received on January 26, 2004; revised and accepted on April 22, 2004
Advance Access publication May 14, 2004

ABSTRACT
Motivation: Statistical methods based on controlling the false discovery rate (FDR) or positive false discovery rate (pFDR) are now well established in identifying differentially expressed genes in DNA microarray. Several authors have recently raised the important issue that FDR or pFDR may give misleading inference when specific genes are of interest because they average the genes under consideration with genes that show stronger evidence for differential expression. The paper proposes a flexible and robust mixture model for estimating the local FDR which quantifies how plausible each specific gene expresses differentially.

Results: We develop a special mixture model tailored to multiple testing by requiring the P-value distribution for the differentially expressed genes to be stochastically smaller than the P-value distribution for the non-differentially expressed genes. A smoothing mechanism is built in. The proposed model gives robust estimation of local FDR for any reasonable underlying P-value distributions. It also provides a single framework for estimating the proportion of differentially expressed genes, pFDR, negative predictive values, sensitivity and specificity. A cervical cancer study shows that the local FDR gives more specific and relevant quantification of the evidence for differential expression that can be substantially different from pFDR.

Availability: An R function implementing the proposed model is available at http://www.geocities.com/jg_liao/software
Contact: liaojg@umdnj.edu

1 INTRODUCTION
DNA microarray is a new technology that measures the expression levels of thousands of genes simultaneously. It is often used to identify genes that express differentially between two types of tissues or between two experimental conditions. In the example below, for example, we compare the gene expression levels between normal and cancerous cervical tissues. For a single gene, the question of differential expression can be answered by testing the hypothesis that the gene’s expression levels are the same between two types of tissues. Various forms of t-test or Wilcoxon rank test can be used. The resulting P-value summarizes the evidence against the hypothesis that the gene expression level does not differ between types of tissues. With the simultaneous testing of thousands of genes, however, a large number of non-differentially expressed genes can be falsely classified as differentially expressed if multiple testing is not properly accounted for. A comprehensive review of multiple testing in microarray analysis can be found in Dudoit et al. (2003) and Glonek and Solomon (2003).

The classical approach of controlling for the family-wise type I errors, such as the Bonferroni adjustment, is often too conservative to be useful for microarray data. A new approach of controlling the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995), has gained popularity and is now well developed. To set up the notation, let $H_i$ be the (null) hypothesis that the $i$-th gene does not express differentially. The outcome of testing $m$ genes simultaneously can be summarized as follows (Benjamini and Hochberg, 1995):

<table>
<thead>
<tr>
<th></th>
<th>Accepts $H_i$</th>
<th>Rejects $H_i$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes for which $H_i$ is true</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Genes for which $H_i$ is false</td>
<td>$T$</td>
<td>$S$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$W$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

*To whom correspondence should be addressed.
Benjamini and Hochberg (1995) defined the FDR as
\[
FDR = E\left[\frac{V}{R} \mid R > 0\right] \Pr(R > 0).
\]
They provide a multiple testing procedure that guarantees \( FDR \leq \alpha m \alpha /m \leq \alpha \) for a desired \( \alpha \) when the P-values from testing the non-differentially expressed genes are independent and uniformly distributed on \([0, 1]\). Storey (2002); Storey et al. (2003) and Storey (2003) propose a new class of procedures that provide a tighter FDR control and thus a higher power by incorporating information of \( m \alpha /m \) in the procedure. Storey (2003) also proposes a modified FDR, the positive FDR,
\[
pFDR = E\left[\frac{V}{R} \mid R > 0\right]
\]
and argues that it is more conceptually sound than DFL. For microarray data with a large \( m \) and many differentially expressed genes, the difference between pFDR and FDR is generally small as the extra factor in FDR, \( \Pr(R > 0) \), is very close to 1 except for extremely small \( t \). Storey’s methods are implemented in his R function \( q \) value available at http://faculty.washington.edu/~jstorey/. More discussion on FDR can be found in Benjamini and Yekutieli (2001); Finner and Roters (2002); Genovese and Wasserman (2002) and Reiner et al. (2003).

Efron et al. (2001) and Efron and Tibshirani (2002) define the local FDR to quantify the plausibility of a particular hypothesis being true, given its specific test statistic or P-value. To motivate this concept, let \( z_i \) be 1 if the \( i \)-th gene expresses differentially and 0 if it does not. It is natural to model \( z_i, i = 1, \ldots, m \), as Bernoulli trials with probability \( 1 - \pi_0 \), where \( \pi_0 = m \alpha /m \). Let \( \pi_i \) be the (raw) P-value from testing hypothesis \( H_i \). Let \( f_0 \) be the density of \( y_i \) distribution given \( z_i = 0 \) and \( f_1 \) be the density of \( y_i \) distribution given \( z_i = 1 \). It follows that \( f_0 \) is the uniform distribution on \((0, 1)\). The P-values \( y_1, \ldots, y_m \) then come from the two component mixture model:
\[
f(y) = \pi_0 + (1 - \pi_0) f_1(y).
\]
The local FDR is the posterior probability of \( H_i \) being true given its specific P-value \( y_i = t \)
\[
fdr(t) \equiv \Pr(H_i \text{ being true } \mid y_i = t) = \frac{\pi_0}{\pi_0 + (1 - \pi_0) f_1(t)}.
\]
For comparison, the pFDR for rejecting all hypotheses with \( y_i \leq t \), is (Storey, 2003)
\[
pFDR(t) \equiv \Pr(H_i \text{ being true } \mid y_i \leq t) = \frac{\pi_0 f}{\pi_0 f + (1 - \pi_0) F_1(t)},
\]
where \( F_1 \) is the cumulative distribution for density \( f_1 \) and \( t \) is the cumulative distribution of \( f_0 \). Efron and Tibshirani (2002) establish the averaging theorem that \( pFDR(t) = E_{y \sim f}(fdr(y) \mid 0 < y < t) \).

Several authors have recently raised the importance issue that FDR and pFDR can give misleading inference when particular genes are of interest. Finer and Roters (2002) discuss cheating with FDR. Suppose that one wishes to reject a particular hypothesis \( H_i \). One can simply group \( H_i \) with 99 other hypotheses \( H_2, \ldots, H_{100} \) that are false and will certainly be rejected. The FDR for the family of 100 hypotheses is then no greater than 1/100. Glonek and Soloman (2003) give more realistic examples. In their example one, the pFDR is 0.17 if we reject all the hypotheses with testing statistic \( Z \geq 2 \). Given the testing statistic \( Z \) in the small proximity of 2, however, the local FDR is a huge 0.99972. See also the rejoinder of Ge et al. (2003). These examples show that the averaging mechanism in pFDR may not be desirable. Suppose that we want to identify genes that show some evidence of differential expression for further biological study. The local FDR quantifies the gene-specific evidence for each gene. The pFDR, however, averages over other genes with stronger evidence. The local FDR should thus be preferred in such situation. See the cervical cancer example below for further discussion. The research for multiple comparisons has so far focused on FDR and pFDR. The methods for local FDR are much less developed. Statistically, it is more difficult to estimate the local FDR than FDR or pFDR. The FDR or pFDR can be formulated in terms of \( F \), the cumulative distribution of \( f \), for which the empirical distribution of \( y_1, \ldots, y_n \) is a consistent and stable estimator (Storey, 2003; Genovese and Wasserman, 2002). To estimate the local FDR, however, it is necessary to estimate the density \( f \) or \( f_1 \) and their stable estimation requires either a parametric model or some smoothing method (Efron et al., 20001). Efron et al. (20001) use the kernel density smoothing. Allison et al. (2002) formulate \( f_1 \) as a mixture of beta distributions. Pounds and Morris (2003) and G. Heller and J. Qing (Technical Report) parameterize \( f_1 \) as a beta distribution. The mixture model [Equation (1)], in terms of the test statistics instead of the \( P \)-values, is also studied in several papers. Pan et al. (2003) formulate the components of the mixture as a mixture of normal densities. K.-A. Do, P. Mueller and F. Tang (Technical Report) use a variation of the traditional Dirichlet process mixture models. These proposed methods are mostly generic two-component mixture models without incorporating special structures of multiple testing. As such, they tend to be either rigid and stable (such as a single beta distribution for \( f_1 \) or rich and fluctuating (such as a mixture of betas for \( f_1 \) with many mixture components). For example, a beta density parametrization of \( f_1 \) fails to fit the data in Pounds and Cheng (2004). This paper proposes a special parametric model tailored to multiple testing by requiring \( f_1 \) to be stochastically smaller than \( f_0 \), a structure appropriate for multiple testing. A smoothing mechanism is built in. The proposed model provides stable and robust estimation of the local FDR for any reasonable form of \( f_1 \). Our proposed method also gives a single framework for estimating the proportion
of differentially expressed genes, pFDR, negative predictive values, sensitivity and specificity.

2 PROPOSED MIXTURE MODEL

We now set to derive a good model for \( f_1 \) by requiring \( f_1 \) to be stochastically smaller than \( f_0 \):

\[
F_1(t) > F_0(t) \quad \text{for all } t \in (0, 1).
\]

This is very reasonable as \( F_1 \) is the \( P \)-value distribution from testing differentially expressed genes and \( F_0 \) is the \( P \)-value distribution from testing non-differentially expressed genes. Two popular models exist that impose a stochastic order (Agresti, 2002, Chapter 7). One is the proportional odds model (or cumulative logit model):

\[
\frac{F_1(y)}{1 - F_1(y)} = \frac{\theta F_0(y)}{1 - F_0(y)} = \frac{\theta y}{1 - y} \quad \text{for } \theta > 1.
\]

with solution

\[
F_1(y) = \frac{\theta y}{1 - y + \theta y} \quad \text{and} \quad f_1(y) = \frac{\theta}{(1 - y + \theta y)^2}. \tag{4}
\]

The other is the proportional hazard model:

\[
\frac{f_1(y)}{1 - F_1(y)} = \frac{\theta f_0(y)}{1 - F_0(y)} = \frac{\theta}{1 - y} \quad \text{for } \theta > 1.
\]

The solution is

\[
F_1(y) = 1 - (1 - y)^\theta \quad \text{or} \quad f_1(y) = \theta (1 - y)^{\theta - 1}. \tag{5}
\]

which is the one-parameter beta family beta \((1, \theta)\).

The two families of \( f_1 \), given in Equations (4) and (5), respectively, are both decreasing functions of \( y \). The local FDR, \( fdr(t) \), is thus an increasing function of the raw \( P \)-value \( t \), which is eminently reasonable. For the \( f_1 \) in Equation (4) we have \( \min f_1 > 0 \) and for the \( f_1 \) in Equation (5) we have \( \min f_1 = 0 \). Note that any \( f_1 \) with \( \min f_1 > 0 \) contains a component of \( f_0 \) in it (Pounds and Morris, 2003), which we believe should be absorbed in the null component in mixture model (1). We thus will choose the beta family in Equation (5) over the family in Equation (4). The beta \((1, \theta)\) distribution can model extreme concentration on the lower end of \([0,1]\) with a large \( \theta \) and the uniform distribution \( f_0 \) with \( \theta = 1 \). With a common hazard rate on the entire \([0,1]\), however, this beta family can be overly restrictive. It is natural to expand the model by allowing for different hazard ratios on different intervals. Let \( 0 = t_0 < t_1 < \cdots < t_k = 1 \) be pre-specified cut points and let \( \lambda_i > 1 \) be the hazard ratio of \( f_1 \) over \( f_0 \) on interval \([t_{i-1}, t_i], i = 1, \ldots, k \). Let

\[
\theta(y) = \sum_{i=1}^{k} \lambda_i I[t_{i-1}, t_i),
\]

where \( I[t_{i-1}, t_i) \) is the indicator function for interval \([t_{i-1}, t_i)\). Our piecewise proportional hazard model is

\[
\frac{f_1(y)}{1 - F_1(y)} = \frac{\theta(y)}{1 - y}
\]

with the solution

\[
F_1(y | \lambda) = 1 - \exp \left(- \int_0^y \frac{\theta(t)}{1 - t} dt \right),
\]

where \( \lambda = (\lambda_1, \ldots, \lambda_k) \). Define \( l \) to satisfy \( y \in [t_{l-1}, t_l) \). We then have

\[
f_1(y | \lambda) = \lambda_l (1 - y)^{\lambda_l - 1} \prod_{i=1}^{l-1} (1 - t_i)^{1 - \lambda_i}, \tag{6}
\]

and

\[
f_1(y | \lambda) = \lambda_l (1 - y)^{\lambda_l - 1} \prod_{i=1}^{l-1} (1 - t_i)^{1 - \lambda_i}. \tag{7}
\]

It is easy to see that

\[
f_1(y | \lambda) > F_0(y) \quad \text{for } y \in (0, 1)
\]

since \( F_0(y) \) is simply \( F_1(y | \lambda) \) with all \( \lambda_i \to 1 \). It also follows that

\[\min f_1(y | \lambda) = f_1(1 | \lambda) = 0\]

so that \( f_1 \) contains no component of \( f_0 \) in it. We choose a piecewise constant function for \( \theta(y) \) because of its simplicity. Other forms of \( \theta(y) \) we investigated lead to an intractable \( F_1 \).

The family \( f_1(y | \lambda) \) in Equation (7) can approximate, to any desired accuracy, any given underlying \( f_1 \) that is stochastically smaller than \( f_0 \) if \( k \) is large enough. A larger \( k \), however, can lead to unstable estimation of \( \lambda_i \)’s due to increased number of parameters. To solve this dilemma, we introduce a smoothing model on \( \lambda_i \)’s so that two neighboring \( \lambda_i \)’s do not change too rapidly. The parameter \( \lambda_i \) is defined on \((1, \infty)\) and is thus not a good scale to work on. We shall use the transformation

\[\tau_i = \log(\lambda_i - 1) \in (-\infty, \infty).\]

Our smoothing model is then

\[\tau_i - \tau_{i-1} \sim N(0, \sigma^2), \tag{8}\]

independently for \( i = 2, \ldots, k \). A smaller \( \sigma^2 \) pulls neighboring \( \tau_i \) closer to each other and has a stronger smoothing effect. The same smoothing model was successfully used by Liao and Brookmeyer (1995) to stabilize the estimated HIV infection rates in the United States.
3 STATISTICAL INFERENCE

We now propose a Bayesian inference for our proposed mixture model as specified by Equations (1), (7) and (8). To complete the model for Bayesian inference, we assign a diffuse prior on \( \tau_1 \). Our prior for \( \pi_0 \) is the standard conjugate prior beta(1,1). Following Gelman et al. (1995, p. 47), the prior distribution of \( \sigma^2 \) is taken to be the inverse chi-square distribution with scale \( \zeta_0 = 0.0001 \) and degree of freedom \( v \), which is the distribution of
\[
\frac{\xi_0^2 v}{\chi^2_v},
\]
where \( \chi^2_v \) is chi-square random variable with degree of freedom \( v \). Such distribution will be denoted as Inv-\( \chi^2(v, \xi_0) \).

This prior for \( \sigma^2 \) has little prior information with \( v = 1 \) but increasingly pulls \( \tau_1 \) toward each other as \( v \) increases (Gelman et al., 1995, p. 47). We will thus use \( v \) as a smoothing parameter as described below.

We fit the proposed Bayesian model using a block-at-a-time Metropolis–Hasting algorithm (Chip and Greenberg, 1995). Recall that the latent variable \( z_i \) takes value 1 if the \( i \)-th gene truly expresses differentially and takes value 0 if it does not. Assume that \( y_1, \ldots, y_m \) are independent of each other given \( z_1, \ldots, z_n \). We can work out the following conditional distributions:
\[
\begin{align*}
\pi_0 | \lambda_1, \ldots, \lambda_k & \sim \text{Bernoulli}(1 - \text{fdr}(y_i)), \\
\pi_0 | z_1, \ldots, z_m & \sim \text{beta} \left( m - \sum z_i + 1, \sum z_i + 1 \right), \\
\sigma^2 | \tau_1, \ldots, \tau_k & \sim \text{Inv-\( \chi^2(v + k - 1, \zeta) \)} \quad \text{with} \\
\zeta & = \frac{v \xi_0 + \sum_{i=1}^k (\tau_i - \tau_{i-1})^2}{v + k - 1}.
\end{align*}
\]
The conditional density of \( \tau_j \), given \( \pi_0, \sigma^2, z_1, \ldots, z_m \) and other \( \tau_j \) with \( j \neq i \), is proportional to
\[
\exp \left\{ -\frac{1}{2 \sigma^2} \sum_{j=2}^{i} (\tau_j - \tau_{j-1})^2 \right\} \prod_{j=1}^{m} f(y_j)^{z_j}
\]
as a function of \( \tau_j \). Our block-at-a-time Metropolis–Hasting algorithm continuously updates \( z_i, \pi_0, \sigma^2, \tau_i, i = 1, \ldots, k \), from these conditional distributions with \( \tau_i \) drawn using random walk Metropolis–Hasting. The program, localFDR, is implemented as an R (R Development Core Team, 2004; http://www.R-project.org) function and available at http://www.geocities.com/gq_liao. The convergence of the Metropolis–Hasting algorithm has proved pretty fast as runs with different starting values produce very close result after 5000 of iterations. The R package coda (Plummer, M., Best, N., Cowles, K., and Vines, K., http://r-project.org) can be conveniently used for further convergence diagnostics. The localFDR function computes the posterior distribution of \( \pi_0 \) and \( \lambda_1, \ldots, \lambda_k \) and use the respective posterior means for point estimates \( \hat{\pi}_0 \) and \( \hat{\lambda}_1, \ldots, \hat{\lambda}_k \). The \( \hat{\lambda}_1, \ldots, \hat{\lambda}_k \) are plugged in Equations (6) and (7) for the estimated density \( f_1 \) and cumulative distribution \( F_1 \). The \( f_1, F_1 \) and \( \hat{\pi}_0 \) are then plugged in Equations (2) and (3) for the estimated \( \text{fdr}(t) \) and \( \text{pFDR}(t) \). Estimates of the positive and predictive values and the sensitivity and specificity are obtained similarly. Note that the posterior mean of \( \text{fdr}(t) \) and \( \text{pFDR}(t) \) can also be used directly for point estimation. The cost in computing time and storage would be much larger, however. The program also estimates the following quantities of interest, for the rule of rejecting all hypotheses with \( y_i \leq t \),

| Negative predictive value | \( \text{Pr}(H_i \text{ being true } | y_i > t) = \frac{\pi_0(1 - t)}{1 - F(t)} \) |
| Sensitivity | \( \text{Pr}(y_i \leq t | H_i \text{ being false}) = F_1(t) \) |
| Specificity | \( \text{Pr}(y_i > t | H_i \text{ being true}) = 1 - t \) |

The negative predictive value is closely related to Genovese and Wasserman’s (2002) false non-discovery rate. Note that the positive predictive value, \( \text{Pr}(H_i \text{ being false } | y_i \leq t) \), is simply 1 − \( \text{pFDR}(t) \).

The statistical inference implemented in localFDR assumes that \( P \)-values \( y_1, \ldots, y_m \) are independent given \( z_1, \ldots, z_n \). Suppose \( y_1, \ldots, y_m \) are dependent but the empirical distribution of these \( P \)-values converges to \( F_0 \) for non-differentially expressed genes and converges to \( F_1 \) for differentially expressed genes. Our model and inference will continue to estimate \( \pi_0 \) and \( \lambda_1, \ldots, \lambda_k \) consistently since it is based on the empirical cumulative distribution of \( y_1, \ldots, y_m \) which converges to \( F \). The Bayesian credible intervals, however, will be too narrow.

We now address several practical issues. The first is how to choose the number of knots \( k \) and the location of knots \( t_1, \ldots, t_{k-1} \). A larger \( k \) makes a more flexible \( f_1 \) but is more computationally intensive. We use \( k = 100 \) in the examples below which should give a very flexible \( f_1 \). For the knots \( t_1, \ldots, t_{k-1} \), we choose \( t_j \) to be the \( jm/k \) quantile of \( y_1, \ldots, y_m \) which places the knots more densely where the \( P \)-values \( y_1, \ldots, y_m \) are concentrated and is desirable based on smoothing model (8). The second issue is how to determine the value of the smoothing parameter \( v \). We start with \( v = 0.10 k \) and increase it to \( v = 0.20 k, v = 0.30 k \), etc until the estimated \( \text{fdr} \) is generally a monotonically increasing function with no visible fluctuation. Our experience is that \( v = 0.30 k \) usually works well. It would be helpful to automate this process in future research. We note, however, the choice of the smoothing parameter in the simpler kernel density estimation is still unsettled after intensive research and visual inspection of the estimated density curves remains an important tool (Härdle et al., 2004). The third issue is model
checking. Let 
\[ F_m(t) = \#\{y_i \leq t\}/m \]
be the empirical cumulative distribution of \( y_1, \ldots, y_m \), which converges to \( F(t) \) uniformly over \( t \in [0, 1] \) by the Glivenko–Cantelli theorem (Billingsley, 1995). The estimated \( F(t) \) from our model is
\[ \hat{F}(t) = \hat{\pi}_0 t + (1 - \hat{\pi}_0)\hat{F}_1(t). \]
The examples in the next section all show very close agreement between \( \hat{F} \) and \( F_m \), indicating excellent fit of our model.

4 APPLICATION TO CERVICAL CANCER DATA
Dr K.V. Chin studies the gene expression levels of 10 391 genes on 26 cervical cancer patients and 9 normal subjects. Among the 26 cervical cancer patients, 11 are at stage 1 of the disease and 15 at more advanced stages (Wong et al., 2003). The research question is to identify differentially expressed genes between cancer and normal tissues and between stage one and more advanced stages of the cancer. Patient specimen and clinical data were from the Department of Obstetrics and Gynecology, Prince of Wales Hospital, at The Chinese University of Hong Kong. The cDNA targets were synthesized from the isolated total RNA with 33P–dCTP by oligo-dT-primed polymerization using Superscript II reverse transcriptase (Gibco/BRL) (Zheng et al., 2002). The microarray was done using Human GeneFilter arrays, GF223, from research Research Genetics, Inc. (Huntsville, AL) with its built-in imaging scanning and normalization procedures.

For each gene, the two-sample \( t \)-statistics are computed for comparing between cancerous and normal tissues and for comparing between different stages of the cancer. The raw \( P \)-value is then obtained based on the permutation distribution of the \( t \)-statistic as implemented in the multtest package (Ge and Dudoit, 2004, http://www.bioconductor.org). The two sets of \( P \)-values are then fitted by our smoothing piecewise proportional hazard model with \( k = 100 \) and \( t_1, \ldots, t_{k-1} \) as described in the previous section. To get a smoothed estimate of \( \text{fdr}(t) \), the smoothing parameter \( v = 0.30 \) \( k \) is used. For the cancer versus normal model, the estimate \( \hat{\lambda}_j \) starts very high with \( \hat{\lambda}_1 = 175.85 \) and declines rapidly to \( \hat{\lambda}_{100} = 2.17 \), confirming the need for piecewise hazard model. The posterior distribution of \( \pi_0 \) is summarized in Table 1. The estimated \( fdr(t) \) and \( \text{pFDR}(t) \) are plotted against \( t \) in Figure 1. The estimated negative predictive values and sensitivities for rejecting all hypotheses with \( y_i \leq t \) are plotted in Figure 2. The estimated \( \hat{F} \) from our model and the empirical cumulative distribution \( F_m \) are plotted in Figure 3 and they are almost identical, indicating excellent model fitting.

For comparing between normal and cancerous tissues, a total of 1043 genes are declared differentially expressed if we base on \( \text{pFDR} < 10\% \). The 516 genes among them, however, have \( \text{fdr} > 10\% \) and the smaller \( \text{pFDR} \) values are the result of averaging over genes with stronger evidence for differential expression. Only 527 genes will be declared differentially expressed if we choose base on \( \text{fdr} < 10\% \). We believe that whether a specific gene should be selected for further biological investigation should depend on the evidence for that specific gene, not other genes with stronger evidence. The inference based on \( \text{fdr} \) should thus be preferred. In working with biologists, we find it easier to communicate the concept of \( \text{fdr} \) than that of \( \text{pFDR} \) or \( \text{FDR} \).

For comparison, the two sets of \( P \)-values are also analyzed using Storey’s \( q \)-value function, which gives estimates of \( \text{pFDR}(t) \) and \( \pi_0 \) but not \( fdr(t) \). The estimated \( \text{pFDR}(t) \) from the \( q \)-value function (using \( \pi_0 \) estimate by the smoother method) is shown in Figure 1. The \( \text{pFDR}(t) \) estimates from our model and Storey’s method agree extremely well for the cancerous versus normal comparison. For the comparison between different stages of the cancer, the \( \text{pFDR}(t) \) estimates from the \( q \)-value function are unstable when \( t \) is near 0. The estimates of \( \pi_0 \) from \( q \)-value function are given in Table 1 and they agree well with our method.

4.1 A simulation study
We now apply our proposed model to some simulated datasets. The experiment is designed to have \( m = 10 \) 000 genes in total, with \( m_0 \) of them non-differentially expressed and \( m - m_0 \) of them differentially expressed. Each of the two comparison groups (cancer versus normal, for example) has 10 subjects.

We generate, for \( j = 1, \ldots, 10 \),
\[
\begin{align*}
x_{ij}^{[1]} & \sim N(0, 1), \quad i = 1, \ldots, m, \\
x_{ij}^{[2]} & \sim N(0, 1), \quad i = 1, \ldots, m_0, \\
x_{ij}^{[2]} & \sim N(\delta, 1), \quad i = m_0 + 1, \ldots, m.
\end{align*}
\]
The \( P \)-value \( y_i, i = 1, \ldots, m \), is computed from the two sided \( t \)-test comparing \( x_{ij}^{[1]}, j = 1, \ldots, 10 \) with \( x_{ij}^{[2]}, j = 1, \ldots, 10 \). Four such sets of \( P \)-values are generated with different \( m_0 \) and \( \delta \) as given in Table 2. Each set of \( P \)-values is then analyzed by our localFDR program with \( k = 100 \) and \( v = 0.30 k \) and also by Storey’s \( q \)-value function. Table 1 summarizes the posterior

| Table 1. Estimation of \( \pi_0 \) for Dr K. V. Chin’s cervical cancer data |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Percentiles of \( \pi_0 \) posterior | Estimated \( \pi_0 \) by Storey’s method | Smoother | Bootstrap |
| 2.5th | 50th | 97.5th |
| Cancer versus normal tissues | 0.726 | 0.752 | 0.768 | 0.738 | 0.738 |
| Stage one versus other stages | 0.844 | 0.867 | 0.896 | 0.860 | 0.840 |
Local false discovery rate

Fig. 1. The estimated fdr(t) and pFDR(t) from our model and estimated pFDR(t) using Storey’s method for Dr K. V. Chin’s cervical cancer data. The x-axis t is the raw $P$-value. The left panel is for cancer versus normal and the right panel is between stage one and more advanced stages of the cancer.

Fig. 2. The estimated sensitivity and negative predictive value for rejecting all $H_i$ with $y_i \leq t$ for Dr K. V. Chin’s cervical cancer data. The left panel is for cancer versus normal and the right panel is for comparing stage one and more advanced stages of the cancer.

distribution of $\pi_0$ from our model and the $\pi_0$ estimates using q value. The fdr(t) estimate from our model and two pFDR(t) estimates, one from our model and the other from Storey’s program, are plotted against t in Figure 4. Again, fdr(t) can be much larger than pFDR(t) and the two pFDR(t) estimates are very close.

5 DISCUSSION

This paper develops a flexible and robust mixture model tailored to multiple testing problem. The main motivation is to provide a tool for accurate estimation of local FDR which is more specific and relevant when particular genes are of interest. Our model also estimates the proportion of differentially expressed genes, pFDR, negative predictive values, sensitivity and specificity. Previously available tools such as Storey’s q value software do not estimate local FDR. For estimating pFDR and proportion of non-differentially expressed genes $\pi_0$, our model and Storey’s method tend to give very similar results although Storey’s pFDR estimation can be unstable when the raw $P$-value is near 0. This instability is also noted by Pounds and Cheng (2004).
Fig. 3. The estimated $\hat{F}(t)$ from our model and the empirical cumulative distribution $F_m(t)$ of $y_1, \ldots, y_m$ for Dr K. V. Chin’s cervical cancer data. The left panel is for cancer versus normal and the right panel is between stage one and more advanced stages of the cancer.

Fig. 4. The fdr$(t)$ and pFDR$(t)$ estimates from our model and pFDR$(t)$ estimates using Storey’s method for the four simulated datasets in Section 4.1 (datasets 1–4 from left to right and from top to bottom).
We have followed the popular approach of basing our multiple testing procedure on the raw $P$-values for individual genes (Dudoit et al., 2003; Glonek and Solomon, 2003). The benefit of this approach is it fits well with the current statistical practice in which the $P$-values are used and reported to summarize statistical evidence. It is also very flexible in easily allowing different tests, such as $t$-tests and non-parametric tests, to be used for the raw $P$-values. More sophisticated analysis that takes into account the dependence structure of different groups of genes may be carried out in the future as our knowledge of microarray data accumulates. Nevertheless, our proposed method will remain a useful tool for basic analysis before more complicated modeling is attempted.

ACKNOWLEDGEMENTS

The authors are grateful to Dr K.V. Chin for allowing us to use his cervical cancer data. The two reviewers’ comments help to significantly improve the paper. The work was supported by NCI grant 2P30 CA 72720-04.

REFERENCES


Table 2. Estimation of $\pi_0$ for the four simulated sets of $P$-values in Section 4.1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\delta$</th>
<th>$m_0/m$</th>
<th>Posterior percentiles of $\pi_0$</th>
<th>$\pi_0$ estimate by Storey’s methods</th>
<th>Smoother</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.850</td>
<td>0.827 0.862 0.893 0.851 0.837</td>
<td>0.851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>0.700</td>
<td>0.687 0.719 0.738 0.680 0.681</td>
<td>0.680</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>0.850</td>
<td>0.812 0.847 0.863 0.846 0.841</td>
<td>0.846</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.4</td>
<td>0.700</td>
<td>0.688 0.712 0.728 0.715 0.707</td>
<td>0.715</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>