University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 22

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

Debashis Ghosh*

*University of Michigan, ghoshd@psu.edu

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

Debashis Ghosh

Abstract

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. In this article, 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 modelling the data; a combination of methods of moments procedures and the expectation-maximization (EM) algorithm are used to estimate the model parameters. Links with the false discovery rate are discussed. The methods are applied to two microarray datasets from cancer studies as well as to simulated data.

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

Debashis Ghosh
Department of Biostatistics, University of Michigan
1420 Washington Heights
Ann Arbor, MI 48109-2029

Corresponding author:

Debashis Ghosh, Ph.D.
Department of Biostatistics
School of Public Health, University of Michigan
1420 Washington Heights, Room M4057
Ann Arbor, Michigan 48109-2029
Phone: (734) 615-9824
Fax: (734) 763-2215
Email: ghoshd@umich.edu

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 article, 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 modelling the data; a combination of methods of moments procedures and the expectation-maximization (EM) algorithm are used to estimate the model parameters. Links with the false discovery rate are discussed. The methods are applied to two microarray datasets from cancer studies as well as to simulated data. R code for analyzing the datasets can be downloaded from the following URL:

http://www.sph.umich.edu/∼ghoshd/COMPBIO/COMPMIX/

Contact: ghoshd@umich.edu

Keywords: Bioinformatics, DNA Microarrays, Gene Expression, Mixed Cell Populations, Multiple Comparisons, Variable Selection.

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 (Lipshutz et al., 1999; Schena, 2000). They have been utilized tremendously in cancer profiling studies (Alizadeh et al., 2001; Khan et al., 2001; Dhanasekaran et al., 2001). Goals there have included 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 Efron et al. (2001), Newton et al. (2001), Ideker et al. (2001), Baggerly et al. (2001), Dudoit et al. (2002), Pan (2002), Ibrahim et al. (2002) and Parmigiani et al. (2002).

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). If there are epithelial cells with tumorigenic potential located in this environment, then they can start to proliferate. Thus, the tumor grows within the context of this microenvironment, and the sample taken from the patient represents a mixture of tumor and normal cells. This problem also relates to the 'seed and soil' hypothesis formulated by Paget (1989). In a recent experiment, Creighton et al. (2003) attempted to address this hypothesis using a mouse xenograft models, but they concluded their findings to be preliminary. The problem has also been noticed experimentally by Staal (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. None of the methods listed in the previous paragraph accounts for this mixed cell type problem.

Our experience with this issue arises from a recent molecular profiling study in prostate cancer (Dhanasekaran et al., 2001). Some of the microarray experiments there used a normal adjacent prostate pool as the reference sample (i.e. the Cy3-labelled sample), and the investigators thought that this tissue might be influenced by paracrine effects mediated by PCA, and furthermore is exposed to the same environmental and genetic factors as the adjacent PCA. Thus, in determining differential expression, it would be important to "subtract" out these effects. In another example, a recent article by Hampson and Hughes (2001) discussed the problem in the context of determining differentially expressed genes in muscle when there are muscle-specific cells and non-muscle specific cells.

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 underexpressed. In this article, we develop a general framework for assessing differential expression in complex tissues that incorporates sample heterogeneity. 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. It turns out that a natural tool in the probabilistic formulation of this problem is mixture models (McLachlan and Peel, 2002). The structure of this paper is as follows. In **Systems and Methods**, 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, along with comparisons with procedures based on the false discovery rate (Benjamini and Hochberg, 1995). We illustrate the use of the proposed methodologies to two cancer studies and simulated data in **Results**. We conclude with some brief remarks in the **Discussion**.

Systems and Methods

Data Structures and Study Designs

We observe the random samples $(\mathbf{X}_1^t, \ldots, \mathbf{X}_n^t)$ and $(\mathbf{Y}_1^c, \ldots, \mathbf{Y}_m^c)$, where $\mathbf{X}_i^t$ is the p-dimensional gene expression profile for the ith tumor sample, and $\mathbf{Y}_j^c$ is the corresponding profile for the jth control sample, $i = 1, \ldots, n$, $j = 1, \ldots, m$. In addition, we observe $\pi \equiv (\pi_1, \ldots, \pi_n)$, where π_i represents the proportion of the ith 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 ith tumor sample that is normal tissue. Throughout the paper, we will assume that the data have been appropriately preprocessed and normalized. Again, the methods for assessing differential expression previously reported 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 manuscript, 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_{ig}^c to be the expression measurement for the gth gene using the ith control sample, $g = 1, \ldots, G; i = 1, \ldots, m$. Equivalently, Y^c_{ig} is the gth component of $\mathbf{Y}^c_i$. Then the model for gene expression in control samples we are formulating is the following:

$$Y^c_{1g}, \ldots, Y^c_{mg} \sim f^c_g(y) \tag{1}$$

$$f^c_1(y), \ldots, f^c_G(y) \overset{iid}{\sim} f^c(y) \tag{2}$$

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:

$$X^t_{1g}, \ldots, X^t_{mg} \sim (1 - \pi_i) f^c_g(x) + \pi_i f^t_g(x) \tag{3}$$

$$f^t_g(x) \overset{iid}{\sim} p_+ f^t_+(x) + p_- f^t_-(x) + (1 - p_- - p_+) f^c(x) \tag{4}$$

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 modelled 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.

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:

$$
\begin{aligned}
\begin{pmatrix} X_{gi}^t \\ Y_{gi}^c \end{pmatrix} |\mu_i &\sim \begin{bmatrix} (1-\pi_i) f_g^c(x) + \pi_i f_g^t(x) \\ f_g^c(y) \end{bmatrix} && (5) \\
\begin{pmatrix} f_g^c(x) \\ f_g^t(x) \end{pmatrix} &\overset{iid}{\sim} \begin{pmatrix} f^c(x) \\ p_+ f_+^t(x) + p_- f_-^t(x) + (1 - p_- - p_+) f^c(x) \end{pmatrix} && (6) \\
\mu_1, \ldots, \mu_n &\overset{iid}{\sim} M && (7)
\end{aligned}
$$

This model takes into account the pairing of samples. In (5), we assume that the control and tumor expression measurement for the gth gene from the ith 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_g^t 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 Efron et al. (2001) and Ibrahim et al. (2002).

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 difficult complex tissue problem. We will later link this quantity to the false discovery rate (Benjamini and Hochberg, 1995).

Model Specifications

We first start by considering the unpaired study design model (1)-(4). We specify that f_g^c is the density function of a normal random variable with mean μ_{gc} and variance σ_{gc}^2 in (1) and that f_g^t is that from a normal distribution with mean μ_{gt} and variance σ_{gt}^2. In the second stage of the model (equations (2) and (4)), σ_{gt}^2 and σ_{gc}^2 are assumed to be from distributions with mean σ_t^2 and σ_c^2. We will assume that the distribution of μ_{gc} $(g = 1, \ldots, G)$ at the second stage is from a normal distribution with mean μ_c and variance σ^2. For the distribution of μ_{gt}, we will formulate the following model:

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

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:

$$
\begin{aligned}
Y_{gi}^{c} &\sim N(\mu_{gc}, \sigma_{gc}^{2}) && (9)\\
X_{gi}^{t} &\sim \pi_i N(\mu_{gt}, \sigma_{gt}^{2}) + (1-\pi_i) N(\mu_{gc}, \sigma_{gc}^{2}) && (10)\\
\mu_{gc} &\overset{iid}{\sim} N(\mu_c, \sigma_c^2) && (11)\\
\mu_{gt} &\overset{iid}{\sim} p_+ N(\mu_+, \sigma_+^2) + p_- N(\mu_-, \sigma_-^2) + (1 - p_+ - p_-) N(\mu_c, \sigma^2) && (12)\\
\sigma_{gc}^2 &\overset{iid}{\sim} F_c && (13)\\
\sigma_{gt}^2 &\overset{iid}{\sim} F_t, && (14)
\end{aligned}
$$

where F_c and F_t are the distribution functions corresponding to σ_{gc}^2 and σ_{gt}^2, 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

$$\begin{pmatrix} \sigma_{gc}^2 & \sigma_{tc} \\ \sigma_{tc} & \pi_i^2 \sigma_{gt}^2 + (1-\pi_i)^2 \sigma_{gc}^2 \end{pmatrix}.$$

For the models corresponding to (6), we assume that (μ_{gc}, μ_{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:

$$\begin{pmatrix} Y^c_{gi} \\ X^t_{gi} \end{pmatrix} |\mu_i \sim N_2 \left[\begin{pmatrix} \mu_i + \mu_{gc} \\ \mu_i + \pi_i \mu_{gt} + (1-\pi_i)\mu_{gc} \end{pmatrix}, \begin{pmatrix} \sigma^2_{gc} & \sigma_{gct} \\ \sigma_{gct} & \pi_i^2 \sigma^2_{gt} + (1-\pi_i)^2 \sigma^2_{gc} \end{pmatrix} \right] \tag{15}$$

$$\begin{aligned} \begin{pmatrix} \mu_{gc} \\ \mu_{gt} \end{pmatrix} \overset{iid}{\sim} \; & p_+ N_2 \left[\begin{pmatrix} \mu_c \\ \mu_+ \end{pmatrix}, \begin{pmatrix} \sigma^2_c & \sigma_{c+} \\ \sigma_{c+} & \sigma^2_+ \end{pmatrix} \right] \\ & + p_- N_2 \left[\begin{pmatrix} \mu_c \\ \mu_- \end{pmatrix}, \begin{pmatrix} \sigma^2_c & \sigma_{c-} \\ \sigma_{c-} & \sigma^2_- \end{pmatrix} \right] \\ & + (1 - p_- - p_+) N_2 \left[\begin{pmatrix} \mu_c \\ \mu_c \end{pmatrix}, \begin{pmatrix} \sigma^2_c & \sigma_{c0} \\ \sigma_{c0} & \sigma^2_c \end{pmatrix} \right] \end{aligned} \tag{16}$$

$$\sigma_{gct} \overset{iid}{\sim} F_{ct} \tag{17}$$

$$\sigma^2_{gc} \overset{iid}{\sim} F_c \tag{18}$$

$$\sigma^2_{gt} \overset{iid}{\sim} F_t, \tag{19}$$

$$\mu_i \overset{iid}{\sim} M \tag{20}$$

In (15),we allow for the normal and tumor components to interact; this is summarized by the covariance term σ_{gct}. Having stated the semiparametric models that we are fitting in the unpaired and paired design cases, respectively, we are now ready to describe the necessary estimation procedures. Our procedures will involve a combination of methods of moment estimators and the EM algorithm.

Estimation Procedures

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

1. Estimate μ_{gc} and σ^2_{gc} in (9) by

   $$\hat{\mu}_{gc} = n^{-1} \sum_{i=1} Y^c_{gi}$$

   and $\hat{\sigma}^2_{gc} = (n-1)^{-1} \sum_{i=1}^n (Y^c_{gi} - \hat{\mu}_{gc})^2$, $g = 1, \ldots, G$.

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

3. Estimate (μ_c, μ_+, μ_-), (p_-, p_+) and $(\sigma^2_c, \sigma^2_-, \sigma^2_+)$ 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 σ^2_{gc} and σ^2_{gt} from steps 1 and 2, $g = 1, \ldots, G$.

For the paired study design (equations (15)-(20)), the following algorithm is used:

1. Estimate μ_i by
$$\hat{\mu}_i = (2G)^{-1} \sum_{g=1}^{G} (Y^c_{gi} + X^t_{gi});$$
subtract $\hat{\mu}_i$ from the gene expression measurements for the ith individual, $i = 1, \ldots, n$.

2. Estimate μ_{gt} and σ^2_{gt} in (15) by method of moments estimation, using the estimates of μ_{gc} and μ_{gt} from step 1.

3. Estimate (μ_c, μ_+, μ_-), (p_-, p_+) and $(\sigma^2_c, \sigma^2_-, \sigma^2_+)$ by the EM algorithm where the "data" are $(\hat{\mu}_{gc}, \hat{\mu}_{gt})$, $g = 1, \ldots, G$.

4. Estimate the distribution of the variance parameters in (17)-(19) by the empirical distribution of the estimated variance components.

A general description of the EM algorithm in normal mixture models can be found in Ghosh and Chinnaiyan (2002). 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, overexpressed 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 $D = 3$.

A method for modelling cell type composition heterogeneity was proposed by Venet et al. (2001). Unlike our method, they do no make any assumptions on the number of component cell populations; they consider different numbers of components. Their goal is to attempt to reconstruct the measurement for the component cell population. In this article, we will be pursuing a different goal; finding differentially expressed genes explicitly incorporating sample heterogeneity.

False Discovery Rates: Analytical Comparisons

Before describing the application of the proposed methods to the real datasets, we briefly discuss the relationship between the methods proposed here with differential expression methods that control the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Let us return to case of $\pi_i = 1$ for $i = 1, \ldots, n$. Many methods have been developed in which univariate testing is done (i.e., G tests are performed), and the false discovery rate (Benjamini and Hochberg, 1995) is controlled. We now give a formal definition. We consider the following 2×2 contingency table:

Thus, from Table 1, of the G hypotheses being tested, for g_0 of them, the null is true. The definition of FDR as put forward by Benjamini and Hochberg (1995) is

$$FDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/R is not well-defined when $Q = 0$. Storey (2002) points out the problems with controlling this quantity and suggests use of the positive false discovery rate (pFDR), defined as

$$pFDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right].$$

Conditional on rejecting at least one hypothesis, the pFDR is defined to be the fraction of rejected hypotheses that are in truth null hypotheses. In other words, the pFDR is the rate at which discoveries are false.

Suppose we have independent test statistics $T_1, \ldots, T_G$ for testing G hypotheses. Define corresponding indicator variables $H_1, \ldots, H_G$ where $H_i = 0$ if the null hypothesis of no differential is true and $H_i = 1$ if the alternative hypothesis of differential expression is true, both for the ith gene. Note that $H_i = 1$ corresponds to $D_i = 1$ or $D_i = 2$ from the previous section. We assume that $H_1, \ldots, H_G$ are a random sample from a Bernoulli distribution where for $i = 1, \ldots, G$, $P(H_i = 0) = \pi_0$. We assume that $T_i|H_i = 0 \sim f_0$ and $T_i|H_i = 1 \sim f_1$ for densities f_0 and f_1 $(i = 1, \ldots, G)$. Suppose we use the same rejection region R for testing each of the G hypotheses. By a theorem from Storey (2002), we have that

$$\begin{aligned} pFDR(R) &= P(H=0|T \in R) \\ &= \frac{\pi_0 P(T \in R|H=0)}{P(T \in R)}. \end{aligned}$$

Thus, the interpretation of pFDR is as the posterior probability of no differential expression, conditional on the test statistic being in the rejection region. By contrast, we calculate $P(D=3)$, or equivalently, $P(H=0)$ given the full data. Thus, there is a partial conditioning in the definition of pFDR as put forward by Storey, while our posterior probability measure of differential expression is fully conditional on the observed data. However, our measure of differential expression has an interpretation similar to that of a false discovery rate.

Results

We now discuss the application of the proposed methodology to two 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 on which paired normal and cancer samples. 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 in the original dataset, downloaded from the following URL:

http://www.molbio.princeton.edu/oncology.

Based on the data from there, 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). A summary is provided in Figure 1. We find that the under the null distribution, we would expect to find approximately 18% of genes to be differentially expressed. A table listing numbers of genes called significant at various q-values is found in Table 2. The numbers of genes called significant can also be inferred using Figure 1.

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

1. Estimate μ_i by

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

   subtract $\hat{\mu}_i$ from the gene expression measurements for the ith individual, $i = 1, \ldots, n$.

2. μ_{gc}, σ_{gc}^2 and μ_{tc}, σ_{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 19% in Figure 1 to 10% in Figure 2. Based on Table 2, 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 Section 3 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 false discovery rate of 5% in the framework of Storey (2002).

In comparing the proposed method to 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 4 and 5. It is clear that there are genes that our method determines to be differentially expressed that have more conservative q-values. For example, in comparison with the first method of analysis, where π is assumed to be 1 for all tumor samples, there are 458 genes with a q-value bigger than 0.05 that are determined to have greater than 95% posterior probability of being differentially expressed.

Prostate cancer data

The second data example is from a recent prostate cancer study by Luo et al. (2001). In their experiments, prostate cancer and 9 normal samples were profiled using spotted cDNA microarrays. Note that this is an unpaired design. While there were 6500 original genes on the microarray in Luo et al. (2001), the gene data we are working with is from the study of Rhodes et al. (2002). Details on preprocessing can be found there. In addition, we excluded genes with missing values; this left a total of $G = 5364$.

Our first analysis involves treating cancer samples as being $\pi = 1$ and computing 5364 t-tests and estimating q-values by the method of Storey (2002). This analysis found no differentially expressed genes. We will instead work with the unadjusted p-value here. We apply the proposed method from **Systems and Methods** for the unpaired design. Based on the data, 11.4% of genes are estimated to be differentially overexpressed in tumor relative to normal, while 25% are determined to be underexpressed. Looking at Figure 6, the p-values and posterior probabilities are giving much different results as to which genes are differentially expressed.

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 $\mu_c = 0$ 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 nondifferential 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 nondifferentially 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

In this article, we have proposed a general model-based framework for the analysis of complex cell populations using microarray data. The methods in the paper are primarily motivated by cancer studies, where mixture of cancer and normal adjacent cells in the cancer samples are common. However, the proposed framework developed in the paper could apply to other settings in which multiple types of cell populations exist. While methods for assessing differentially expressed genes are becoming quite commonplace in the microarray literature, the issue of adjusting for mixed cell types is not as established. More generally, it is important to adjust for all important and available covariates in determining differentially expressed genes so that differences in gene expression will not be confounded.

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 modelling methods (Diebolt and Robert, 1994) for estimation.

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. He would also like to thank Ronglai Shen for bringing the dataset of Luo et al. (2001) to his attention.

Appendix

Expectation-Maximization (EM) algorithm

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denote the observations, where $\mathbf{y}_i$ is a p-dimensional vector $(i = 1, \ldots, n)$. Assume that the observed data are independent and identically distributed realizations from the density

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_n) \equiv \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i | \mu_k, \mathbf{\Sigma}_k), \qquad (A1)$$

where $\pi_k \quad (k = 1, \ldots, K)$ is the probability that an observation belongs to the kth group, and

$$f_k(\mathbf{y}_i | \mu_k, \mathbf{\Sigma}_k) \equiv |2\pi \mathbf{\Sigma}_k|^{-p/2} \exp\{-(\mathbf{y}_i - \mu_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{y}_i - \mu_k)\} \qquad (A2)$$

is a multivariate normal density with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$. Thus in (A1), we have formulated that the gene expression data arise from a mixture of K multivariate normal populations.

The distributions of the K components are fully specified by μ_k and $\mathbf{\Sigma}_k$, $k = 1, \ldots, K$. Let $\theta_k = (\mu_k, \mathbf{\Sigma}_k)$, $k = 1, \ldots, K$. We apply the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) for estimating $\theta_1, \ldots, \theta_K$ and $\pi_1, \ldots, \pi_K$. We begin by formulating the estimation problem as a missing data problem. For this setting, the complete data are $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$ $(i = 1, \ldots, n)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ is defined by

$$z_{ik} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to group } k \\ 0 & \text{otherwise.} \end{cases}$$

The $\mathbf{z}_i$ $(i = 1, \ldots, n)$ represent the cluster assignments. We assume that $\mathbf{z}_i$, $i = 1, \ldots, n$, are iid realizations from a multinomial distribution with probabilities $\pi_1, \ldots, \pi_K$ $(\sum_{k=1}^{K} \pi_k = 1)$ and that the density of $\mathbf{y}_i$ given $\mathbf{z}_i$ is

$$\prod_{i=1}^{n} f_k(\mathbf{y}_i|\theta_k)^{z_{ik}}.$$

Then it is easy to then derive the likelihood and log likelihood for the complete data:

$$L^{CD}(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k f_k(\mathbf{y}_i|\theta_k)\}^{z_{ik}},$$

and

$$l^{CD}(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\{\pi_k f_k(\mathbf{y}_i|\theta_k)\}. \qquad (A3)$$

Given estimates $\widehat{\pi}_1, \ldots, \widehat{\pi}_K$ and $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$, the E-step of the EM algorithm involves estimating $\mathbf{z}_i$ by $E[\mathbf{z}_i|\mathbf{y}_i, \widehat{\pi}_1, \ldots, \widehat{\pi}_K, \widehat{\theta}_1, \ldots, \widehat{\theta}_K]$. The estimator here has a simple form:

$$\widehat{z}_{ik} = \frac{\widehat{\pi}_k f_k(\mathbf{y}_i|\widehat{\theta}_k)}{\sum_{j=1}^{K} \widehat{\pi}_j f_j(\mathbf{y}_i|\widehat{\theta}_j)} \qquad (i = 1, \ldots, n; k = 1, \ldots, K).$$

The estimated $\mathbf{z}_i$ $(i = 1, \ldots, n)$ are then plugged into (A3), and the complete data log-likelihood is then maximized as a function of θ_k and π_k, $k = 1, \ldots, K$. This is the M-step of the EM algorithm. Estimates of π_k and θ_k $(k = 1, \ldots, K)$ are output from the M-step and are then input into the E-step. The two steps (E-step and M-step) are then iterated until convergence is reached. Many authors have shown (Wu, 1983; Boyles, 1983) that under general regularity conditions, the solution from the EM algorithm will converge to a local maximum. In practice, the results from fitting the algorithm has proven to be acceptable. One of the potential problems with the EM algorithm is its rate of convergence in practice. We used the k-means algorithm (MacQueen, 1967) for determining initial values.

References

Alizadeh, A. A, Ross, D. T., Perou, C. M. and van de Rijn, M. (2001). Towards a novel classification of human malignancies based on gene expression patterns. *J. Pathology*, **195**, 41 – 52.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering anlaysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA*, **96**, 6745 – 6750.

Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V. and Zhang, W. (2001). Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639 – 659.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.

Bissell, M. J. and Radisky, D. (2001). Putting tumours in context. *Nature Reviews Cancer*, **1**, 46 – 54.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A. and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Stat. Soc. Ser. B*, **56**, 363 – 375.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111 – 140.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Stat. Assoc.*, **96**, 1151 – 1160.

Fend, F. and Raffeld, M. (2000). Laser capture microdissection in pathology. *J. Clin. Pathol.*, **53**, 666 – 672.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275 – 286.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *J. Amer. Stat. Assoc.*, **97**, 88–99.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. (2000). Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805 – 817.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.*, **5**, 299 – 314.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673 – 679.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Amer. Stat. Assoc.*, **84**, 881 – 896.

Lonnestedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M., and Isaacs, W. B. (2001). Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Research*, **61**, 4683 – 4688.

McLachlan, G. J. and Krishnan, T. (2000) *The EM Algorithm and Extensions.* New York: John Wiley and Sons.

McLachlan, G. J. and Peel, D. (2002) *Finite Mixture Models.* New York: John Wiley and Sons.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37 – 52.

Paget, S. (1989). The distribution of secondary growths in cancer of the breast. *Cancer Metastasis Reviews*, **8**, 98 – 101.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **12**, 546 – 554.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for molecular-based classification in cancer. *J. Roy. Stat. Soc. Ser. B*, **64**, 717 – 736.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, **62**, 4427 – 4433.

Schena, M. (2000). *Microarray Biochip Technology.* Sunnyvale, CA: Eaton.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B*, **64**, 479 – 498.

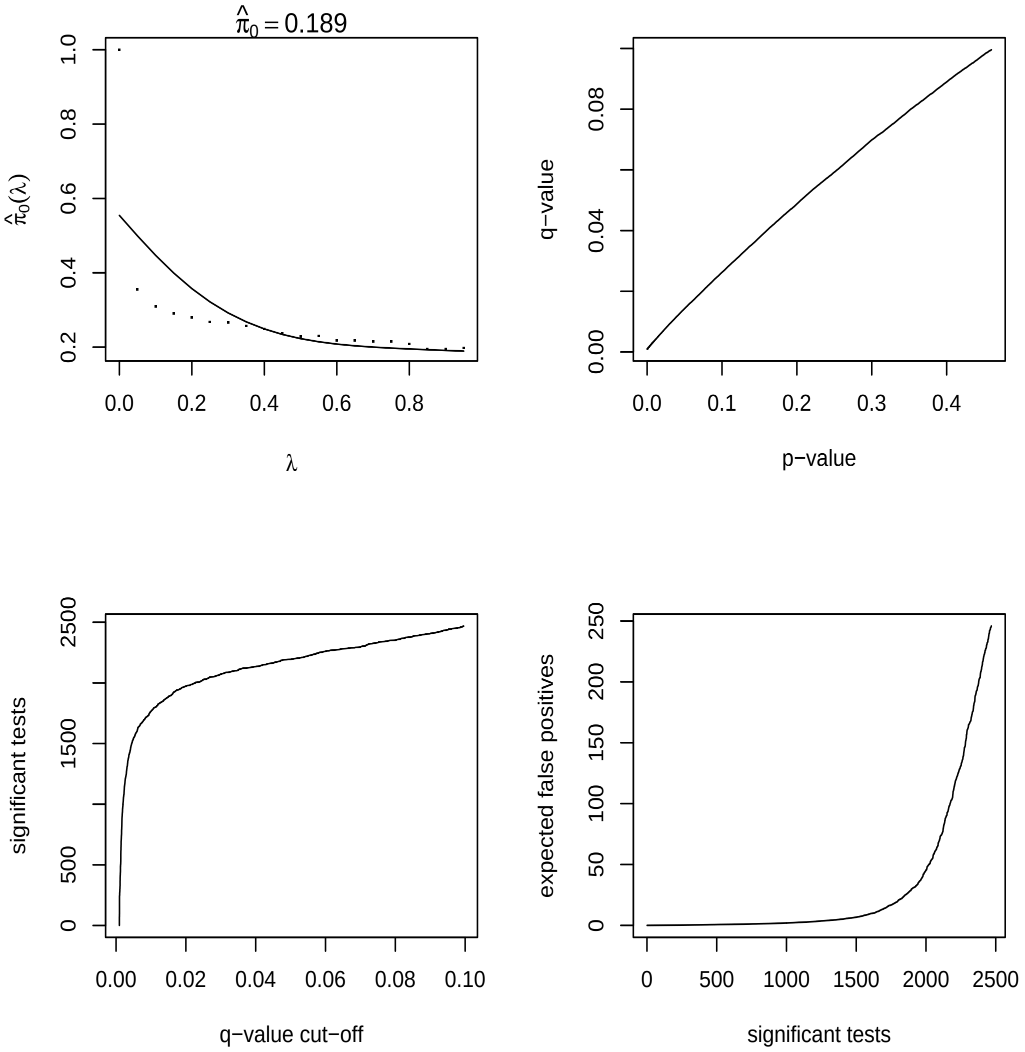

Figure 1: Output of SAM analysis (Storey, 2002) for assessing differential expression between paired normal and colon cancer tissue from Alon et al. (1999) study in which tumor is assumed to be homogeneous.

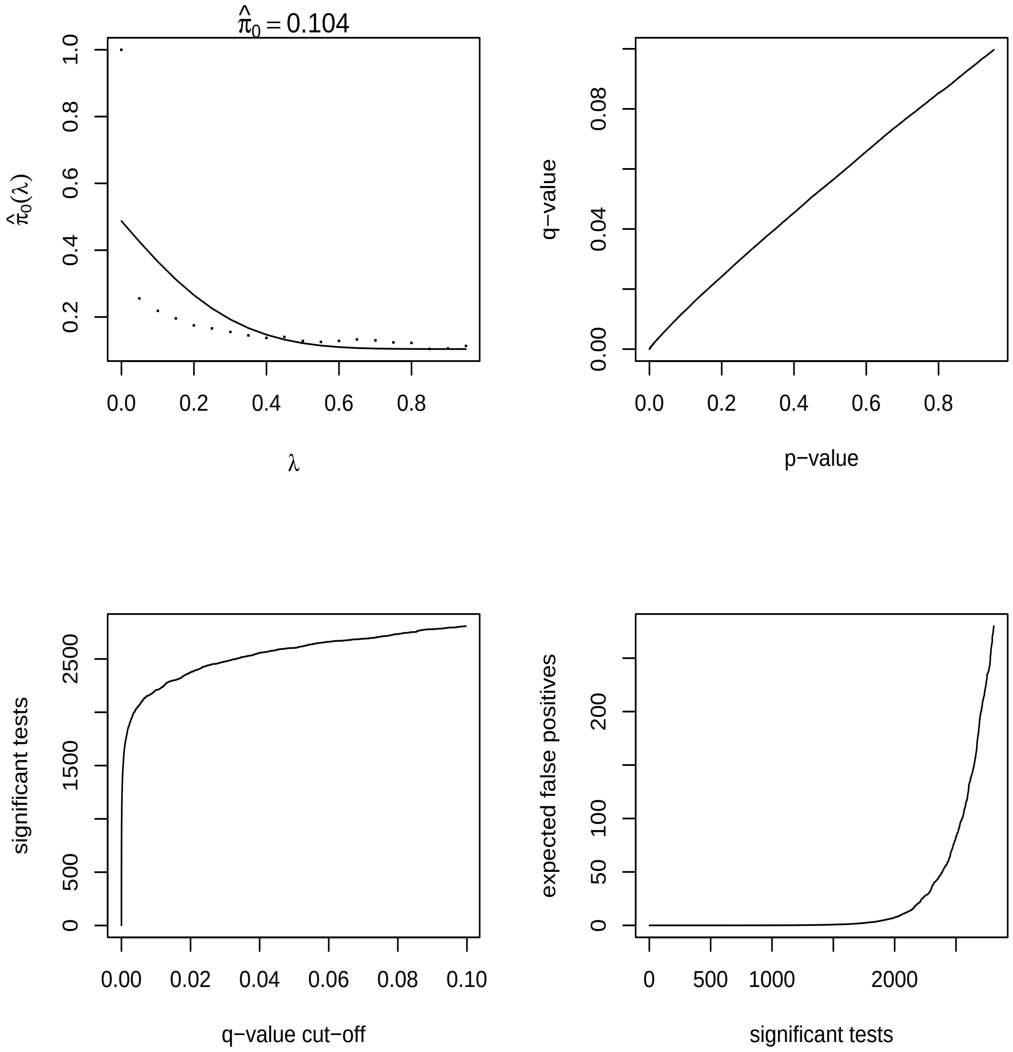

Figure 2: Output of SAM analysis (Storey, 2002) for assessing differential expression between paired normal and colon cancer tissue from Alon et al. (1999) study in which tumor is assumed to be nonhomogeneous.

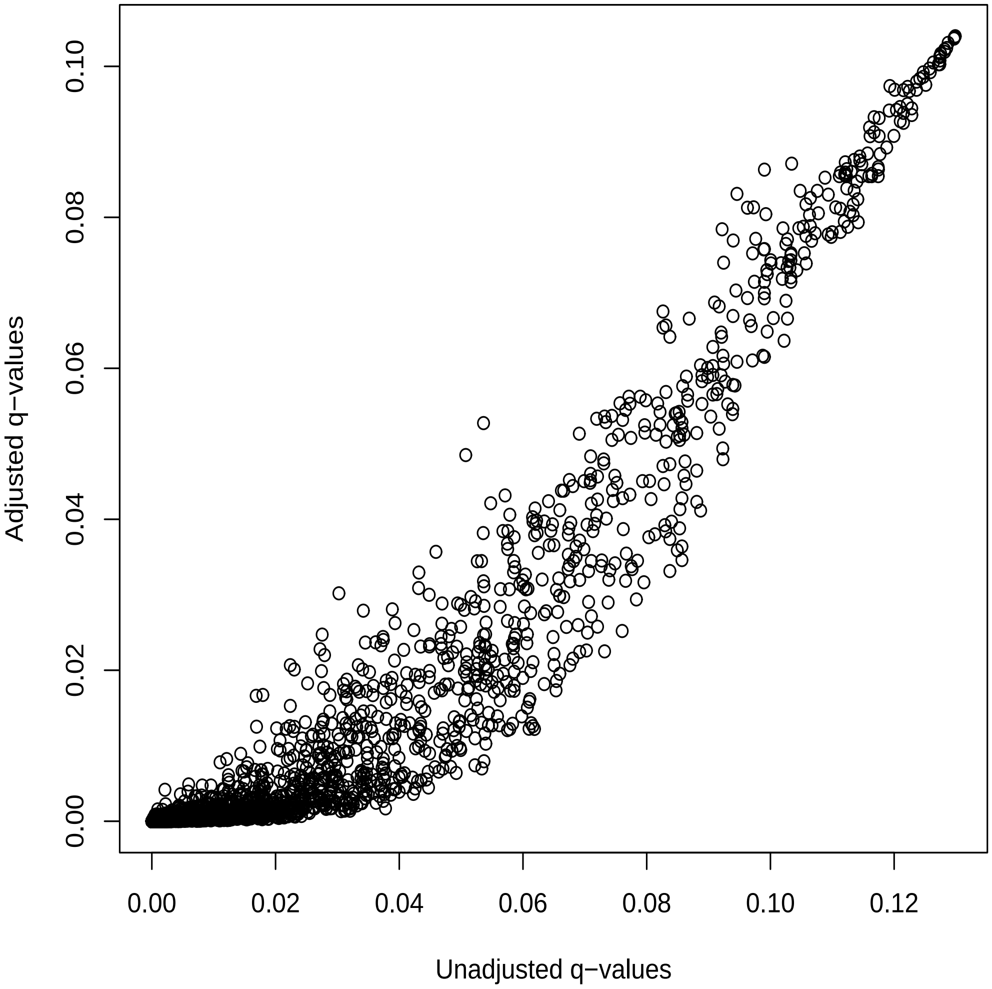

Figure 3: Comparison of unadjusted and adjusted p-values from paired colon cancer data of Alon et al. (1999).

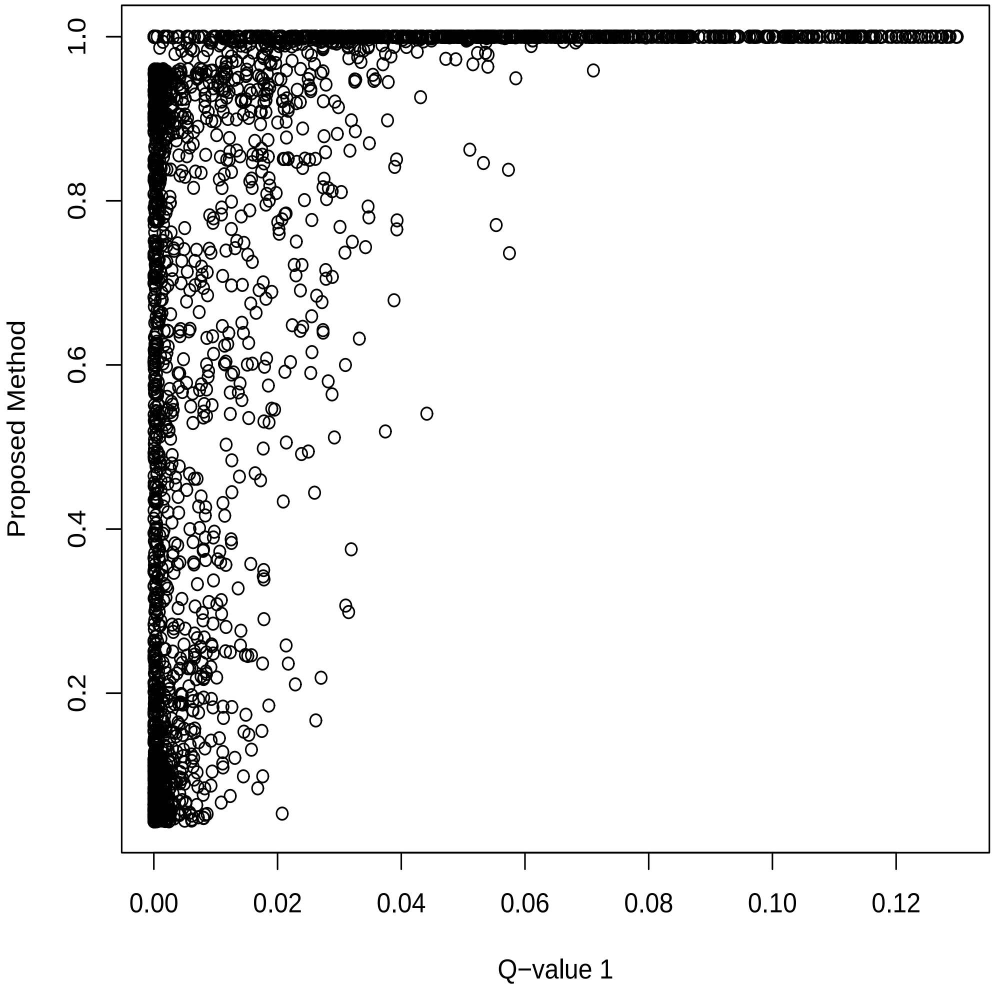

Figure 4: Comparison of estimated gene-specific false discovery rate from method in Figure 1 with posterior probability of differential expression using proposed method for colon cancer data of Alon et al. (1999).

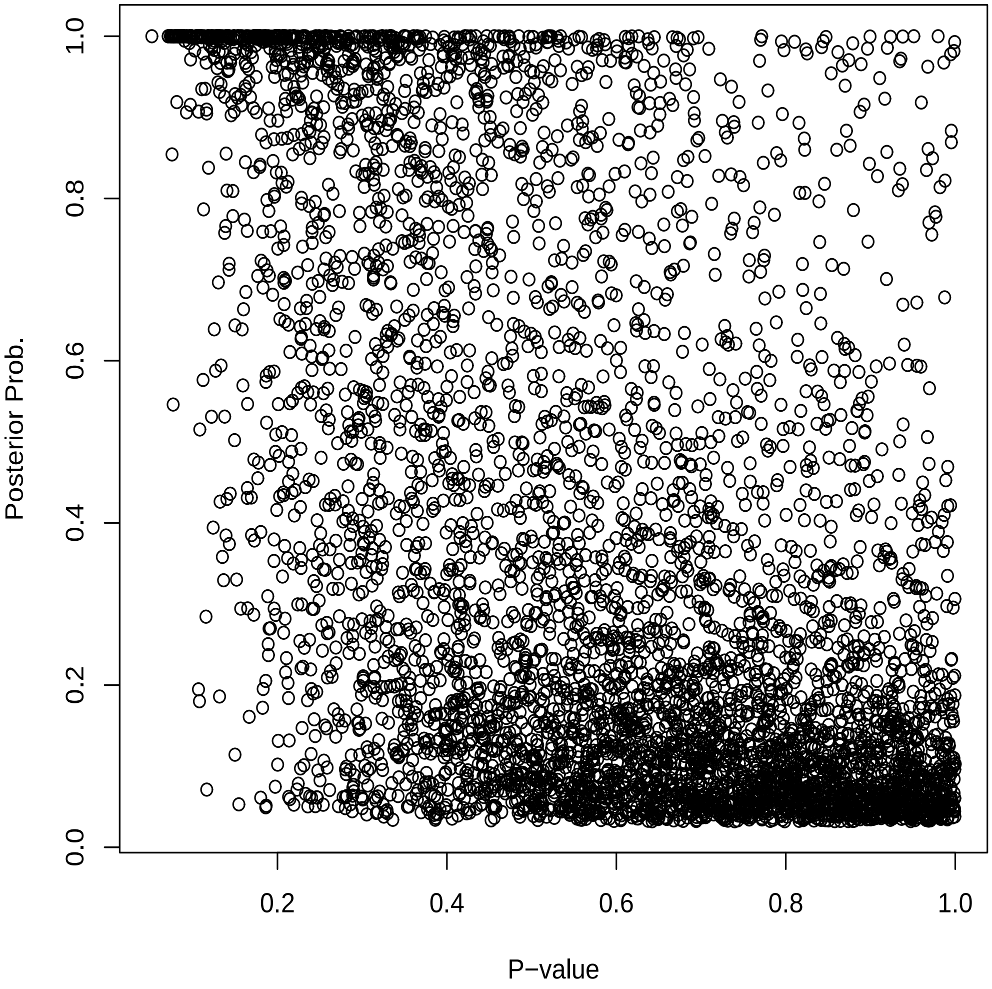

Figure 5: Comparison of estimated gene-specific false discovery rate using false-discovery rate method with posterior probability of differential expression using proposed method for unpaired prostate cancer data of Luo et al. (2001).

Table 1: Outcomes of G tests of hypotheses

	Accept	Reject	Total
True Null	U	V	g_0
True Alternative	T	S	g_1
	W	Q	G

Table 2: Number of genes called significant ($\hat{k}$) for various values of q^* based on Storey (2002) procedure for paired normal and colon cancer data from Alon et al. (1999)

q^*	No adjustment	Adjustment
0.2	2824	2824
0.1	2704	2809
0.05	2360	2604
0.02	1906	2375
0.01	1635	2210
0.005	1441	2061
0.001	966	1693

Table 3: Summary of simulation results for unpaired design. 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.

		Proposed Method			T-test method		
N	(μ_-, μ_+)	Under-sens.	Over-sens.	Spec.	Under-sens.	Over-sens.	Spec.
60	(-0.4,0.4)	0.64	0.62	0.78	0.02	0.02	0.65
60	(-0.8,0.8)	0.67	0.69	0.80	0.04	0.04	0.67
60	(-1.2,1.2)	0.71	0.71	0.83	0.06	0.06	0.74
100	(-0.4,0.4)	0.82	0.81	0.95	0.50	0.50	0.62
100	(-0.8,0.8)	0.88	0.89	0.98	0.55	0.56	0.67
100	(-1.2,1.2)	0.97	0.97	0.99	0.59	0.59	0.70

Table 4: Summary of simulation results for paired design. 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.

		Proposed Method			T-test method		
N	(μ_-, μ_+)	Under-sens.	Over-sens.	Spec.	Under-sens.	Over-sens.	Spec.
60	(-0.4,0.4)	0.54	0.52	0.68	0.12	0.11	0.65
60	(-0.8,0.8)	0.67	0.69	0.73	0.14	0.04	0.67
60	(-1.2,1.2)	0.72	0.71	0.78	0.16	0.16	0.74
100	(-0.4,0.4)	0.74	0.73	0.51	0.19	0.18	0.71
100	(-0.8,0.8)	0.77	0.79	0.55	0.26	0.24	0.77
100	(-1.2,1.2)	0.82	0.82	0.58	0.32	0.33	0.81