Funct Integr Genomics (2005) 5: 18–27
DOI 10.1007/s10142-004-0120-3

ORIGINAL PAPER

Debashis Ghosh · Arul M. Chinnaiyan

Covariate adjustment in the analysis of microarray data from clinical studies

Received: 8 December 2004 / Revised: 4 August 2004 / Accepted: 4 August 2004 / Published online: 17 September 2004

Abstract There is tremendous scientific interest in the analysis of gene expression data in clinical settings, such as oncology. In this paper, we describe the importance of adjusting for confounders and other prognostic factors in order to select for differentially expressed genes for follow-up validation studies. We develop two approaches to the analysis of microarray data in non-randomized clinical settings. The first is an extension of the current significance analysis of microarray procedures, where other covariates are taken into account. The second is a novel covariate-adjusted regression modelling based on the receiver operating characteristic (ROC) curve for the analysis of gene expression data. The ideas are illustrated using data from a prostate cancer molecular profiling study.

Keywords Differential expression · Gene expression · Multiple comparisons · Simultaneous inference

Introduction

With the advent of high-throughput gene assay technologies, scientists are now able to measure genome-wide mRNA expression levels in a variety of settings. An example of this are DNA microarrays (Schena 2000). One of the major 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-regulated or down-regulated in cancerous tissue relative to healthy tissue. Typically in these experiments, the number of genes, represented as spots on the biochip, is much larger than the number of independent samples in the study. Consequently, assessing differential expression in this setting leads to performing several thousand hypothesis tests, which leads to the problem of multiple comparisons.

Our work is motivated by a gene expression profiling study in prostate cancer. The goal is to determine if gene expression profiles can be used to classify various types of prostate cancer. In our studies (Dhanasekaran et al. 2001; Varambally et al. 2002), we have profiled tissue samples from various stages of prostate cancer (normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, advanced metastatic prostate cancer) using 10K cDNA microarrays. This gene expression database is linked to a clinical and tissue microarray database and housed in the Chinnaiyan lab at the University of Michigan. Consequently, in addition to the gene expression profiles for a sample, the investigators have access to several other clinical parameters, such as Gleason score, survival time and status, and time to PSA recurrence. Throughout the profiling studies, investigators have made the following hypotheses:

1. There exists a set of genes that distinguish lethal prostate cancer from non-lethal prostate cancer.
2. Distinct sets of genes and proteins dictate progression from precursor lesion, to localized disease, and finally to metastatic disease.

The importance of the first hypothesis is for prognostic purposes. Distinguishing indolent prostate cancer from aggressive disease will impact treatment decisions. The hypotheses address the potential of microarray technologies to develop a molecular classification for cancer that has a higher resolution than traditional histopathological staging systems. The second hypothesis is more biological in nature and is focused upon learning about which genes are involved in cancer progression. The cDNA micro-

D. Ghosh (✉)
Department of Biostatistics, School of Public Health,
University of Michigan,
Room M4057, 1420 Washington Heights,
Ann Arbor, MI, 48109-2029, USA
e-mail: ghoshd@umich.edu
Tel.: +1-734-6159824
Fax: +1-734-7632215

A. M. Chinnaiyan
Departments of Pathology and Urology, University of Michigan,
1301 Catherine Road,
Ann Arbor, MI, 48109-1063, USA

arrays also serve a screening role in that a subset of the genes that show significant differential expression will be assayed on a proteomic level using tissue microarrays (Kononen et al. 1998). Because of the fact that our gene expression database is linked to the tissue microarray database, it is relatively easy to validate the gene expression results at the protein level using tissue microarrays.

The analyses in our previous studies have focused on comparing gene expression profiles from two conditions. In the setting of a single study, differential expression for microarray data is a well-studied problem: see, for example, the work of Efron et al. (2001), Dudoit et al. (2002), Lonnestedt and Speed (2002) and Ibrahim et al. (2002). Recently, several authors have advocated use of the false discovery rate (FDR; Benjamini and Hochberg 1995) for the problem of testing multiple hypotheses simultaneously (Efron et al. 2001; Storey 2002). This quantity is different from the familywise error rate (FWER) that is typically controlled in multiple testing problems (Westfall and Young 1993).

While much of this multiple testing literature is geared towards controlling the rate of false positives using a proper calibration, in practice the outputs of these analyses are used as a screening procedure for investigators in order to find candidate biomarkers to validate on a protein level. A related argument was also given by Pepe et al. (2003), who suggested a method of ranking genes based on measures of discrimination using the receiver operating characteristic (ROC) curve.

A feature common in cancer studies is the availability of additional clinical information, such as a staging variable, survival time, baseline covariates and treatment. In virtually all of the scientific literature dealing with microarrays, the major comparison done is between cancerous and non-cancerous tissue. However, if the biomarker does not offer an improvement in terms of discriminative ability relative to the usual staging systems, then it does not serve well as a diagnostic biomarker. Potential biomarkers will be of benefit if they can provide this additional information.

More generally, the samples used for microarray profiling are typically collected in the context of an observational study. Unlike a clinical trial, where subjects are randomized to a treatment group, no such randomization assignment occurs in observational studies. In most published analyses of microarray data, the assessment of differential expression is not adjusted for potential confounders such as age and race. While this is commonly done in epidemiological studies, it has not been utilized very much in the analysis of microarray data. If the potential confounders are not adjusted for, then differences in gene expression between cancerous and non-cancerous samples will be confounded with differences between tumor sample characteristics.

In this paper, we discuss the importance of covariate adjustment in the analysis of microarray data from clinical experiments and develop some new analytical methodologies for selecting genes. The structure of this paper is as follows. In "Data description," we give a brief background on the data used to illustrate the ideas in the paper. We describe the concept of FDR and develop its link to quantities from the diagnostic testing literature in "Multiple testing procedures." In "Statistical methods," we present two methods for gene selection. The first is a covariate-adjusted FDR estimation procedure. The second is estimation of a covariate-adjusted ROC curve. The methods are applied to data from the prostate cancer study in "Results". Finally, we conclude with some brief discussion in "Discussion."

Data description

The dataset we will be using to illustrate the ideas in the paper is from an ongoing molecular profiling study in prostate cancer. The benign and malignant prostate tissues were analyzed using a 9,984-element (10K) human cDNA microarray. The glass slide cDNA microarrays developed for this study include approximately 5,000 known, named genes from the research genetics human cDNA clone set, 4,400 ESTs, and 500 control elements (which include

Table 1 Primary tumor clinical definitions (taken from American Joint Committee on Cancer 2002, p. 340)

Value	Definition
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
T1	Clinically unapparent tumor neither palpable nor visible by imaging
T1a	Tumor incidental histologic finding in 5% or less of tissue resected
T1b	Tumor incidental histologic finding in 5% or more of tissue resected
T1c	Tumor identified by needle biopsy
T2	Tumor confined within prostate
T2a	Tumor involves one-half of one lobe or less
T2b	Tumor involves more than one-half of one lobe but not both lobes
T2c	Tumor involves both lobes
T3	Tumor extends through the prostate capsule
T3a	Tumor invades seminal vesicles
T4	Tumor is fixed or invades adjacent structures other than seminal vesicles: bladder neck, external sphincter rectum, levator muscles, and/or pelvic wall

genomic human, rat, and yeast DNAs, and yeast genes). As is common with other spotted cDNA microarrays, the test and reference samples were labeled with Cy5 and Cy3 dyes and competitively hybridized to the microarray. While there are 9,984 genes on the original array, we did some preprocessing to reduce the number of genes considered. We removed genes that were missing on more than 10% of the samples as well as those having variance across all samples less than 0.05. This left a total of G =4,880 genes profiled on n =78 samples.

The data consist of (Y_{gi}, Z_i, S_i, A_i), where Y_{gi} is the gene expression measurement on the gth gene (g =1,⋯, 4,880) for the ith subject, Z_i denotes presence or absence of tumor (1 denotes tumor present, 0 denotes tumor absent), S_i is the clinical staging of the tumor from which the sample came and A_i is the age of the patient that provided the ith sample (i =1,⋯, 78). Age is included in the analysis because it is a potential confounder that we will want to adjust for in the analysis. We now describe the stage covariate further.

All tumor specimens have a mixture of cancerous and normal tissue. They are stored in paraffin-embedded blocks, a slice of which is used for the microarray experiments. The variable S refers to the stage for the specimen, but the given slice used for the microarray experiment may or may not have tumor. The variable Z refers to the slice of tissue used. Because Z and S were not perfectly correlated and S has been shown to a prognostic factor for biochemical recurrence, we decided to include it in the analysis.

The standard prostate cancer staging system is provided by the American Joint Committee on Cancer (2002, pp. 337–345). It consists of scoring the primary tumor (T), the regional lymph nodes (N) and distant metastases (M). The tumor falls into one of four stages based on the combination of T, N and M scores. In Table 1, we provide the definitions of the values of this variable.

Going back to the hypotheses formulated by the investigators, it is reasonable that biomarkers of interest should be those that have predictive power above and beyond the stage variable S. In "Statistical methods," we describe methods for adjusting for covariates in the analysis. We first describe some multiple testing methods and their links to classification ideas.

Multiple testing procedures

In most analyses of microarray data in cancer studies, the major inferential tool for assessing differential expression is using multiple testing procedures. Examples include the methods of Efron et al. (2001) and Dudoit et al. (2002).

Background

Suppose we are interested in testing a set of m hypotheses. Of these m hypotheses, suppose that for m_0 of them, the null is true. To guard against making too many type I errors, the FWER has typically been controlled. A review of methods for controlling this quantity can be found in Shaffer (1995). To better understand the FWER and FDR, we consider the 2×2 contingency table given in Table 2.

Using the definitions from Table 2, the FWER is defined to be P ($V \geq 1$), which is the probability that the number of false positives is greater than 1. The definition of FDR as put forward by Benjamini and Hochberg (1995) is

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

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

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

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 words, the pFDR is the rate at which discoveries are false. This quantity is analogous to type I error rates in single hypothesis testing problems.

The FDR and pFDR refer to one type of mistake that can be made during the hypothesis testing process. The other class of mistake that can be made is that while the alternative hypothesis is true, in practice we fail to reject the null hypothesis. This is similar to making a type II error. Thus, we define the false non-discovery rate (FNR) and positive FNR (pFNR) to be

$$\mathrm{FNR} \equiv E\left[\frac{T}{W}\middle| W > 0\right] P(W > 0)$$

and

$$\mathrm{pFNR} \equiv E\left[\frac{T}{W}\middle| W > 0\right].$$

Conditional on failing to reject at least one hypothesis, the pFNR is the fraction of accepted hypotheses that are in truth alternative hypotheses. As with pFDR, we condition on [W >0] because T/W is not well-defined when W =0. Heuristically, pFNR can be thought of as the rate at which discoveries are missed.

Suppose we have independent test statistics $T_1,\cdots, T_m$, for testing m hypotheses. Define corresponding indicator

Table 2 Outcomes of m tests of hypotheses

	Accept	Reject	Total
True null	U	V	m_0
True alternative	T	S	m_1
	W	Q	m

variables $H_1,\cdots, H_m$ where H_i =0 if the null hypothesis is true and H_i =1 if the alternative hypothesis is true. We assume that $H_1,\cdots, H_m$ are a random sample from a Bernoulli distribution where for i =1,$\cdots$, m, $P(H_i = 0) = \pi_0$. 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,$\cdots$, m). Suppose we use the same rejection region R for testing each of the m hypotheses. By a theorem from Storey (2002), we have that

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

This development has assumed that expression measurements were independent across genes. This is not very realistic, as we expect dependence between genes to occur because of involvement in common pathways. Storey (2002) mentions that FDR estimation procedures are not very sensitive to dependence among genes.

For a fixed R, we see pFDR and pFNR as being related to well-known quantities in diagnostic testing. If we define the events $[T = 1] \equiv [T \in R]$ and $[T = 0] \equiv [T \in R^c]$, then the positive and negative predictive values of T are given by $\mathrm{PPV} = P(H = 1|T = 1)$ and $\mathrm{NPV} = P(H = 0|T = 0)$. Note that we have suppressed dependence of PPV and NPV on R. Then simple algebra yields $\mathrm{pFDR}(R) = 1 - \mathrm{PPV}$ and $\mathrm{pFNR}(R) = 1 - \mathrm{NPV}$. Thus, the procedures of Storey (2002) for estimating pFDR(R) can be thought of as estimating a type of positive predictive value. This link has also been observed by Storey (2003). However, we do not know ahead of time the "diseased" and "undiseased" populations, which are the set of true alternative and null alternative hypotheses. Note that π_0 in these formulae is analogous to prevalence in PPV and NPV calculations. We have to estimate this quantity here as well. There are several approaches one can consider. In the work of Efron et al. (2001) and Storey (2001), π_0 is estimated via permutation methods. A natural Bayesian method is to place a prior on π_0; this corresponds to the biologist's knowledge as to the percentage of differentially expressed genes in the experiment. The prior for π_0 will depend very much on the setting in which the investigator is applying microarrays. For the cancer data we are considering here, the investigators expect many genes to be differentially expressed.

While the pFDR and pFNR are useful quantities for estimation in high-throughput studies, their values depend on the estimated proportion of non-differentially expressed genes, similar to the manner in which estimates of PPV and NPV depend on estimated prevalence.

Another useful quantity for discriminating between diseased and healthy populations is the ROC curve. An advantage of the ROC curve is that its estimation does not require having to estimate the proportion of non-differentially expressed genes. Suppose Y_g^D represents the gene expression measurement for the gth gene for a typical cancer specimen, i.e., D =1, and $Y_g^{\bar{D}}$ is the corresponding measurement for a randomly chosen benign specimen, i.e., D =0. Assume that higher values of Y_g correspond to having the disease. One relevant quantity is the false positive rate based on a cutoff c, defined to be $\mathrm{FP}(c) = P(Y_g > c|D = 0)$. Similarly, the true positive rate is $\mathrm{TP}(c) = P(Y_g > c|D = 1)$. The true and false positive rates can be summarized by the ROC curve, which is a graphical presentation of $\{\mathrm{TP}(c), \mathrm{FP}(c) : -\infty < c < \infty\}$. The ROC curve shows the tradeoff between increasing true positive and false positive rates. Tests that have {TP(c), FP(c)} values close to (0,1) indicate perfect discriminators, while those with {TP(c), FP(c)} values close to the 45° line in the (0,1)×(0,1) plane are tests that are unable to discriminate between the diseased and healthy populations. Examples of ideal and non-informative ROC curves are given in Fig. 1.

In the next section, we develop two procedures for assessing the explanatory power of biomarkers over and above existing clinical information. The first involves a generalization of FDR estimation procedures, while the second deals with use of the ROC curve, adjusting for covariates.

Statistical methods

Linear model-based false discovery rate

We now present a method for assessing differential expression based on direct estimation of the FDR where potential confounders are adjusted. We fit the model

$$E[Y_{ig}] = \beta_{0g} + \beta_{1g}Z_I + \beta_{2g}A_I + \beta_{3g}S_I \quad (1)$$

Our scientific focus in Eq. 1 is making inference about β_{1g}, which represents the difference in gene expression between cancerous and healthy tissue for the gth gene, adjusting for age and staging of the specimen. It is obvious

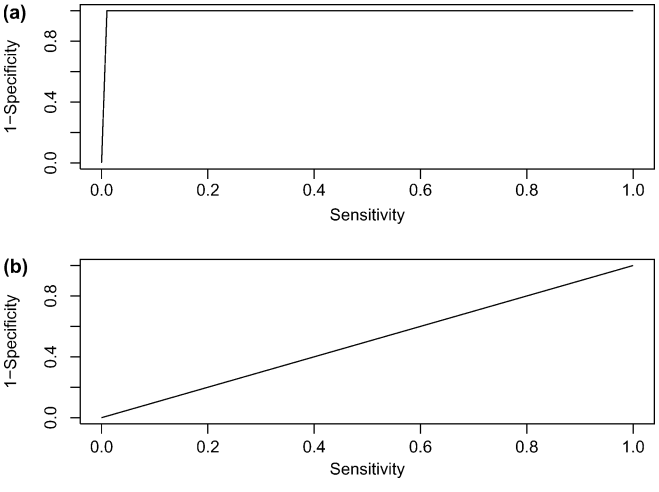

Fig. 1 Receiver operating characteristic (ROC) curves for ideal (**a**) and non-informative (**b**) tests

that fitting Eq. 1 is equivalent to fitting univariate linear models on a gene-by-gene basis. The model in Eq. 1 can be fit using ordinary least squares (OLS), yielding a set of statistics $T_{11},\cdots, T_{1G}$, where T_{1g} is the least squares estimator of β_{1g} divided by its estimated standard error. If we use a normal distribution with mean 0 and variance 1 as the null distribution for testing $H_{0g}: \beta_{1g} = 0$, then we have *Gp*-values $p_1,\cdots, p_G$. We then can apply Algorithm 1 of Storey (2002) to estimate the gene-specific FDR; it is summarized in the Appendix.

Note that because we are working with the *t*-statistics, this is a conditional method, where the conditioning is on the *t*-statistic values themselves. The unconditional method involves permuting the *Z* labels and refitting Eq. 1 to the permuted dataset. There are two potential problems with the use of the unconditional approach here. First, for large sample sizes, permutation methods on the full dataset may be quite computationally intensive. Second, the validity of permutation tests relies on the fact that under the null hypothesis of no differential expression, the distribution of permuted group assignments is exchangeable. Because we are explicitly incorporating the observational nature of the study through covariate adjustment in Eq. 1, it is not clear whether the distribution will be exchangeable. Thus the validity of permutation techniques for the gene expression data is questionable.

Receiver operating characteristic curve estimation

Pepe et al. (2003) discussed the use of ROC methods for finding candidate genes in the two-sample problem. In this section, we extend their method to allow for adjustment with covariates. Pepe (2000) showed that the ROC curve at a value t ($t\in[0,1]$) has an interpretation of $P\left(Y_g^D > Y_g^{\bar{D}} | F_{\bar{D}g}\left(Y_{\bar{D}g}\right) = t\right)$. Based on this fact, we can formulate the following approach for ranking genes based on the ROC curve.

1. Estimate the residuals from fitting Eq. 1 for each gene; this yields the estimates $\hat{e}_{ig}$, I= 1,⋯, 78, g =1,⋯, 4,880. This will determine the appropriate direction to consider for calculating the ROC curve.
2. Determine if the sample mean of $\hat{e}_{ig}$ for the cancer samples are higher or less than $\hat{e}_{ig}$ for the healthy samples.
3. Estimate the ROC curve based on the residuals.

We work by estimating the residuals based on point 2 and calculating univariate ROC curves based on them. It should be mentioned that because of the form of the model, the mean residuals will not be zero within tumor type. If we generalized Eq. 1 to allow for interactions between A and S with tissue type, then the average of residuals would be zero. Based on estimating the ROC curves for each gene, we can consider ranking the genes on several possible quantities. One is the area under the curve:

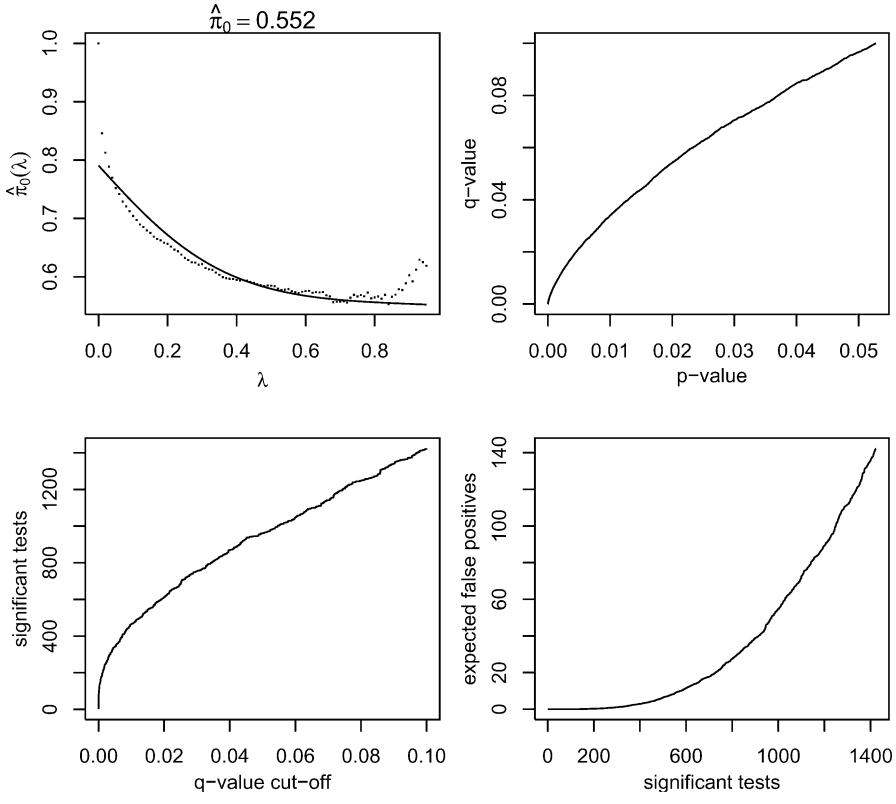

Fig. 2 Results from multiple testing analyses between cancer and non-cancer samples using the method of Storey (2002)

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)\mathrm{d}t = P\left(Y_g^D > Y_g^{\bar{D}}\right). \quad (2)$$

A second quantity is the partial area under the curve, restricted in the range from 0 to t_0:

$$\mathrm{pAUC}(t_0) = \int_0^{t_0} \mathrm{ROC}(t)\mathrm{d}t, \quad (3)$$

where t_0 is some small false positive rate. This is typically done because the value of the entire ROC curve is typically not of interest to investigators for the purposes of selecting genes. Instead, what is of interest is the value of ROC for small false positive rates, or equivalently, for small sensitivity values. Another measure that focuses the ROC curve for small false positive rates is ROC (t_0):

$$\mathrm{ROC}(t_0) = P\left(Y_g^D \geq y^C(1 - t_0)\right). \quad (4)$$

One way in which the ROC values (Eqs. 2–4) provide complementary information to that given by the p-values for β_{1g} in Eq. 2 is that ROC-based quantities summarize the potential discriminative abilities of biomarkers.

Results

We now describe the application of the proposed methodology to the prostate cancer data described in "Data description." We first began by performing a simple analysis using two-sample t-tests (i.e., comparing gene expression samples with Z =0 and Z =1) and estimating the gene-specific FDR (Storey 2002); the results are provided in Fig. 2. Based on this plot, we see that approximately 43.3% of genes are being estimated as non-differentially expressed between cancerous and healthy samples. Based on the plot, it also appears that the false positive rate is fairly low even for moderately large numbers of significance tests; this information is provided by the figure in the bottom right corner of Fig. 1.

In the next analysis, we fit Eq. 1 and estimate the gene-specific FDRs using the methods outlined in "Linear model-based false discovery rate"; the resulting plot is given in Fig. 3. We find that a greater percentage of genes are estimated as non-differentially expressed. We see that the reduction in number of genes called significant between the two analyses ranges from about 30% to 50%. This suggests that differences in gene expression from the previous analysis were partially due to differences in clinical stage and/or age of the tissue specimen. If we were to focus our attention on individual gene lists and rank the genes based on the p-values, we find that there is some overlap between the two. There is 80% overlap between the two lists for the top 20 genes. This overlap drops to 74% for the top 100 genes and 69.5% for the top 1,000 genes.

We next ranked genes based on AUC, pAUC and ROC (t_0), where we took t_0 =0.1. The lists of top 20 genes based

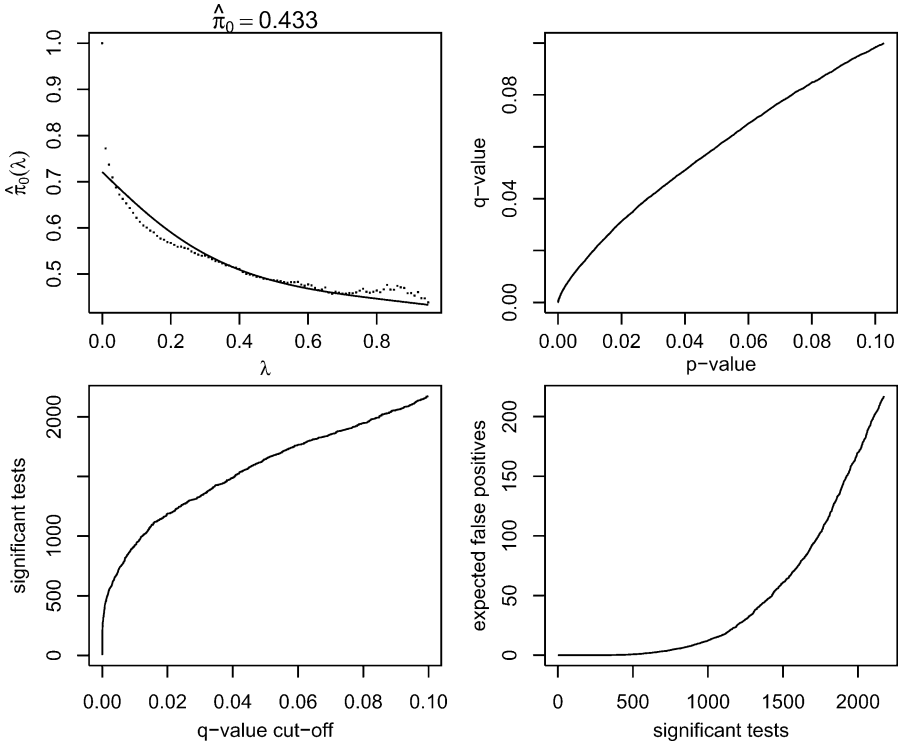

Fig. 3 Results from multiple testing analyses assessing differential expression between cancer and no cancer adjusting for age and clinical stage of tumor

Table 3 Top 20 differentially expressed genes based on adjusted q-values. The data were fit using Eq. 1 so that an adjustment for age and stage was made

Unigene ID	Gene name
Hs.85155-or-Hs.33905	Zinc finger protein 36, C3H type-like 1-or-ESTs
Hs.83951	Hermansky-Pudlak syndrome 1
Hs.139851	Caveolin 2
Hs.76252	Endothelin receptor type A
Hs.171731	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)
Hs.342874	Transforming growth factor, beta receptor III (betaglycan, 300 kDa)
Hs.352554	*Homo sapiens* cDNA FLJ31353 fis, clone MESAN2000264
Hs.343522	ATPase, Ca++ transporting, plasma membrane 4
Hs.4909	Dickkopf homolog 3 (*Xenopus laevis*)
Hs.6838	RAS homolog gene family, member E
Hs.374441	Human calmodulin-I (CALM1) mRNA, 3′UTR, partial sequence
Hs.124029	Inositol polyphosphate-5-phosphatase, 40 kDa
Hs.93005	Snail homolog 2 (*Drosophila*)
Hs.301853	RAB34, member RAS oncogene family
Hs.75350	Vinculin
Hs.342874	Transforming growth factor, beta receptor III (betaglycan, 300 kDa)
Hs.272927	Sec23 homolog A (*S. cerevisiae*)
Hs.104105	Meis1, myeloid ecotropic viral integration site 1 homolog 2 (mouse)
Hs.100554	ESTs
Hs.79339	Lectin, galactoside-binding, soluble, 3 binding protein

on these scores are given in Tables 3, 4, 5, and 6. If we compare them to those using q-values, we now find minimal overlap. In fact, there is only one gene common to both. The overlap between genes based on the ROC measures and q-values is about 5% for the top 100 and 18.5% for the top 1,000. This suggests the q-value is not adequately able to capture the discriminative power of the biomarker. Using arguments similar to those in Pepe et al. (2003), what is happening is that the linear model method is picking up the genes that show the smallest variation, while the ROC method is picking up the genes with higher discriminatory power.

Table 4 Top 20 differentially expressed genes based on AUC

Unigene ID	Gene name
Hs.351622	Pyruvate dehydrogenase complex, lipoyl-containing component X; E3-binding protein
Hs.325825	*Homo sapiens* cDNA FLJ20848 fis, clone ADKA01732
Hs.279798	Placenta-specific 7
Hs.86437	*Homo sapiens* gastric cancer-related protein GCYS-20 (gcys-20) mRNA, complete cds
Hs.170285	Nucleoporin 214 kDa (CAIN)
Hs.30120	ESTs, weakly similar to NUCL_HUMAN NUCLEOLIN [*H. sapiens*]
Hs.287820 or Hs.211579	Fibronectin 1 or melanoma cell adhesion molecule
Hs.165216	ESTs
Hs.301451	ESTs, weakly similar to GNMSLL retrovirus-related reverse transcriptase homolog—mouse retrotransposon [*M. musculus*]
Hs.2910	Phosphoribosyl pyrophosphate synthetase 2
Hs.28700	*Homo sapiens*, clone IMAGE: 3685952, mRNA
Hs.19377	ESTs, highly similar to B30142 pepsin A [*H. sapiens*]
Hs.43910	CD164 antigen, sialomucin
Hs.99423 or Hs.145020	ATP-dependent RNA helicase or ESTs, weakly similar to KIAA1205 protein [*H. sapiens*]
Hs.9663	Programmed cell death 6 interacting protein
Hs.301612	FOS-like antigen 2
Hs.70704	Chromosome 20 open reading frame 129
Hs.143434	Contactin 1
Hs.46849	ESTs
Hs.278385 or Hs.6088	ESTs or a disintegrin and metalloproteinase domain 11

Table 5 Top 20 differentially expressed genes based on pAUC (the AUC was restricted to false positive rates less than or equal to 0.1)

Unigene ID	Gene name
Hs.351622	Pyruvate dehydrogenase complex, lipoyl-containing component X; E3-binding protein
Hs.279798	Placenta-specific 7
Hs.86437	*Homo sapiens* gastric cancer-related protein GCYS-20 (gcys-20) mRNA, complete cds
Hs.170285	Nucleoporin 214 kDa (CAIN)
Hs.30120	ESTs, weakly similar to NUCL_HUMAN NUCLEOLIN [*H. sapiens*]
Hs.287820 or Hs.211579	Fibronectin 1 or melanoma cell adhesion molecule
Hs.325825	*Homo sapiens* cDNA FLJ20848 fis, clone ADKA01732
Hs.165216	ESTs
Hs.28700	*Homo sapiens*, clone IMAGE: 3685952, mRNA
Hs.9663	Programmed cell death 6 interacting protein
Hs.19377	ESTs, highly similar to B30142 pepsin A [*H. sapiens*]
Hs.301451	ESTs, weakly similar to GNMSLL retrovirus-related reverse transcriptase homolog—mouse retrotransposon [*M. musculus*]
Hs.99423 or Hs.145020	ATP-dependent RNA helicase or ESTs, weakly similar to KIAA1205 protein [*H. sapiens*]
Hs.2910	Phosphoribosyl pyrophosphate synthetase 2
Hs.143434	Contactin 1
Hs.78065	Complement component 7
Hs.21594	RAS-like, estrogen-regulated, growth-inhibitor
Hs.70704	Chromosome 20 open reading frame 129
Hs.43910	CD164 antigen, sialomucin
Hs.301612	FOS-like antigen 2

Table 6 Top 20 differentially expressed genes based on ROC (0.1)

Unigene ID	Gene name
Hs.107000	Hypothetical protein FLJ11294
Hs.13997	ESTs
Hs.288057 or Hs.351291	Hypothetical protein FLJ22242 or *Homo sapiens* cDNA FLJ32731 fis, clone TESTI2001134
Hs.348955	ESTs, weakly similar to A43932 mucin 2 precursor, intestinal [*H. sapiens*]
Hs.172788	ALEX3 protein
Hs.256398	*Homo sapiens* mRNA; cDNA DKFZp434E0528 (from clone DKFZp434E0528)
Hs.13993	TBP-like 1
Hs.107707 or Hs.78436	Mitochondrial ribosomal protein S15 or EphB1
Hs.155560	Calnexin
Hs.39982	ESTs
Hs.106728	*Homo sapiens*, clone IMAGE: 4686377, mRNA
Hs.175955	ESTs, weakly similar to hypothetical protein FLJ11267 [*Homo sapiens*]
Hs.50966	Carbamoyl-phosphate synthetase 1, mitochondrial
Hs.5822	*Homo sapiens* cDNA: FLJ22120 fis, clone HEP18874
Hs.39785	ESTs
Hs.24654	*Homo sapiens*, clone MGC: 10198 IMAGE: 3909581, mRNA, complete cds
Hs.75360	Carboxypeptidase E
Hs.194140	ESTs
Hs.74346 or Hs.144240	Hypothetical protein MGC14353 or EST
Hs.49727	ESTs

Discussion

In this manuscript, we have stressed two ideas in the analysis of microarray data. The first is adjustment for variables that might be confounders or that might increase precision of the estimated difference in gene expression between samples from different conditions (e.g., cancer versus non-cancer tissue). This is important because in most settings the samples are collected in an observational study, which is subject to various biases. In most analyses of gene expression data, confounders and precision variables are not usually taken into account in assessing differential expression. We have done this through use of model/Eq. 1. Alternative methods for achieving the adjustment include matching and stratification.

The second idea emphasized here is the use of ROC curves for gene selection in microarray experiments. ROC curves have been utilized heavily in classification problems. One of the advantages of this method is that it does not depend on the estimated proportion of non-differentially expressed genes, while the approaches of Efron et al. (2001) and Storey (2002) do. The methodology we have described here is a generalization of that given in Pepe et al. (2003).

The analytic approaches described in the paper were primarily based on Eq. 1. This model can be generalized in several ways. First, we could include additional covariates on the right-hand side of Eq. 1, along with interactions with cancer status. Second, the linear model can be extended to the class of generalized linear models (McCullagh and Nelder 1999; Fahrmeir and Tutz 2001) in a normal fashion in order to accommodate other types of clinical responses.

For example, an alternative analysis to the one proposed here would be to fit a logistic regression model where the dependent variable is presence or absence of cancer, and the predictors are the gene expression level, age, and stage. The method can then be applied to the coefficient of *Y*. This model should achieve the same end that the authors seek, namely, to account for the effects of age and stage on the discovery of biomarkers.

While we have focused on the analysis of gene expression data, it is also useful to incorporate statistical considerations in the design of gene expression experiments. It is an area we are currently pursuing.

In a recent paper, Pepe et al. (2001) provide guidelines as to the development of new biomarkers for the early detection of cancer. They argue eloquently for the use of the ROC curve. In addition, they make the case that the potential exists for confounders to obscure the relationship between candidate biomarkers with clinical outcome. This work serves as a practical implementation of some of those ideas.

Appendix: Proposed algorithm for estimating pFDR and FDR

Fit model/Eq. 1 for each gene g, $g = 1,\cdots, G$.

Calculate a p-value using $\hat{\beta}_{1g}/\widehat{SE}\left(\hat{\beta}_{1g}\right)$, $g = 1,\cdots, G$.

Let $p_1,\cdots, p_G$ denote the Gp-values. Estimate π_0, the proportion of differentially expressed genes and $F_P(x)$, the cdf of the p-values by

$$\hat{\pi}_0 = \frac{W(\lambda)}{(1-\lambda)m}$$

and

$$\hat{F}_P(x) = \frac{\min\{R(\gamma), 1\}}{G},$$

where $R(\gamma) = \#\{p_I \le \gamma\}$ and $W(\gamma) = \#\{p_I > \lambda\}$.

For any rejection region of interest $[0,\gamma]$, estimate pFDR as

$$\widehat{\mathrm{pFDR}}(\gamma) = \frac{\hat{\pi}_0\gamma}{\hat{F}_P(\gamma)\{1-(1-\gamma)^m\}}.$$

Estimate FDR as

$$\widehat{\mathrm{FDR}}(R) = \frac{\hat{\pi}_0\gamma}{\hat{F}_P(\gamma)}$$

Note: For details on choosing γ, see Section 9 of Storey (2002).

References

American Joint Committee on Cancer (2002) Cancer staging handbook, 6th edn. Springer, New York Berlin Heidelberg

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

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412:822–826

Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin 12:111–140

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Fahrmeir L, Tutz G (2001) Multivariate statistical modelling based on generalized linear models, 2nd edn. Springer, Berlin Heidelberg New York

Ibrahim JG, Chen MH, Gray RJ (2002) Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 97:88–99

Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4:844–847

Lonnestedt I, Speed TP (2002) Replicated microarray data. Stat Sin 12:31–46

McCullagh J, Nelder JA (1999) Generalized linear models, 2nd edn. Chapman & Hall, London

Pepe MS (2000) An interpretation for the ROC curve and inference using GLM procedures. Biometrics 56:352–359

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y (2001) Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 93:1054–1061

Pepe MS, Longton G, Anderson GL, Schummer M (2003) Selecting differentially expressed genes from microarray experiments. Biometrics 59:133–142

Schena M (2000) Microarray biochip technology. Eaton, Sunnyvale

Shaffer J (1995) Multiple hypothesis testing. Annu Rev Psychol 46:561–584

Storey JD (2002) A direct approach to false discovery rates. J R Stat Ser B 64:479–498

Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the *q*-value. Ann Stat 31:2013–2035

Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP, Rubin MA, Chinnaiyan AM (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419:624–629

Westfall PH, Young SS (1993) Resampling-basedmultiple testing: examples and methods for P-value adjustment. Wiley, Chichester, N.Y.