Mixture models for assessing differential expression in complex tissues using microarray data

Debashis Ghosh

Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Received on September 15, 2003; revised on November 17, 2003; accepted on December 4, 2003
Advance Access publication February 26, 2004

ABSTRACT

Motivation: The use of DNA microarrays has become quite popular in many scientific and medical disciplines, such as in cancer research. One common goal of these studies is to determine which genes are differentially expressed between cancer and healthy tissue, or more generally, between two experimental conditions. A major complication in the molecular profiling of tumors using gene expression data is that the data represent a combination of tumor and normal cells. Much of the methodology developed for assessing differential expression with microarray data has assumed that tissue samples are homogeneous.

Results: In this paper, we outline a general framework for determining differential expression in the presence of mixed cell populations. We consider study designs in which paired tissues and unpaired tissues are available. A hierarchical mixture model is used for modeling the data; a combination of methods of moments procedures and the expectation–maximization algorithm are used to estimate the model parameters. The finite-sample properties of the methods are assessed in simulation studies; they are applied to two microarray data-sets from cancer studies. Commands in the R language can be downloaded from the URL http://www.sph.umich.edu/~ghoshd/COMPBIO/COMPMIX/.

Contact: ghoshd@umich.edu

INTRODUCTION

With the advent of high-throughput gene assay technologies, scientists are now able to measure genomewide mRNA expression levels in a variety of settings. One example is DNA microarrays (Schena, 2000). They have been utilized extensively in cancer profiling studies (Khan et al., 2001; Dhanasekaran et al., 2001), including the development of molecular classification systems based on the gene expression profile, discovery of cancer subtypes using such data and elucidation of the genomic differences in the progression of disease.

One of the major statistical tasks in studies involving these technologies is to find genes that are differentially expressed between two experimental conditions. The simplest example is to find genes that are up- or down-regulated in cancerous tissue relative to healthy tissue. In the setting of a single study, differential expression for microarray data is a well-studied problem. A brief but inexhaustive list of the work in this area includes the methods of Ideker et al. (2000), Efron et al. (2001), Newton et al. (2001), Baggeley et al. (2001), Dudoit et al. (2002), Pan (2002), Ibrahim et al. (2002) and Parmigiani et al. (2002). The methods of Efron et al. (2001), Newton et al. (2001), Baggeley et al. (2001), Ibrahim et al. (2002) and Parmigiani et al. (2002) utilize mixture models assuming that cancer and non-cancer samples are homogeneous.

A complication that has not been addressed as much in cancer profiling studies is that the tumor specimen profiled using microarrays is typically a mixture of different cell types. In normal tissues, proper differentiation and development of cells leads to organs in which the component tissues are relatively homogeneous. Tissues interact with neighboring cells by cell-contact, cytokines and the extracellular matrix. When these signals become disrupted, abnormal cell growth and alterations in epithelial cells may occur, leading to functional disorder (Bissell and Radisky, 2001). The problem has been noticed experimentally by Staal et al. (2003); some efforts to incorporate this into the analysis has been initiated by Venet et al. (2001).

In almost all analyses of microarray data, the tumor sample is treated as homogeneous. Consequently, in the analyses of the differential expression, there is a fundamental confounding with cell type. In the future, it may be possible to isolate pure populations of tumor cells using laser capture microdissection techniques (Fend and Raffeld, 2000); however, this technology is currently not in widespread use.

What is typically available is an assessment by the pathologist as to the percentage of the sample that is composed of tumor cells. We seek to utilize this information in the determining which genes are over- and under-expressed. In this paper, we develop a general framework for assessing differential expression in complex tissues that incorporates...
sample heterogeneity using mixture models (McLachlan and Peel, 2002). While we are primarily motivated by cancer studies, this issue is applicable to a variety of biological areas in which it is not possible to generate pure samples. Because of space limitations, we defer many technical details to Ghosh (2003, http://www.bepress.com/umbiostat/). The structure of this paper is as follows. In the Systems and Methods section, we describe the data structure and define two general probabilistic models for gene expression depending on whether an unpaired or paired study design is used. Estimation procedures for the models will be discussed here as well. We illustrate the use of the proposed methodologies to data from a cancer study and simulated data in the Results section. We conclude with some brief remarks in the Discussion section.

SYSTEMS AND METHODS

Data structures and study designs

We observe the random samples \((X_i^1, \ldots, X_i^p)\) and \((Y_j^1, \ldots, Y_j^m)\), where \(X_i^j\) is the \(p\)-dimensional gene expression profile for the \(i\)-th tumor sample, and \(Y_j^j\) is the corresponding profile for the \(j\)-th control sample, \(i = 1, \ldots, n; j = 1, \ldots, m\). In addition, we observe \(\pi = (\pi_1, \ldots, \pi_n)\), where \(\pi_i\) represents the proportion of the \(i\)-th tumor sample representing tumor cells \(i = 1, \ldots, n\). We will assume here that there are two cell types: tumor and normal. Thus, \((1 - \pi_i)\) \((i = 1, \ldots, n)\) will represent the percentage of the \(i\)-th tumor sample that is normal tissue. Throughout the article, we will assume that the data have been appropriately preprocessed and normalized. Again, the methods for assessing differential expression reported previously in the literature have implicitly made the assumption that \(\pi_i = 1, i = 1, \ldots, n\).

Before describing the model formulations, we briefly discuss the study designs appropriate for analysis using this model. In some cancer studies, normal and tumor samples come from separate patients. For these settings, we treat the cancerous and healthy samples as statistically independent, and it will not necessarily be the case that \(m = n\). In other experiments, normal and cancerous tissues come from the same patient. The normal tissue is taken from a region near the cancer; a consequence will be that \(m = n\). The analysis of these studies should take into account the pairing of samples. In this paper, we will consider formulations for both types of studies.

Model for unpaired study design

In this section, the model for an unpaired study design is described. We start by considering the gene expression profiles for the normal samples. Define \(Y_i^g\) to be the expression measurement for the \(g\)-th gene using the \(i\)-th control sample, \(g = 1, \ldots, G; i = 1, \ldots, m\). Equivalently, \(Y_i^g\) is the \(g\)-th component of \(Y_i\). Then, the model for gene expression in control samples we are formulating is the following:

\[
Y_i^1, \ldots, Y_i^G \sim f^c_i(Y), \quad (1)
\]

\[
f^c_i(y), \ldots, f^c_i(Y) \sim f^c_i(y). \quad (2)
\]

In Equation (1), we assume that the expression measurements from individual genes are random samples from a gene-specific probability model, conditional on a gene-specific effect, while the second stage of the model (2) states that the gene-specific densities are random samples from a probability distribution. For the tumor samples, we formulate the following model:

\[
X_i^1, \ldots, X_i^m \sim (1 - \pi_i) f^c_i(x) + \pi_i f^t_i(x), \quad (3)
\]

\[
f^t_i(x) \sim p_+ f^t_1(x) + p_- f^t_2(x) + (1 - p_+ - p_-) f^c_i(x). \quad (4)
\]

In the model for tumor expression, the measurements are no longer independent and identically distributed at the first stage (3); we are incorporating heterogeneity of the tumor specimens \(\pi_1, \ldots, \pi_n\). In addition, we model the gene expression measurements as mixtures of tumor and control gene expression densities. However, the tumor-specific densities of the \(G\) genes are modeled as being a random sample at the second stage of the model, in Equation (4). The first component in the mixture on the right-hand side of (4) represents the population of genes that are overexpressed in tumors relative to the control samples. The second component is the corresponding density for those genes that are underexpressed in tumors relative to controls. The proportion of genes in these two populations are \(p_+\) and \(p_-\), respectively. The remaining proportion of genes, \(1 - p_+ - p_-\), are from the usual control gene population, which represents the population of expression measurements in healthy tissue. Thus, Equations (1)–(4) specify a model for mixed tissues with cells of two different health conditions when samples are unpaired.

We next give the appropriate probability model for paired study designs.

Model for paired study design

We now formulate the probability model for tumor and control samples in the case of paired specimens. We have the following model:

\[
\begin{pmatrix}
X_i^1 \\
Y_i^1 \\
Y_i^G
\end{pmatrix} \mid \mu_i \sim \begin{pmatrix}
(1 - \pi_i) f^c_i(x) + \pi_i f^t_i(x) \\
f^c_i(x) \\
f^t_i(x)
\end{pmatrix}, \quad (5)
\]

\[
\begin{pmatrix}
f^c_i(x) \\
f^t_i(x)
\end{pmatrix} \sim \begin{pmatrix}
p_+ f^t_1(x) + p_- f^t_2(x) \\
(1 - p_+ - p_-) f^c_i(x)
\end{pmatrix}, \quad (6)
\]

\[
\mu_1, \ldots, \mu_n \sim M. \quad (7)
\]
This model takes into account the pairing of samples. In (5), we assume that the control and tumor expression measurement for the $g$-th gene from the $i$-th sample, conditional on a gene effect, is a random sample from a bivariate distribution, where the first component involves the tumor heterogeneity, and the second component is a gene-specific density for control samples. The second stage of the model is given in (6) and (7), where the densities are a random sample from a bivariate distribution. The density for $f^c$ corresponds to that given in (4). Finally, we also need a model formulation for the sample effects $\mu_1, \ldots, \mu_n$; the distribution of these effects is given by $M$ in (7). While there is a multi-stage formulation in both models (1)–(4) and (5)–(7), the latter model is fundamentally bivariate, while the former model models the distributions of the gene expression profiles for control and tumor samples separately. Both models (1)–(4) and (5)–(7) are examples of mixture models. It should also be noted that these models imply that there is a dependence in gene expression measurements between genes. This is because of the two-stage hierarchical formulation we have adopted and is also implied by the models of Newton et al. (2001), Efron et al. (2001), Ibrahim et al. (2002) and Parmigiani et al. (2002). However, these methods assume that $\pi_i = 1$ for $i = 1, \ldots, n$.

In thinking about probabilistic specifications for the models described previously, we want to incorporate the fact that the number of samples $(n, m)$ will be much smaller than the number of genes $(G)$ in gene expression studies. What this implies is that we want to be parametric in the first stage and less parametric in the second stage of the models. By utilizing the hierarchical specifications, this allows us to share information across genes in a natural way. This approach was also incorporated by other authors, such as Newton et al. (2001), Efron et al. (2001) and Parmigiani et al. (2002). However, they were not dealing with the complex tissue scenario addressed here.

The ultimate goal here is to calculate a quantity summarizing differential expression of a gene in tumor tissue relative to healthy tissue. Because we have formulated the problem using mixture models, a natural output in this procedure is the posterior probability of differential expression given the observed data. Similar measures have been developed quite extensively in the situation where $\pi_i = 1$ for $i = 1, \ldots, n$, but not in the more complex tissue problem.

**Model specifications**

We first start by considering the unpaired study design model (1)–(4). We specify that $f^c$ is the density function of a normal random variable with mean $\mu_{gc}$ and variance $\sigma^2_{gc}$ in (1) and that $f^t$ is that of a normal distribution with mean $\mu_{gt}$ and variance $\sigma^2_{gt}$. In the second stage of the model [Equations (2) and (4)], $\sigma^2_{gc}$ and $\sigma^2_{gt}$ are assumed to be from distributions with mean $\sigma^2$ and $\sigma^2$. We will assume that the distribution of $\mu_{gc}$ ($g = 1, \ldots, G$) at the second stage is from a normal distribution with mean $\mu_c$ and variance $\sigma^2$. For the distribution of $\mu_{gt}$, we will formulate the following model:

\[
\mu_{gt} \sim p_+ N(\mu_+, \sigma^2) + p_- N(\mu_-, \sigma^2) + (1 - p_+ - p_-) N(\mu_c, \sigma^2).
\] (8)

In (8), we state that the average gene expression level in tumors comes from a mixture of three distributions. The first distribution on the right-hand side of (8) is for the genes that are overexpressed in tumor relative to normal tissue. The second mixture component corresponds to the population of genes that are underexpressed in tumor cell populations relative to normal cell populations. The last mixture component in (8) represents the genes that do not change between normal and tumor tissue. Notice that we are examining and testing for differential expression using means. The proportion of genes that fall into the three gene populations are given by $p_+$, $p_-$ and $p_0 \equiv (1 - p_+ - p_-)$. A natural constraint to impose is that $\mu_- < \mu_c < \mu_+$. We can reformulate the general model in Equations (1)–(4) in the following manner:

\[
Y_{gi}^c \sim N(\mu_{gc}, \sigma^2_{gc}),
\] (9)

\[
X_{gi}^t \sim \pi_i N(\mu_{gt}, \sigma^2_{gt}) + (1 - \pi_i) N(\mu_{gc}, \sigma^2_{gc}),
\] (10)

\[
\mu_{gc} \sim N(\mu_c, \sigma^2),
\] (11)

\[
\mu_{gt} \sim p_+ N(\mu_+, \sigma^2) + p_- N(\mu_-, \sigma^2) + (1 - p_+ - p_-) N(\mu_c, \sigma^2),
\] (12)

\[
\sigma^2_{gc} \sim F_c,
\] (13)

\[
\sigma^2_{gt} \sim F_t,
\] (14)

where $F_c$ and $F_t$ are the distribution functions corresponding to $\sigma^2_{gc}$ and $\sigma^2_{gt}$, respectively, $g = 1, \ldots, G$.

We next consider the paired study design model (5)–(7). We will make similar specifications as those for the unpaired design, while taking into account the pairing of the samples. At the first stage of the model [corresponding to Equation (5)], we assume that $(Y_{gi}^c, X_{gi}^t) (g = 1, \ldots, G; i = 1, \ldots, n)$ has a bivariate normal distribution with mean vector

\[
\begin{pmatrix}
\mu_i + \mu_{gc} \\
\mu_i + \pi_i \mu_{gt} + (1 - \pi_i) \mu_{gc}
\end{pmatrix}
\]

and variance–covariance matrix

\[
\begin{pmatrix}
\sigma^2_{gc} & \pi_i \sigma^2_{gt} + (1 - \pi_i)^2 \sigma^2_{gc} \\
\pi_i \sigma^2_{gt} & \sigma^2_{gt}
\end{pmatrix}
\]

For the models corresponding to (6), we assume that $(\mu_{gc}, \mu_{gt})$ are iid observations from a bivariate mixture normal distribution, which we state below in (17). Finally, the subject effects $\mu_1, \ldots, \mu_n$ are assumed to come from a distribution function $M$. We will also assume that all variance and
covariance parameters come from distributions that we leave unspecified. Thus, the model can be stated in the following manner:

\[
\begin{align*}
(Y_{gi}^c & )_{gi} \sim N_2 \left[ \mu_i + \sigma_{gc} \right], \\
\sigma_{gc}^2 & = \pi_1 \sigma_{gct}^2 + (1 - \pi_1) \sigma_{gc}^2,
\end{align*}
\]

where \( \sigma_{gct} \) is the covariance term in (15), we allow for the normal and tumor components to interact; this is summarized by the covariance term \( \sigma_{gct} \).

We consider the model for the unpaired study design first. We are ready to describe the necessary estimation procedures. Our method, they do no make any assumptions on the number of component cell populations; they vary the number of components in their algorithm. In addition, their goal is to attempt to reconstruct the measurement for the component cell, while we do not perform such a reconstruction.

**Estimation procedures**

We consider the model for the unpaired study design first. We employ the following multi-stage algorithm for estimation:

1. Estimate \( \mu_{gc} \) and \( \sigma_{gc}^2 \) in (9) by

\[
\hat{\mu}_{gc} = n^{-1} \sum_{i=1}^{n} Y_{gi}^c,
\]

and \( \hat{\sigma}_{gc}^2 = (n - 1)^{-1} \sum_{i=1}^{n} (Y_{gi}^c - \hat{\mu}_{gc})^2, \)

2. Estimate \( \mu_{gt} \) and \( \sigma_{gt}^2 \) in (10) by method of moments estimation, using the estimates of \( \mu_{gc} \) and \( \sigma_{gc}^2 \) from step 1.

3. Estimate \( \mu_{g+}, \mu_{g-}, \mu_1, \mu_2 \) and \( \sigma_{g+}^2, \sigma_{g-}^2, \sigma_{g1}^2, \sigma_{g2}^2 \) by the EM algorithm where the ‘data’ are \( (\hat{\mu}_{gc}, \hat{\mu}_{gt}) \), \( g = 1, \ldots, G \).

4. Estimate \( F_c \) and \( F_t \) by the empirical distributions of the estimators of \( \sigma_{gc}^2 \) and \( \sigma_{gt}^2 \) from steps 1 and 2, \( g = 1, \ldots, G \).

For the paired study design [Equations (15)–(20)], we first start by estimating \( \hat{\mu}_i \) by

\[
\hat{\mu}_i = (2G)^{-1} \sum_{g=1}^{G} (Y_{gi}^c + X_{gi}^t);
\]

\( \hat{\mu}_i \) is then subtracted from the gene expression measurements for the \( i \)-th individual, \( i = 1, \ldots, n \). After this adjustment, we follow the steps of the estimation algorithm for the unpaired design. We also tried a procedure based on the differences between the paired samples, but this method had relatively poor finite-sample properties in simulation studies.

A general description of the EM algorithm in normal mixture models can be found in Ghosh and Chinnaiyan (2002) and the Appendix of Ghosh (2003). The methods here have been implemented using the R language (Ihaka and Gentleman, 1996). If we define the random variables \( D = 1, 2 \) and 3 and corresponding to the populations of underexpressed, over-expressed and non-differentially expressed genes in tumor relative to healthy tissue, then for each gene we will have a measure of the posterior probability that \( D = 1, D = 2 \) and \( D = 3 \), conditional on the observed data. Genes with large posterior probabilities of \( D = 1 \) are likely to be underregulated genes; similar interpretations hold with \( D = 2 \) and \( D = 3 \). It turns out that these posterior probabilities are related to the false discovery rate (Benjamini and Hochberg, 1995); further details are found in Ghosh (2003).

In the method of Venet et al. (2001), the goal is to reconstruct gene expression for component cell populations. Unlike our method, they do not make any assumptions on the number of component cell populations; they vary the number of components in their algorithm. In addition, their goal is to attempt to reconstruct the measurement for the component cell, while we do not perform such a reconstruction.

**RESULTS**

We now discuss the application of the proposed methodology to both real and simulated datasets.

**Colon cancer data**

Our first example comes from a recently reported study by Alon et al. (1999), in which Affymetrix HuGeneFL oligonucleotide microarrays were used to probe colon adenocarcinoma samples. While the initial study reported on differential expression between 40 cancerous and 22 normal samples, we focus on a subgroup of 18 patients in which normal and cancer samples were paired. The pairing was not taken into account in the analysis performed by Alon et al. (1999). In addition, the samples involved contamination with normal adjacent tissue. Based on the framework described in the paper, this corresponds to \( \pi \neq 1 \); however, the samples were analyzed as if \( \pi = 1 \).

Before describing the analysis, we describe the data preprocessing steps that were taken. There were 7471 genes...
Assessing differential expression with microarray data

Table 1. Number of genes called significant (\(k\)) for various values of \(q^*\) based on the Storey (2002) procedure for paired normal and colon cancer data from Alon et al. (1999).

<table>
<thead>
<tr>
<th>(q^*)</th>
<th>No adjustment</th>
<th>Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>2824</td>
<td>2824</td>
</tr>
<tr>
<td>0.1</td>
<td>2704</td>
<td>2809</td>
</tr>
<tr>
<td>0.05</td>
<td>2360</td>
<td>2604</td>
</tr>
<tr>
<td>0.02</td>
<td>1906</td>
<td>2375</td>
</tr>
<tr>
<td>0.01</td>
<td>1635</td>
<td>2210</td>
</tr>
<tr>
<td>0.005</td>
<td>1441</td>
<td>2061</td>
</tr>
<tr>
<td>0.001</td>
<td>966</td>
<td>1693</td>
</tr>
</tbody>
</table>

in the original dataset, output from Affymetrix MAS4.0, downloaded from the URL http://microarray.princeton.edu/oncology/. Based on the data from the website, genes with any missing values or with negative expression were excluded from further consideration. This left a total of 2824 genes. Afterwards, logarithms of base two were taken, which is the data we work with.

The first analysis consisted of assuming that \(\pi = 1\) for all tumor samples and performing a univariate analysis using \(p\)-values based on a \(t\)-distribution with a \(q\)-value calculation based on the FDR method of Storey (2002). We find that under the null distribution, we would expect to find \(\sim 18\%\) of genes to be differentially expressed. Table 1 lists the numbers of genes called significant at various \(q\)-values (Table 1).

A second analysis consisted of incorporating the tissue composition information into the analysis using the following method:

1. Estimate \(\mu_i\) by:
   \[
   \hat{\mu}_i = (2G)^{-1} \sum_{g=1}^{G} (Y_{gi}^c + X_{ig}^t),
   \]
   subtract \(\hat{\mu}_i\) from the gene expression measurements for the \(i\)-th individual, \(i = 1, \ldots, n\).

2. \(\mu_{gc}, \sigma_{gc}^2\) and \(\mu_{tc}, \sigma_{tc}^2\) are estimated using methods of moments.

3. A test statistic
   \[
   T_g = \frac{\hat{\mu}_{gc} - \hat{\mu}_{tc}}{(\hat{\sigma}_{gc}^2 + \hat{\sigma}_{tc}^2)^{1/2}},
   \]
   \(g = 1, \ldots, G\), is constructed.

4. The \(q\)-values are calculated based on \((T_1, \ldots, T_G)\) using the method of Storey (2002).

By incorporating the tissue information in this way, we reduce the rate of non-differentially expressed genes from 18 to 10%.

Based on Table 1, we also find that the number of genes being counted as differentially expressed increases sharply when we incorporate the tissue information based on the above algorithm.

We now apply the method proposed in Systems and Methods to the data, taking into account the paired nature of the data. Based on the fitting procedure, we find that 30% of the genes are found to be non-differentially expressed. If we use a posterior probability of 95% for determining differentially expressed genes, we find that 98 genes are underexpressed in cancer relative to normal, while 939 genes are correspondingly overexpressed. This corresponds to using a FDR of 5% in the framework of Storey (2002).

In comparing the proposed method with the other methods, there are genes that are called differentially expressed by our method that are not being determined to be differentially expressed by FDR methods. The comparisons with the first two analysis methods are given graphically in Figures 1 and 2.

It is clear that there are genes that our method determines to be differentially expressed that have more conservative \(q\)-values. For example, in Figure 1, there are 458 genes with a \(q\)-value greater than 0.05 that are determined to have \(>95\%\) posterior probability of being differentially expressed.

In Ghosh (2003), we also report on a similar comparison between the usual SAM analysis with the proposed method for data from an unpaired prostate cancer study by Luo et al. (2001). R code for the analysis of the two datasets can be downloaded from the URL http://www.sph.umich.edu/~ghoshd/COMPBIO/COMPMIX/.
For a given gene, the vertical axis represents the posterior probability of differential expression, while the horizontal axis represents the q-value using the second proposed method for colon cancer data of Alon et al. (1999).

**Simulation studies**

To assess the finite-sample properties of the proposed methodology, we conducted some simulation studies. We compared the mixture model-oriented procedures with the t-test analysis that ignores tumor heterogeneity. Both the paired and unpaired designs were considered. Data were generated using the paired and unpaired models, Equations (15)–(20) and (9)–(14), respectively. We took $N = 100$ and $N = 60$ for the unpaired design; $\pi_1, \ldots, \pi_M$ are a random sample from the uniform (0,3,1) distribution, where $M = N/2$. We took all variance components to be 1. We set $\mu_e = 0$ and examined $(\mu_+, \mu_-) = (1.2, -1.2), (0.8, -0.8)$ and (0.4, -0.4). For proportion of differentially expressed genes, we considered $(p_+, p_-) = (0.05, 0.05)$; similar results held for $(p_+, p_-) = (0.2, 0.2), (0.1, 0.2)$ and (0.2, 0.1) (data not shown). Finally, for the paired design, we took $M$ to be a normal distribution with mean 1.2 and variance 1. One thousand simulation samples were considered for each setting.

Because we have three populations of genes (overexpressed, underexpressed and no differential expression), we used sensitivity and specificity to assess the properties of the proposed methods and the t-test approach. For the proposed methodology, we defined sensitivity as posterior probability of differential expression greater than 0.9999999 among the differentially expressed genes and specificity as posterior probability of non-differential expression greater than 0.9999999 among the non-differentially expressed genes. For the t-test, we defined sensitivity as a magnitude of t-test greater than 5 among differentially expressed genes and specificity a magnitude of t-test less than 5 among non-differentially expressed genes. The results are given in Tables 2 and 3. Based on the situations considered, we find that the proposed method performs substantially better than the t-test analysis that ignores tissue heterogeneity.

**DISCUSSION**

There are several extensions to the method we have proposed here. One could envision alternative probabilistic specifications for the mixture models. For example, some authors have reported extreme outliers in microarray data. While we have sought to remove outliers in our preprocessing, one could also use mixture models based on the t-distribution (Lange et al., 1989). In this instance, one would also have to resort to more complicated model fitting procedures based on extensions of the EM algorithm (McLachlan and Krishnan, 2000). Another specification of the model is to incorporate parametric distributions for the distributions of the mean and variance parameters. Our estimation procedures were based on the EM algorithm, but one could also potentially use Bayesian mixture modeling methods Diebolt and Robert (1994) for estimation.

**Table 2.** Summary of simulation results for unpaired design

<table>
<thead>
<tr>
<th>$N$ $(\mu_+, \mu_-)$</th>
<th>Proposed method</th>
<th>t-Test method</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 $(-0.4,0.4)$</td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td>60 $(-0.8,0.8)$</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>60 $(1.2,1.2)$</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>100 $(-0.4,0.4)$</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>100 $(0.8,0.8)$</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>100 $(1.2,1.2)$</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Proportion of differentially expressed genes given by $(p_+, p_-) = (0.05, 0.05)$. All variances are set to one. t-Test method denotes t-test analysis treating tumor samples as homogeneous.

**Table 3.** Summary of simulation results for paired design

<table>
<thead>
<tr>
<th>$N$ $(\mu_+, \mu_-)$</th>
<th>Proposed method</th>
<th>t-Test method</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 $(-0.4,0.4)$</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>60 $(-0.8,0.8)$</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>60 $(1.2,1.2)$</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>100 $(-0.4,0.4)$</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>100 $(0.8,0.8)$</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>100 $(1.2,1.2)$</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Proportion of differentially expressed genes given by $(p_+, p_-) = (0.05, 0.05)$. All variances are set to one. t-Test method denotes t-test analysis treating tumor samples as homogeneous.
In terms of dealing with mixtures of cell populations in samples, biologists often speak of methods that ‘subtract’ out the effects of the contaminating mixture components. The technique of laser capture microdissection (Fend and Raffeld, 2000) is a lab-based method that does this. We have demonstrated that through the use of statistical methods, it is possible to develop analysis procedures that also can subtract out the effects of individual mixture components.

ACKNOWLEDGEMENTS
The author would like to thank the referees for their comments, which substantially improved the presentation.

REFERENCES