Empirical Bayes Identication of Tumor Progression Genes from Microarray Data

Debashis Ghosh[*,1] and **Arul M. Chinnaiyan**[2]

[1] Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.
[2] Departments of Pathology and Urology, University of Michigan, U.S.A.

Received 20 May 2006, revised 2 June 2006, accepted 22 September 2006

Summary

The use of microarray data has become quite commonplace in medical and scientific experiments. We focus here on microarray data generated from cancer studies. It is potentially important for the discovery of biomarkers to identify genes whose expression levels correlate with tumor progression. In this article, we propose a simple procedure for the identification of such genes, which we term tumor progression genes. The first stage involves estimation based on the proportional odds model. At the second stage, we calculate two quantities: a q-value, and a shrinkage estimator of the test statistic is constructed to adjust for the multiple testing problem. The relationship between the proposed method with the false discovery rate is studied. The proposed methods are applied to data from a prostate cancer microarray study.

Key words: Gene Expression; Metastasis; Mixture Models; Multiple Comparisons; Prostate Cancer.

1 Introduction

The use of DNA microarray technology has allowed for new understanding of various cancers. The hybridization of cDNA to arrays containing thousands of genes and ESTs permits a global genome-wide evaluation of tumor samples. This technology has has led to development of statistical methodology in various areas of microarray data analysis, such as methods for differential expression (Efron et al., 2001; Dudoit et al., 2002b), clustering (Eisen et al., 1998) and classification (Hastie et al., 2000; Dudoit et al., 2002a).

The motivating example is from a microarray experiment in prostate cancer (Dhanasekaran et al., 2001). We have profiled tissue samples from various stages of prostate cancer (e.g., normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, advanced metastatic prostate cancer). The samples are linked to a patient clinical database that has other parameters, such as Gleason score, survival time and status, and time to PSA recurrence. One of the main hypotheses of interest to scientists is that there exist distinct sets of genes and proteins dictate progression from precursor lesion, to localized disease, and finally to metastatic disease. This hypothesis is biological in nature and is focused upon learning about which genes are involved in cancer pathways. We will refer to genes satisfying this hypothesis as tumor progressor genes.

The ideal design for studying development of gene expression profiles in tumors would be a longitudinal experiment. The tumor is commonly thought to originate as a progenitor cell and goes through several stages of progression (e.g., benign hyperplasia, in situ). Such a model for tumor progression has been postulated by Fearon and Vogelstein (1990). If it were possible to sample the same

[*] Corresponding author: e-mail: ghoshd@umich.edu, Phone: 007346159824, Fax: 007347632215

tumor in these various stages of development and to generate gene expression profiles for each of the time points, then this would provide the optimal setting for studying the effect of gene expression profiles on tumor progression. While this is possible for studying tumor volume progression in mouse models (Ferrante et al., 2000), this is not feasible for humans as tumor tissue is completely resected from the patient. The data typically available are the gene expression profiles for the tumor sampled at one point of time in the tumor progression for a given patient. One could perform such an experiment using tumor cell lines, but then there is the question of whether these results would generalized to human tissue.

One can view the gene expression profile as a high-dimensional biological property of the tumor. There has been a rich literature existing on statistical models for tumor progression in which the property considered was size of the tumor (Kimmel and Flehinger, 1991; Xu and Prorok, 1997). However, no such development has occurred for gene expression profile and its effects on tumor progression. By incorporating clinical information on stage of the tumor (e.g., precursor lesion, localized prostate cancer and metastatic lesion), one can utilize microarray data potentially in a more efficient fashion. However, an important feature that must be considered is that many of the genes are noninformative about tumor progression. In this article, we seek to develop statistical methods for characterizing the relationship of gene expression profile on tumor progression. The gene expression profile is treated as the feature of the tumor that we wish to associate with clinical progression. While there have many proposals for assessing differential expressioneteps in the setting of two or more groups, there has been relatively little literature on assessing association in the situation where the phenotype is an ordinal response with more than two levels.

We develop and describe two simple methods to address this goal. The procedure involves gene-by-gene estimation using the proportional odds model (Agresti, 2002) at the first stage. At the second stage, one of two approaches are used. One is to calculate q-values (Storey, 2002; Storey and Tibshirani, 2003) based on the univariate p-values. The second is to calculate shrinkage estimators of the test statistics are constructed in order to adjust for the multiple testing problem. This is done using James-Stein type estimators; a novel consideration is that shrinkage must be done towards two targets in the current setting. The shrinkage estimators are meant to provide complementary information to the q-values; we discuss this in the context of the prostate cancer example. The eteps structure of this paper is as follows. In Section 2, we describe the data structures and the statistical procedures for analyzing the effects of gene expression on tumor stage. The methods are applied to the previously mentioned prostate cancer data in Section 3. We conclude with some discussion in Section 4.

2 Systems and Methods

2.1. Notations and preliminaries

Let D denote the stage of disease; we assume that it takes values $(1, \ldots, d)$, where increasing numbers corresponding to progressively advanced stages of disease. Thus, D will be treated as an ordinal variable here and in the sequel. We will assume that $d > 2$. Let $\boldsymbol{X}$ denote the G-dimensional gene expression profile. We observe the data $(D_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, *iid* observations from the joint distribution of $(D, \boldsymbol{X})$. In most situations we consider, G is typically much larger than n. We will assume throughout the paper that the gene expression data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ have been suitably preprocessed and normalized both within and across slides.

2.2. Proportional odds model

Define $\Pr(A)$ to be the probability of the event A. One simple model for associating gene expression with stage of disease is the proportional odds model (Agresti, 2002, § 7.2.2): for $r = 1, \ldots, d$,

$$\log\left\{\frac{\Pr(D_i \leq r)}{\Pr(D_i > r)}\right\} = \alpha_{rg} + \beta_g X_{ig}, \tag{1}$$

where $(\alpha_{0g}, \ldots, \alpha_{dg})$ are gene-specific cutpoints, β_g is a gene-specific regression coefficient, and X_{ig} is the g-th component of $\boldsymbol{X}_i$ $(i = 1, \ldots, n;\ g = 1, \ldots, G)$. Note that α_{rg} is increasing in r since $\Pr(D_i \leq r \,|\, X_{ig})$ is increasing in r. Positive values of β_g indicate that higher values of gene expression are associated with increased odds that stage D is an early stage, while negative values of β_g demonstrate the converse.

An alternative motivation of model (1) is to use a latent underlying random variable Z_{ig}, where $Z_{ig} + \beta_g X_{ig}$ has a standard logistic distribution, i.e. $\Pr(Z_{ig} - \beta_g X_{ig} \leq u) = \exp(u)/\{1 + \exp(u)\}$. Then the event $\{D_i = d\}$ corresponds to the event $\alpha_{d-1} < Z_{ig} \leq \alpha_d$. This implies that

$$\begin{aligned}\Pr(D_i \leq d) &= \Pr(Z_i \leq \alpha_d) \\ &= \Pr(Z_i + \beta_g X_{ig} \leq \alpha_d + \beta_g X_{ig}) \\ &= \frac{\exp(\alpha_d + \beta_g X_{ig})}{1 + \exp(\alpha_d + \beta_g X_{ig})}.\end{aligned}$$

The proportional odds model can be fit using many standard software packages, such as SAS or S-Plus. Here and in the sequel, we will derive procedures for making inferences using $T_g = |\hat{\beta}_g|/\widehat{\mathrm{SE}}(\hat{\beta}_g)$ $(g = 1, \ldots, G)$, the usual Wald statistic based on $\hat{\beta}_g$.

2.3. Multiple testing adjustment and false discovery rates

In most microarray studies, G is much larger than n. The model in (1) is univariate and does not incorporate information across genes. One method of doing this is to incorporate a second stage in which $T_1, \ldots, T_G$ is a random sample from a mixture distribution:

$$T_g \overset{iid}{\sim} \pi_0 F_U + (1 - \pi_0) F_V. \tag{2}$$

In model (2), π_0 represents the proportion of genes that are noninformative about tumor progression, while the remaining percentage, $1 - \pi_0$ are indicators of gene progression. F_U and F_V are the distribution functions for the noninformative and informative tumor progressor genes, respectively. We will take F_U to be a standard normal distribution, here and in the sequel.

It turns out that the model (1)–(2) has a connection with multiple testing procedures based on the false discovery rate (Benjamini and Hochberg, 1995; Storey, 2002). We consider the G univariate null hypotheses H_{0g}: $\beta_g = 0$, $g = 1, \ldots, G$. Mimicking the arguments of Theorem 1 in Storey (2002), we have that based on the two-stage model (1)–(2), the gene-specific false-discovery rate is given by $\mathrm{FDR}_g = Pr(H_0 \,|\, T_g \in R)$, where R is the rejection region for the gth test statistic T_g, $g = 1, \ldots, G$. One could then use the following algorithm for the estimation of gene-specific false-discovery rates, following Storey (2002):

Algorithm 1:

1. Fit (1) for each gene g using maximum likelihood for $g = 1, \ldots, G$.
2. Calculate a p-value using $T_g \equiv |\hat{\beta}_g|/\widehat{\mathrm{SE}}(\hat{\beta}_g)$, $g = 1, \ldots, G$.
3. Let $p_1, \ldots, p_G$ denote the G p-values. Estimate π_0, the proportion of nondifferentially expressed genes and $F_P(\gamma)$, the cdf of the p-values, by

$$\hat{\pi}_0 = \frac{\#\{p_i > \lambda\}}{(1 - \lambda)G}$$

and

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

where $R(\gamma) = \#\{p_i \leq \gamma\}$.

4. For any rejection region of interest $[0, \gamma]$, estimate the FDR as

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

In this algorithm, λ takes values between 0 and 1 and plays the role of a smoothing parameter. The smaller the value of λ, the fewer the p-values that are included in the calculation of $R(\gamma)$. This leads to decreased bias but increased variability.

There are two issues in this algorithm that need to be resolved. The first is method of calculating the p-value in step 2 of the algorithm. We use permutation methods where the sample labels $D_1, \ldots, D_n$ are permuted. Note the validity of the method depends on the assumption that under the global null hypothesis of no difference in progression groups for any of the genes, the data are exchangeable. The second issue is the choice of λ. Observe that there is a bias-variance tradeoff in the choice of λ. It turns out that the bias of π_0 is minimized when $\lambda = 1$. This leads to the following algorithm to determine π_0, described by Storey and Tibshirani (2003):

Algorithm 2:
1. Order the G p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(G)}$.
2. Construct a grid of L λ values, $\lambda_1, \ldots, \lambda_L$ and calculate

$$\hat{\pi}_0(\lambda_l) = \frac{\#\{p_j > \lambda_l\}}{G(1-\lambda_l)},$$

   $l = 1, \ldots, L$.
3. Fit a cubic smoothing spline to the values $\{\lambda_l, \hat{\pi}_0(\lambda_l)\}$, $l = 1, \ldots, L$.
4. Estimate π_0 by the interpolated value at $\lambda = 1$.

Given this algorithm, one can then estimate gene-specific q-values (Tusher et al., 2001; Storey and Tibshirani, 2003) for the individual genes. The q-value is defined to be the smallest FDR at which a hypothesis can be rejected. For the gene with the largest p-value, the q-value is given by

$$q(p_{(G)}) = \min_{t \geq p_{(G)}} \frac{\hat{\pi}_0 G t}{\#\{p_j \leq t\}} = \hat{\pi}_0 p_{(G)},$$

and for $i = G-1,\ G-2, \ldots, 1,\ q(p_{(i)}) = \min(\hat{\pi}_0 G p_{(i)}/i, \hat{p}_{(i+1)})$. This guarantees that the q-values will be monotonically increasing as a function of p-values.

2.4. Decision-theoretic estimation procedures

An alternative interpretation of the mixture distribution (2) is that it provides two targets for estimation: the target under the null hypothesis and that under the alternative hypothesis. Thus, an alternative to the q-value method of Storey (2002) is to develop shrinkage estimators for the mixture model. One can then view the shrinkage estimators as adjusted estimates of β_g on a standardized scale, accounting for the multiple comparisons issue.

Following the arguments of George (1986), a James-Stein approach involves constructing shrinkage estimators for $P(H = 0|T_i)$. We calculate for $i = 1, \ldots, G$,

$$T_i^{JS} = \pi_0(T_i) T_{0i}^{JS} + \{1 - \pi_0(T_i)\} T_{1i}^{JS}, \tag{3}$$

where

$$T_{0i}^{JS} = T_i - \left[1 \wedge \frac{(G-1)}{\sum_{i=1}^G T_i^2}\right] T_i,$$

$$T_{1i}^{JS} = T_i - \left[1 \wedge \frac{(G-1)\sigma_V^2}{\sum_{i=1}^{n}(T_i - \mu_1)^2}\right](T_i - \mu_V), \tag{4}$$

$$\pi_0(z) = \frac{\pi_0 F_U(z)}{\pi_0 F_U(z) + (1-\pi_0)F_V(z)} = \frac{\pi_0 F_U(z)}{F_T(z)}, \tag{5}$$

where μ_V and σ_V^2 are the mean and variance corresponding to F_V, and F_Z is the population cumulative distribution function for T. These adjusted statistics are shrunken statistics that account for the multiple testing problem.

In fact, there are many choices for the definition of (5). We have defined it in terms of the cumulative distribution functions for the two components of the mixture model. Suppose we consider an alternative definition for (5):

$$\tilde{\pi}_0(p) = \frac{\pi_0 f_U(p)}{\pi_0 f_U(p) + (1-\pi_0)f_V(p)}, \tag{6}$$

where f_U and f_V are the density functions for U and V. Then (6) is precisely the local false discovery rate (Efron et al., 2001) based on p. We prefer the use of (5) to (6) because of variance issues. In particular (6) will have greater variance than (5) because density estimates tend to be much more variable than those based on the cumulative distribution function.

The proposed methodology adjusts the univariate Wald statistics for the multiple testing problem by shrinkage in which the shrinkage weights (5) or (6) are data-adaptive. Here and in the sequel, we consider (5). Suppose that a large fraction of null hypotheses are true, i.e. $\pi_0 \approx 1$. Then based on (5) and (3), the statistics will be shrunk towards 0. By contrast, if a majority of the null hypotheses are false, then the adjusted statistics will be closer to the mean of the distribution of the Wald statistics under the alternative hypothesis. This shows that the methodology is data-adaptive. In addition, one can view the estimated quantities as Empirical Bayes estimators of the estimation targets in (2), similar to what has been done with the location parameter in normal probability models (Berger, 1985, Section 4.5).

Another interpretation of (3) is as a doubly-shrunken test statistic, shrunk towards each component of the mixture. This idea was originally proposed by George (1986) in the context of a normal probability model. There are several differences between his work and ours. First, we are considering a mixture model for the test statistics, which is fundamentally different from the normal model considered by George (1986). In addition, note that there are unknown population quantities in (4) and (5) that need to be estimated. George (1986) provides no estimation procedure from observed data.

Observe that (2) implies the following result for the cumulative distribution:

$$F_Z(t) = \pi_0 F_U(t) + (1-\pi_0)F_V(t). \tag{7}$$

Simple algebraic manipulation of (7) yields

$$F_V(t) = \frac{F_Z(t) - \pi_0 F_U(t)}{1-\pi_0}. \tag{8}$$

We can estimate F_Z in (8) using the empirical distribution function of the observed statistics. Provided we have an estimator of π_0, we can then estimate F_V and subsequently the mean and variance in (4). Thus, the outstanding issue becomes one of estimating π_0. We use the estimator described in Algorithm 2.

3 Prostate Cancer Data

The dataset we will be using to illustrate the ideas in the paper is from a molecular profiling study in prostate cancer (Dhanasekaran et al., 2001). The benign and malignant prostate tissues were analyzed

using a 9984 element (10 K) human cDNA microarray. A two-channel (Cy5/Cy3) scheme was utilized. While there are 9984 genes on the original array and 101 samples from $d = 3$ tumor classes: benign precursor, localized prostate cancer and metastatic prostate cancer. We did some preprocessing to reduce the number of genes considered; namely, we filtered out genes that are reported as missing in more than 10% of the samples. This left a total of $G = 7910$ genes for analysis. We first performed the analysis based on fitting the proportional odds model. A spreadsheet containing gene names, estimated regression coefficients and associated Wald statistics can be downloaded as the file from the following website:

http://www.sph.umich.edu/∼ghoshd/COMPBIO/TPROG/pgenes1.csv.

In addition, a histogram of the t-statistics is given in Figure 1.

Based on permutation methods, we calculated p-values and then applied the false discovery rate estimation procedure of Storey and Tibshirani (2003). The results are summarized in Figure 2. Based on the graphs, 1582 genes have q-values less than 0.001, 753 have those less than 0.0001, and 313 less than 0.00001. One point of note is that the proportion of nondifferentially expressed genes is only 0.3. This is relatively low but is consistent with estimates we have found on other cancer datasets, in our experience. The main reason we think that the estimate is so low is because of confounding of the disease comparison with other factors.

In studies such as these, investigators are typically interested in developing a gene list of candidate biomarkers that they would be interested in performing further validation analyses, such as immunohistochemistry or quantitative RT-PCR. We focus on two results from such an analysis. The first is the identification of homologs of mammalian transcription factors. Among the top 100 genes are a homolog of a yeast transcription factor (Sec23 – Hs. 753381), a homolog of the FAT tumor suppressor in Drosophila (Hs. 591266), a homolog of a transcription factor in Xenopus laevis (Hs. 760299), and another Drosophila transcription factor homolog, frizzled, (Hs. 298122). Given the recent discovery

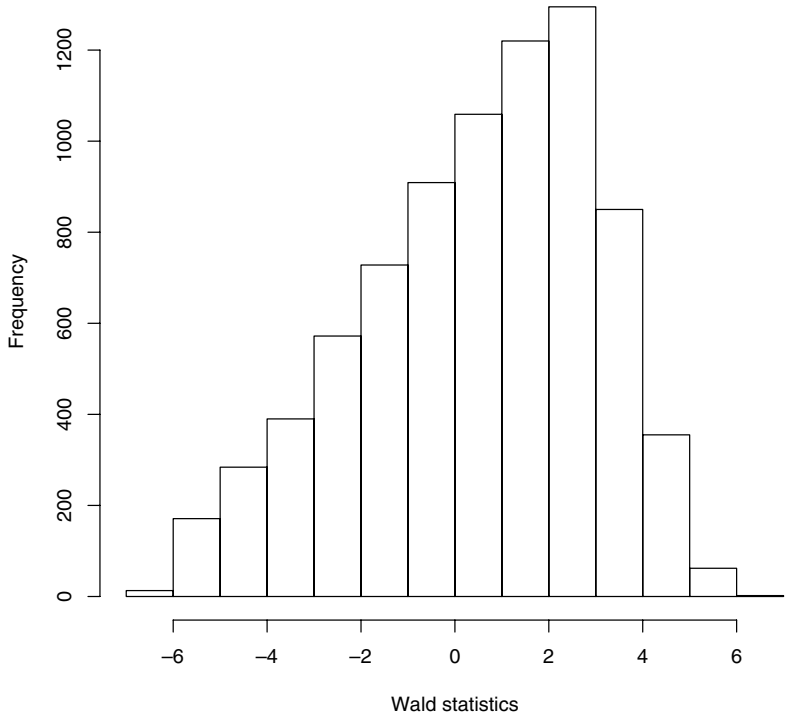

Figure 1 Histogram of Wald statistics for prostate cancer gene expression data.

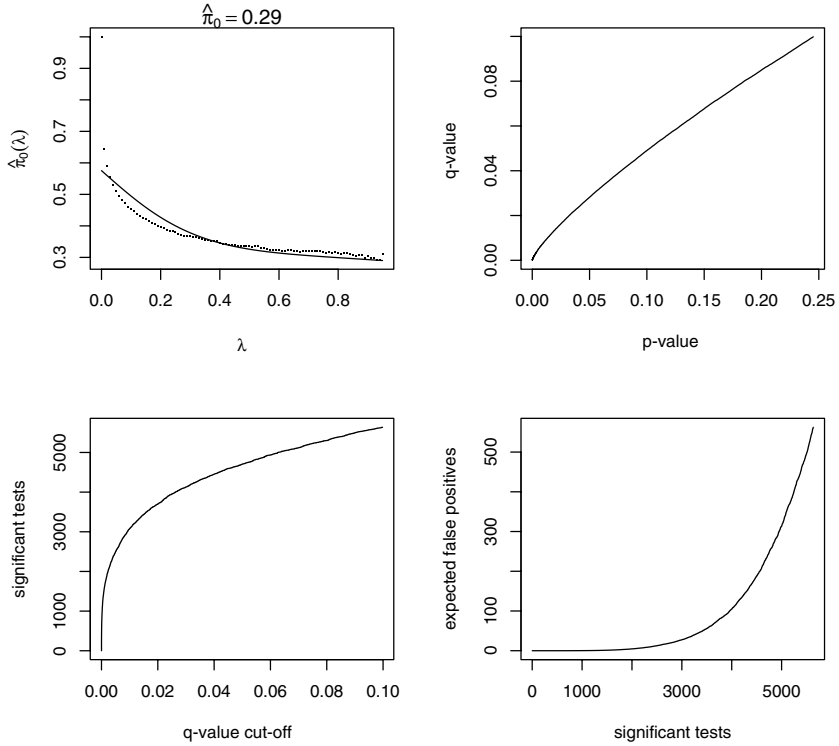

Figure 2 Output of proportional odds method combined with false discovery rate estimation procedures. The plot in the upper left-hand corner shows the estimated false discovery rate using the method of Storey and Tibshirani (2003). The upper right-hand plot shows the conversion of p-values to q-values as discussed in the **Proposed Methods and Results** section. The graph on the lower left-hand side shows the number of significant tests as a function of q-value cut-off. The lower right-hand graph displays the expected false positives as a function of number of significant tests; the estimated false discovery rate is the ratio of these quantities.

(Varambally et al., 2002) of a prostate cancer biomarker that is a homolog of a Drosophila transcription factor, EZH2, these genes are of interest to the investigator to identify other homologs of mammalian transcription factors that might be involved in cancer dysregulation.

Another finding is the decreased expression of cell surface and cell adhesion genes and products in the top 200 list. This includes genes such as catenin (Hs. 364921), moesin (Hs. 131362), integrin (Hs. 502527), and integrin, beta 1 (Hs. 343072). The transformation of a cancer from nonmetastatic to metastatic involves the development of a motile phenotype that allows the cancer cells to migrate from the site of origin to other sites. This involves step such as the epithelial-mesenchymal transition, cellular transformation and loss of cell adhesion to epithelial cells. The results of the analysis here are consistent with such a conceptual model for tumorigenesis. A caveat of these results is that they are not confirmatory but rather hypothesis-generating. Further computational and/or biological experiments would be needed to validate these findings.

Next, we applied the proposed Empirical Bayes method for calculating doubly shrunken Wald statistics in the analysis of the microarray data. A plot comparing their values to those of the original statistics is given in Figure 3. We find several interesting results. First, the scale of significance is drastically shrunk using the adjusted scale (vertical axis) versus the original scale (horizontal axis).

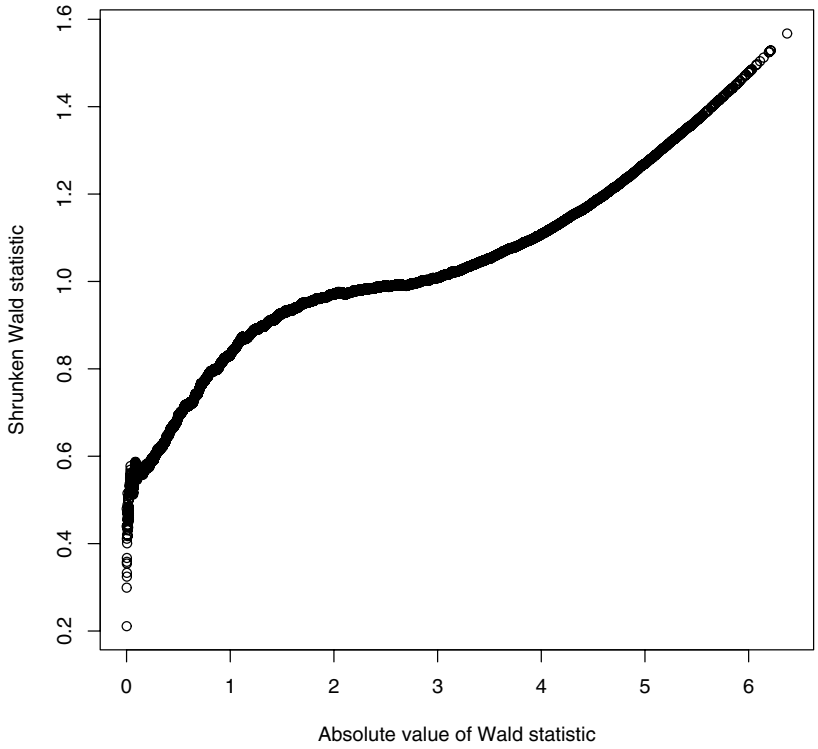

Figure 3 Plot of shrunken Wald statistics (y-axis) against absolute value of Wald statistics (x-axis) for prostate cancer gene expression data.

This serves to illustrate the fact that shrinkage leads to significant downweighting of the strength of evidence in the multiple testing situation. One can imagine combining the q-value procedure with the double shrinkage estimators proposed here in the following procedure:

1. Select all genes that have a q-value less than some cut-off;
2. Report the double shrunken estimators as estimates of the standardized effect size for those genes selected in 1.

Because the q-values have an Empirical Bayes interpretation, as do the double shrinkage estimators, one can view this as an approximate Bayesian testing procedure. The sampling distribution of the test statistics does not depend on the stopping rule specified in 1.

Finally, some mention of costs is warranted. The use of FDR is appropriate if the cost of a false discovery is proportional to the proportion of false discoveries. However, if the cost of false discoveries is proportional to the total number of false discoveries, then one should seek to have control of the generalized familywise error rate (Hommel and Hoffman, 1988; Lehmann and Romano, 2005).

4 Discussion

In this article, we have described a simple approach using Empirical Bayes methods for identifying genes that associate with tumor progression in cancer studies using microarray data. These techniques complement the multiple testing methods currently available for microarray data (Efron et al., 2001; Dudoit et al., 2002b).

The estimation procedure at the first step involves use of a proportional odds model. However, any other model for ordinal categorical data (Agresti, 2002) could also be used there. For the estimation

of the doubly shrunk Wald statistics, we used the estimator of π_0 from the algorithm of Storey and Tibshirani (2003). However, there are several other potential estimators of the proportion of null hypotheses that could be used as well (Pounds and Cheng, 2004; Dalmasso et al., 2005).

There are two crucial assumptions embodied in the mixture model posited for the test statistics in (2). The first assumption in the methods developed so far is that there is no confounding of gene expression by other clinical factors. One could imagine that sample characteristics, such as age of the patient or tissue heterogeneity, could confound the association between gene expression and tumor progression. If these characteristics have been measured, then we would extend the proportional odds model approach by including them as covariates. As discussed in Ghosh and Chinnaiyan (2005), the validity of *p*-values derived from permutation testing in this setting would be questionable. Another reason that the *p*-values might not be quite correct is that the permutation assumes all genes have no differential expression, while in actuality, the model (2) assumes that some fraction of hypotheses testing are in actuality not true. Thus, the permutation procedure would work better if we were to screen out differentially expressed genes. This idea has been suggestetepsed by several authors, among them Xie et al. (2005) and Scheid and Spang (2006).

The second assumption deals with the choice of the null distribution. We assume that the component corresponding to the null hypotheses is a normal distribution with mean zero and variance one. This might not be the right null to use (Efron, 2004). We are currently researching the use of alternative distributions for the null hypothesis.

We have attempted to treat gene expression as a biological property of the tumor sample and correlate it with tumor progression. An alternative approach, not considered in this paper, would be to formulate a stochastic modelling approach in which a mechanistic model for gene expression development is postulated. This has precedents in the mathematical modelling literature (Yakovlev and Tsodikov, 1996).

Acknowledgements The research of the first author is supported by grant GM72007 from the Joint NSF/NIGMS Biological Mathematics Program. He would like the editor and two referees for suggesting improvements to the manuscript.

References

Agresti, A. A. (2002). *Categorical Data Analysis, 2nd edition.* Wiley, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

Dalmasso, C., Broët, P., and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics* **21**, 660–668.

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.

Dudoit, S., Fridyland, J. F., and Speed, T. P. (2002). Comparison of discrimination methods for tumor classification based on microarray data. *Journal of the American Statistical Association* **97**, 77–87.

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. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96–104.

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

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Fearon, E. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767.

Ferrante, L., Bompadre, S., Possati, L., and Leone, L. (2000). Parameter estimation in a Gompertzian stochastic model for tumor growth. *Biometrics* **56**, 1076–1081.

George, E. I. (1986). Minimax multiple shrinkage estimation. *Annals of Statistics* **14**, 188–205.

Ghosh, D. and Chinnaiyan, A. M. (2005). Covariate adjustment in the analysis of microarray data from clinical studies. *Functional and Integrative Genomics* **5**, 18–27.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, RESEARCH0003.

Hommel, G. and Hoffman, T. (1988). Controlled uncertainty. In: Bauer, P. et al. (eds.) Multiple Hypotheses Testing, pages 154–161. Springer, Berlin.

Kimmel, M. and Flehinger, B. J. (1991). Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* **47**, 987–1004.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics* **33**, 1138–1154.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20**, 1737–1745.

Scheid, S. and Spang, R. (2006). Permutation filtering: A novel concept for significance analysis of large-scale genomic data. In: Apostolico, A. et al. (eds.) Research in Computational Molecular Biology: 10th Annual International Conference, Proceedings of RECOMB 2006, pages 338–347. Springer, Heidelberg.

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

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to ionization radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G. A. B., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629.

Xie, Y., Pan, W., and Khodursky, A. B. (2005). A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* **21**, 4280–4288.

Xu, J. L. and Prorok, P. C. (1997). Nonparametric estimation of solid cancer size at metastasis and probability of presenting with metastasis at detection. *Biometrics* **53**, 579–591.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific Press, Singapore.

