University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2005 *Paper* 55

Shrunken p-values for assessing differential expression, with applications to genomic data analysis

Debashis Ghosh*

*University of Michigan, ghoshd@psu.edu

Shrunken p-values for assessing differential expression, with applications to genomic data analysis

Debashis Ghosh

Abstract

n many scientific problems involving high-throughput technology, inference must be made involving several hundreds or thousands of hypotheses. Recent attention has focused on how to address the multiple testing issue; much focus has been devoted towards use of the false discovery rate. In this article, we consider an alternative estimation procedure titled shrunken p-values for assessing differential expression (SPADE). The estimators are motivated by risk considerations from decision theory and lead to a completely new method for adjustment in the multiple testing problem. Some theoretical results are outlined. The proposed methodology is illustrated using simulation studies and with application to data from a prostate cancer gene expression profiling study.

Shrunken p-values for assessing differential expression, with applications to genomic data analysis

Debashis Ghosh

Department of Biostatistics, University of Michigan

1420 Washington Heights

Ann Arbor, MI 48109-2029

Abstract

In many scientific problems involving high-throughput technology, inference must be made involving several hundreds or thousands of hypotheses. Recent attention has focused on how to address the multiple testing issue; much focus has been devoted towards use of the false discovery rate. In this article, we consider an alternative estimation procedure titled shrunken p-values for assessing differential expression (SPADE). The estimators are motivated by risk considerations from decision theory and lead to a completely new method for adjustment in the multiple testing problem. Some theoretical results are outlined. The proposed methodology is illustrated using simulation studies and with application to data from a prostate cancer gene expression profiling study.

Keywords: Gene Expression; Hypothesis Testing; Microarray; Multiple Comparisons; Shrinkage Estimators; Simultaneous Inference.

1. Introduction

Because of technological developments in scientific fields such as genomics, it has become possible to simultaneously assay the biological activities of thousands of genes in parallel. Similarly, in neuroimaging, there is consideration of thousands of voxels as a global map of the human brain. A common problem in this setting is to determine which objects are differentially expressed between two conditions (genes in the microarray setting, voxels in the neuroimaging example). Consideration of all the hypotheses leads to a multiple comparisons problem.

Our work is motivated by a collaborative gene expression profiling study in prostate cancer (Dhanasekaran et al., 2001). The investigators have profiled tissue samples from various stages of prostate cancer (normal adjacent prostate, benign prostatic hyperplasia, localized prostate cancer, advanced metastatic prostate cancer) using microarrays. In addition to the gene expression profiles for a sample, the investigators have access to several other clinical parameters. A key hypothesis made by investigators is that there exists a set of genes that distinguish aggressive prostate cancer from non-lethal prostate cancer. To begin to address this, a fairly standard analysis would be to determine which genes are differentially expressed between aggressive prostate cancer from nonaggressive prostate cancer. Here, we will focus on finding genes that are differentially expressed between metastatic prostate cancer (i.e., cancer that has spread to other organ sites) versus localized prostate cancer.

Methods for dealing with differential expression in the multiple testing setting have been the subject of much research interest in the recent statistical literature. Methods for controlling the familywise error rate (FWER) and related quantities have been proposed by Ge et al. (2003) and by Van der Laan and Dudoit in a series of papers (Dudoit et al., 2004; van der Laan et al., 2004a,b). Several authors have argued that control of the FWER is too stringent and have advocated use of the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995). Methods for both controlling the false discovery rate as well as estimating it directly have appeared in the recent statistical literature (e.g., Efron et al., 2001;

Benjamini and Yekutieli, 2001; Sarkar, 2002; Storey, 2002).

As noted by Storey (2002), there is an explicit mixture model for the distribution of marginal test statistics from which the positive false discovery rate (Storey, 2003) and the false discovery rate can be estimated. In this paper, we consider an alternative statistical method that can be motivated from the same mixture model for dealing with multiple testing. The procedures proposed in this paper, termed shrunken p-values for the assessment of differential expression (SPADE), have links to decision-theoretic considerations (Brown, 1971; Stein, 1981; Lehmann and Casella, 2002). The structure of the paper is as follows. A definition of false discovery rate in the multiple testing situation, along with previous work, is reviewed in Section 2. In Section 3, we outline the SPADE method, which effectively boils down to calculation of James-Stein type, or shrinkage, estimators for the p-values. A key notion here is shrinkage towards the two components of the mixture model. While such an idea has been pursued by George (1986) for data from a normal mixture model, he did not consider the case of a mixture model for p-values, nor did he give approaches for estimation with observed data. Practical implementation of the SPADE methodology are discussed in Section 4. Numerical comparisons with simulated and real data are made in Section 5. We conclude with a brief discussion in Section 6.

2. False Discovery Rate: Background

Our setup is that we have test statistics $T_1, \ldots, T_n$ for testing hypotheses H_{0i}, $i = 1, \ldots, n$. Of the n hypotheses, suppose that for n_0 of them, the null is true. We will assume that the test statistics are independent. Using the following 2×2 contingency table, we can categorize hypotheses by whether they are true or not and whether or not we reject or fail to reject them:

[Note: Table 1 about here.]

Benjamini and Hochberg (1995) proposed a method for controlling the so-called false

discovery rate, defined as

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

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well-defined when $Q = 0$. Several authors have developed single-step methods for controlling the false discovery rate (Benjamini and Hochberg, 1995; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001, Sarkar, 2002).

An alternative approach that has been taken in the recent statistical literature is to fix a rejection region and to estimate FDR. Storey (2002, 2003) considers a mixture model for motivating the false discovery rate. Define indicator variables $H_1, \ldots, H_n$, corresponding to $T_1, \ldots, T_n$, where $H_i = 0$ if the ith null hypothesis is true and $H_i = 1$ if the ith alternative hypothesis is true $(i = 1, \ldots, n)$. $H_1, \ldots, H_n$ are a random sample from a Bernoulli distribution where $P(H_i = 0) = \pi_0$, $i = 1, \ldots, n$. If f_0 and and f_1 correspond to the densities to $T_i|H_i = 0$ and $T_i|H_i = 1$ $(i = 1, \ldots, n)$, respectively, the density corresponding to the marginal distribution of test statistics $T_1, \ldots, T_n$ is

$$f(t) \equiv \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{1}$$

Methods for false discovery rate estimation based on (1) have been developed by several authors (Efron et al., 2001; Storey, 2002; Pounds and Cheng, 2004, Dalmasso et al., 2005). While we assume here that the test statistics are independent, several authors (Storey et al., 2004; Genovese and Wasserman, 2004) have shown that estimates of FDR are fairly robust to various forms of dependence.

A related quantity to FDR is the positive false discovery rate (pFDR) (Storey, 2003), defined as $pFDR = FDR/P(Q > 0)$. For the multiple testing context, an analogous quantity to p-values based on pFDR, proposed by Storey (2002), is the q-value. When inference is performed using p-values, rejection regions for the null hypothesis are intervals of the form $[0, c]$, where $0 < c < 1$. The q-value corresponding to a given p-value p is defined

as

$$q(p) = \inf_{c \geq p} pFDR(\gamma) = \inf_{c \geq p} \left\{ \frac{\pi_0 c}{P(U \leq c)} \right\},$$

where U is the distribution of the p-value. This corresponds to equation (21) in Storey (2002). Q-values are tailored to the multiple comparisons problem, and their use is much like that of the p-value. Smaller q-values correspond to greater evidence against a null hypothesis.

3. SPADE: Proposed methodology

The starting point for our methodology is (1). Observe that (1) specifies a model for the test statistics, which we are assuming to be independent. Even though the test statistics are derived variables, previous authors (Storey, 2002; Genovese and Wasserman, 2002) treat them as if they are random variables with no associated variability. In the original paper by Storey (2002), the test statistics used for testing the hypotheses $H_1, \ldots, H_n$ were the p-values. The false discovery rate was estimated on the basis of the p-values and estimating π_0, the proportion of true null hypotheses, using a permutation scheme.

To motivate our methodology, we need to think of the population quantities being estimated by the test statistics. This is also the tack that is taken in decision theory (Ferguson, 1967). It is a bit more problematic to come up with a population "parameter" estimated by a p-value. We follow the approach of Hwang et al. (1990) and consider the p-value to be an estimator of the probability of the null hypothesis. Equivalently, we can consider estimators of the expected value for the indicator function corresponding to the null hypothesis being true. If we let $p_1, \ldots, p_n$ denote the p-values for testing $H_{01}, \ldots, H_{0n}$, then the model induced by (1) is

$$p_1, \ldots, p_n \overset{iid}{\sim} \pi_0 F_U + (1 - \pi_0) F_V, \tag{2}$$

where F_U is the cdf of a $U \equiv \text{Uniform}(0, 1)$ random variable and F_V is that of a random variable stochastically smaller than U.

Following the arguments of George (1986), a James-Stein approach to constructing shrinkage estimators for $I(\mu_i \in \Theta)$ is to calculate for $i = 1, \ldots, n$,

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

where

$$p_{0i}^{JS} = p_i - \left[1 \wedge \frac{(n-1)}{12 \sum_{i=1}^{n} (p_i - 1/2)^2}\right] (p_i - 1/2), \tag{4}$$

$$p_{1i}^{JS} = p_i - \left[1 \wedge \frac{(n-1)\sigma_1^2}{\sum_{i=1}^{n} (p_i - \mu_1)^2}\right] (p_i - d), \tag{5}$$

$$\pi_0(t) = \frac{\pi_0 p}{\pi_0 p + (1 - \pi_0) F_V(p)}, \tag{6}$$

where μ_1 and σ_1^2 are the mean and variance of V and $(p_1, \ldots, p_n)$. Note that the 1/2 and 1/12 refer to the mean and variance of a Uniform(0,1) distribution. These adjusted p-values are shrunken p-values that account for the multiple testing problem. This describes the essence of the SPADE methodology. Note that the mixture distribution of the p-values is providing two targets for shrinkage.

In fact, there are many choices for the definition of (6). 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 (6):

$$\tilde{\pi}_0(t) = \frac{\pi_0}{\pi_0 + (1 - \pi_0) f_V(p)}, \tag{7}$$

where f_V is the density function for V. Then (7) is precisely the local false discovery rate (Efron and Tibshrani, 2002) based on the p-value. We prefer the use of (6) to (7) because of variance issues. In particular (7) will have greater variance than (6) because density estimates tend to be much more variable than those based on the cumulative distribution function.

Another interpretation of (3) is as a doubly-shrunken p-value, 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 p-values, which is fundamentally different from

the normal model considered by George (1986). In addition, note that there are unknown population quantities in (5) and (6) that need to be estimated. George (1986) provides no estimation procedure from observed data. Estimation methodologies will be dealt with in Section 4.2.

4. SPADE: Theory and Implementation

In this section, we outline some informal theoretical results of the procedure, as well as provide some details on how to implement the method in practice.

4.1. Theoretical motivation

The results in this section pertain to minimaxity of the adjusted p-value estimates. We first start by considering the case of $n = 1$ hypothesis. The following results summarize the work in Hwang et al. (1990):

Lemma 1: *For the problem of testing H_0 versus H_1 where the null and alternative hypotheses are one-sided, the p-value is minimax under squared error loss.*

Lemma 2: *For the problem of testing H_0 versus H_1 where the null hypothesis is simple and the alternative hypothesis is two-sided, the p-value is admissible under squared error loss.*

The proofs of Lemma 1 and 2 start by recognizing the fact that under the squared error loss function

$$L(\mu, d) = \{I(H_0 \text{ is true }) - d(T_1)\}^2, \tag{8}$$

where d is an estimator of the indicator function of H_0 being true. For this situation, the Bayes rule is given by $P(H_0|T_1)$, which has been referred to in recent literature as the local false discovery rate (Efron and Tibshirani, 2002). The p-value can be written as a sequence of these Bayes rules. for the one-sided testing problem. It turns out, however, from Lemma 2, that p-values are inadmissible for the two-sided testing problem. Even though Lemma 2 provides a negative result as to the optimality of the p-value for testing a simple null hypothesis versus a two-sided alternative, Hwang et al. (1990) show that it is impossible to dominate the p-value by any type of Bayes rule.

Note that because of the inability for the p-value to be minimax in the two-sided testing problem, a result analogous to Theorem 1 cannot be achieved for the two-sided p-value. However, by arguments analogous to Hwang et al. (1990), a two-sided p-value (3) should still provide reasonable risk performance. In particular, we expect the Stein phenomenon of risk reduction by borrowing strength from other genes to hold here as well. This will be assessed numerically in Section 5.1. Given that most genomic data analyses focus on testing a simple null versus a two-sided alternative, that will be the one we study here.

4.2. Practical implementation

The major issues in implementing SPADE are twofold. First, π_0 needs to be estimated. This is the proportion of hypotheses estimated to be truly null. Second, the cumulative distribution function for the p-values under the alternative hypothesis also needs to be calculated. This will then provide estimates of d and σ_{p1}^2 in (5). Observe that (2) implies the following result for the cumulative distribution of the p-values:

$$F_P(p) = \pi_0 p + (1 - \pi_0) F_V(p), \tag{9}$$

where F_P is the population cumulative distribution function for the p-values. Simple algebraic manipulation of (9) yields

$$F_V(p) = \frac{F_P(p) - \pi_0 p}{1 - \pi_0}. \tag{10}$$

We can estimate F_P in (10) using the empirical distribution function of the observed p-values. Provided we have an estimator of π_0, we can then estimate F_V and subsequently all the population quantities in (5) and (6). Thus, the outstanding issue becomes one of estimating π_0. We consider three approaches to do this.

The first is the algorithm by Storey and Tibshirani (2003), which has proven quite popular in the analysis of microarray data. It is summarized as follows:

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

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

$l = 1, \ldots, L$.

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 $\lambda = 1$.

We will refer to this procedure as the q-value method.

The second is the SPLOSH algorithm of Pounds and Cheng (2004). Their algorithm proceeds by ordering the p-values and computing a local regression (Cleveland, 1993) where the response variable is a transformed slope of the empirical distribution function of the p-values and the independent variable is a midpoint of the distribution function. Based on the nonparametric regression fit, we get an estimator of π_0. Pounds and Cheng (2004) argue that their method is better than the q-value method of Storey and Tibshirani (2003) because while the estimator of π_0 from the latter method uses information to the left of λ, SPLOSH uses information from both directions. This implies that the SPLOSH estimator should be more stable than the q-value estimator.

The last method for π_0 is based on the algorithm of Dalmasso et al. (2005). The motivation of this method is based on the asymptotic normality of estimators of π_0 using the central limit theorem. Dalmasso et al. (2005) consider an approach in which the p-values are transformed, and an estimator of π_0 is calculated as a backtransformation from the empirical distribution of the transformed p-values. There are many potential transformations that satisfy the necessary technical conditions in Dalmasso et al. (2005); they consider the following estimator of π_0:

$$\tilde{\pi}_0 = \frac{n^{-1} \sum_{i=1}^{n} [-\log(1-p_i)]^m}{m!}, \tag{11}$$

where m is an integer that needs to be estimated. The role of m is similar to that of a bandwidth in nonparametric regression. Larger values of m correspond to decreasing bias in

the estimate of π_0, while smaller values of m lead to decreased variance in the estimate of π_0. Thus, we see a bias-variance tradeoff based on the choice of m. Dalmasso et al. (2005) suggest the following rule for the choice of m:

$$m = \max\left[1, \max\left\{m : \frac{\binom{2m}{m} - 1}{n} \leq l\right\}\right],$$

where l is a postulated value for the variance of π_0. The performance of SPADE with π_0 estimated from the three algorithms is assessed in a simulation study in Section 5.1.

5. Numerical examples

5.1. Simulation studies

To evaluate the proposed procedures, we performed several simulation studies. In these numerical experiments, p-values generated from model (2) with F_0 being the cdf for a uniform [0,1] distribution and F_1 being the cdf for a Beta distribution under three scenarios, which we refer to as Small, Medium, Large. The adjectives refer to the discrepancy of the p-value from the null hypothesis:

1. **Small**: Beta distribution with parameters $\alpha = 3$ and $\beta = 4$. This choice of parameters gives a mean of 3/7 and a variance of 3/98 for the distribution of p-values under the alternative hypothesis.

2. **Medium**: Beta distribution with parameters $\alpha = 3$ and $\beta = 12$. This choice of parameters gives a mean of 3/15 and a variance of 1/100 for the distribution of p-values under the alternative hypothesis.

3. **Large**: Beta distribution with parameters $\alpha = 3$ and $\beta = 50$. This choice of parameters gives a mean of 3/53 and a variance of approximately 0.001 for the distribution of p-values under the alternative hypothesis.

For each simulation setting, we generated 1000 datasets. We considered sample sizes $n = 10000$ and π_0 values of 0.2, 0.5 and 0.8. Our results did not change substantially with sample sizes $n = 2000$ and $n = 5000$; we do not report those results here. The q-value estimation procedure proposed by Storey (2002) was used; the shrunken p-value procedures based on the three algorithms described in Section 4.2 were also studied. The algorithm of Pounds and Cheng (2004) uses a local regression; the span used is the default value of 0.75. For the Dalmasso et al. (2005) method, we chose the default value of $m = 1$. The results using mean-squared error are presented in Table 1, while those using the L_1 analog are given in Table 2.

The simulation sheds light as to the decision-theoretic performance of the q-value method as well as the proposed methods. In Table 2, we see that the L_1 error for the q-value method is lower than for the other three methods; this is based on the fact that the q-value method is based on the positive false discovery rate, which will be the Bayes rule under L_1 error (Storey, 2003). In Table 1, this is not always the case. In particular, when π_0 is small, then the proposed methods are quite competitive with the q-value. In fact, for smaller values of π_0, the q-value estimator of π_0 becomes quite unstable. The instability in the q-value-based estimator of π_0 has also been noticed by Pounds and Morris (2003). If the groups being compared in the differential expression analysis represent grossly different phenotypes, then we would expect π_0 to be small. For more subtle phenotypes, the value of π_0 is larger; this is precisely where the q-value method will perform at its best. In addition, we find that the difference between the the SPADE procedures with the q-value method diminishes for larger numbers of tests. This also suggests that shrinkage will be powerful in high-dimensional situations.

One point of note is that (3) is constructed using squared error loss, which might not be the most appropriate scale for p-values. We also studied a modification of the procedure in Section 3 and 4.2 in which transformed p-values using based on twice the negative log p-values were calculated, double shrinkage estimators were calculated on the transformed

scale and back-transformed. In simulation experiments not reported here, this approach tended to have much worse performance than the procedure described here.

5.2. Microarray example

We now return to the microarray data described in the Introduction. Measurements were made on $n = 9984$ genes for 79 individuals. There are 59 localized prostate cancers and 20 metastatic prostate cancer samples. Before analyzing the data, we took the following preprocessing steps:

1. Genes that were reported as missing in more than 10% of samples were filtered out.

2. Genes that had a sample variation greater than 0.15 across all samples

This left a total of $n = 5241$ genes available for analysis.

We first calculated t-tests comparing gene expression in localized versus metastatic prostate cancer samples; we assumed unequal variances between the two groups. For the purposes of illustration, we used a normal approximation to calculate the p-values. In Table 4, the estimated proportion of true null hypotheses from the various methods are given. We also get the mean and variance of the distribution of p-values under the alternative in this approach; they are also listed in Table 4. Based on the table, we find that qualitative, all three methods find similar means and variances.

We compared the adjustment in p-values using SPADE to the q-value estimates provided by the other procedures (Q-value, SPLOSH, LBE). These are given in Figures 1 - 3. Based on the plots, we find that there is high shrinkage of p-values using SPADE relative to all three methods. With the method of Storey and SPLOSH, the relationship between the shrunken p-values and q-values is monotone but nonlinear. With the LBE-estimated q-values, the shrunken p-values tend to show a more linear relationship.

6. Discussion

In this article, we have argued for a reinterpretation of the mixture model for multiple testing that allows for the consideration of shrinkage procedures. The work of Efron et al. (2001) and Storey (2002) for estimation of false discovery rates represents one method of pooling information across test statistics. We have constructed an alternative procedure based on a James-Stein construction for the p-values. The resulting adjusted p-values from the SPADE procedure represent another method for addressing the multiple testing issue.

While we have proposed new methods for multiple testing, the simulation studies also showed the situations in which the q-value (Storey and Tibshirani, 2003) performs relatively well. Namely, if the proportion of true null hypotheses is large, then the q-value will perform well. If the proportion is small, then the estimate of π_0 will be unstable, which will lead to poor performance of the q-value.

The multiple testing procedure proposed in the paper is based on shrinkage estimation. This is also a common element in Bayesian testing procedures. It has been noted that Bayesian adjustment to the multiple testing problem leads to well-calibrated and more conservative inference procedures than non-Bayesian methods (Gelman and Tuerlinckx, 2000). Based on the results in the real data example, that appears to be the case here as well.

The decision-theoretic framework in which the false-discovery rate procedures have been studied complement the work of Storey (2003), Müller et al. (2004) and Bickel (2004). A natural extension of the SPADE methodology is to construct multiple testing procedures for controlling either FWER or FDR in the spirit of the procedures described at the end of Section 2. The properties of such procedures have not been well-characterized and merit further study.

Because of the availability of software for false discovery rate estimation procedures (Storey, 2002; Pounds and Cheng, 2004; Dalmasso et al., 2005), implementation of the SPADE methodology is very straightforward. R functions implementing the proposed procedures can be obtained from the author.

Acknowledgments

The author would like to thank Tom Nichols and Trivellore Raghunathan for helpful discussions. This research is supported by grant GM72007 from the Joint DMS/DBS/NIGMS Biological Mathematics Program.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* **57**, 289–300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. *J. Stat. Plan. Inf.* **82**, 163 - 170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188.

Bickel, D. R. (2004). Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology* **3**, 8.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Stat.* **42**, 855 – 903.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596 – 610.

Cox, D. R. and Wong, M. Y. (2004). A simple procedure for the selection of significant effects. *J. R. Statist. Soc. B* **66**, 395 – 402.

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., van der Laan, M. J. and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, **3**, 13.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70 – 86.

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

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Boston: Academic Press.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *Test* **12**, 1 – 77.

Gelman, A. and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373 – 390.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B* **64**, 499 – 517.

Genovese, C. and Wasserman, L. (2004). A stochastic approach to false discovery control. *Ann. Statist.* , 1035 – 1061.

George, E. I. (1986). Minimax multiple shrinkage estimation. *Ann. Statist.* **14**, 188 – 205.

Hwang, J. T., Casella, G., Robert, C., Wells, M. T. and Farrell, R. H. (1990). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490 – 509.

Lehmann, E. L and Casella, G. (2002). *Theory of Point Estimation, 2nd Edition.* New York: Springer.

Müller, P., Parmigiani, G., Robert, C. P. and Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. *J. Am. Statist. Assoc.* **468**, 990 – 1001.

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

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of the p-values. *Bioinformatics* **19**, 1236 – 1242.

Sarkar, S. K. (2002). Some results on false discovery rates in multiple testing procedures. *Ann. Statist.* **30**, 239 - 257.

Shaffer, J. (1995). Multiple hypothesis testing. *Ann. Rev. Psych.* **46**, 561–584.

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

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* 31, 2013 – 2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B* **66**, 187 – 205.

Storey, J. D. and Tibshirani R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440 – 9445.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, **3**, 14.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, **3**, 15.

Table 1: Outcomes of n tests of hypotheses

	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_1
	W	Q	m

Table 2: Estimated mean-squared errors from simulation studies

Effect	π_0	Q-value	SPADE1	SPADE2	SPADE3
Small	0.2	0.179	0.186	0.180	0.16
	0.5	0.264	0.333	0.358	0.302
	0.8	0.165	0.333	0.380	0.31
Medium	0.2	0.179	0.183	0.171	0.168
	0.5	0.272	0.328	0.326	0.309
	0.8	0.168	0.330	0.374	0.319
Large	0.2	0.161	0.166	0.173	0.164
	0.5	0.251	0.297	0.275	0.296
	0.8	0.161	0.312	0.310	0.312

Note: Q-value refers to method of Storey and Tibshirani (2003). SPADE1 is the SPADE methodology, where π_0 is estimated using algorithm of Storey and Tibshirani (2003); SPADE2 is based on Pounds and Cheng (2004) method for estimation of π_0; SPADE3 is based on Dalmasso et al. (2005) method for estimation of π_0.

Table 3: Estimated mean absolute deviation from simulation studies

Effect	π_0	Q-value	SPADE1	SPADE2	SPADE3
Small	0.2	0.235	0.223	0.235	0.282
	0.5	0.495	0.500	0.500	0.500
	0.8	0.365	0.570	0.601	0.555
Medium	0.2	0.219	0.225	0.248	0.255
	0.5	0.462	0.494	0.494	0.493
	0.8	0.394	0.566	0.596	0.558
Large	0.2	0.198	0.222	0.217	0.223
	0.5	0.391	0.455	0.449	0.455
	0.8	0.370	0.545	0.544	0.545

Note: See note to Table 2.

Table 4: Estimated proportion of true null hypotheses, mean and variance of distribution of p-values under H_1 for prostate cancer data

Method	π_0	μ_1	σ_1
q-value	0.308	0.040	0.15
SPLOSH	0.460	0.018	0.21
LBE	0.331	0.024	0.15

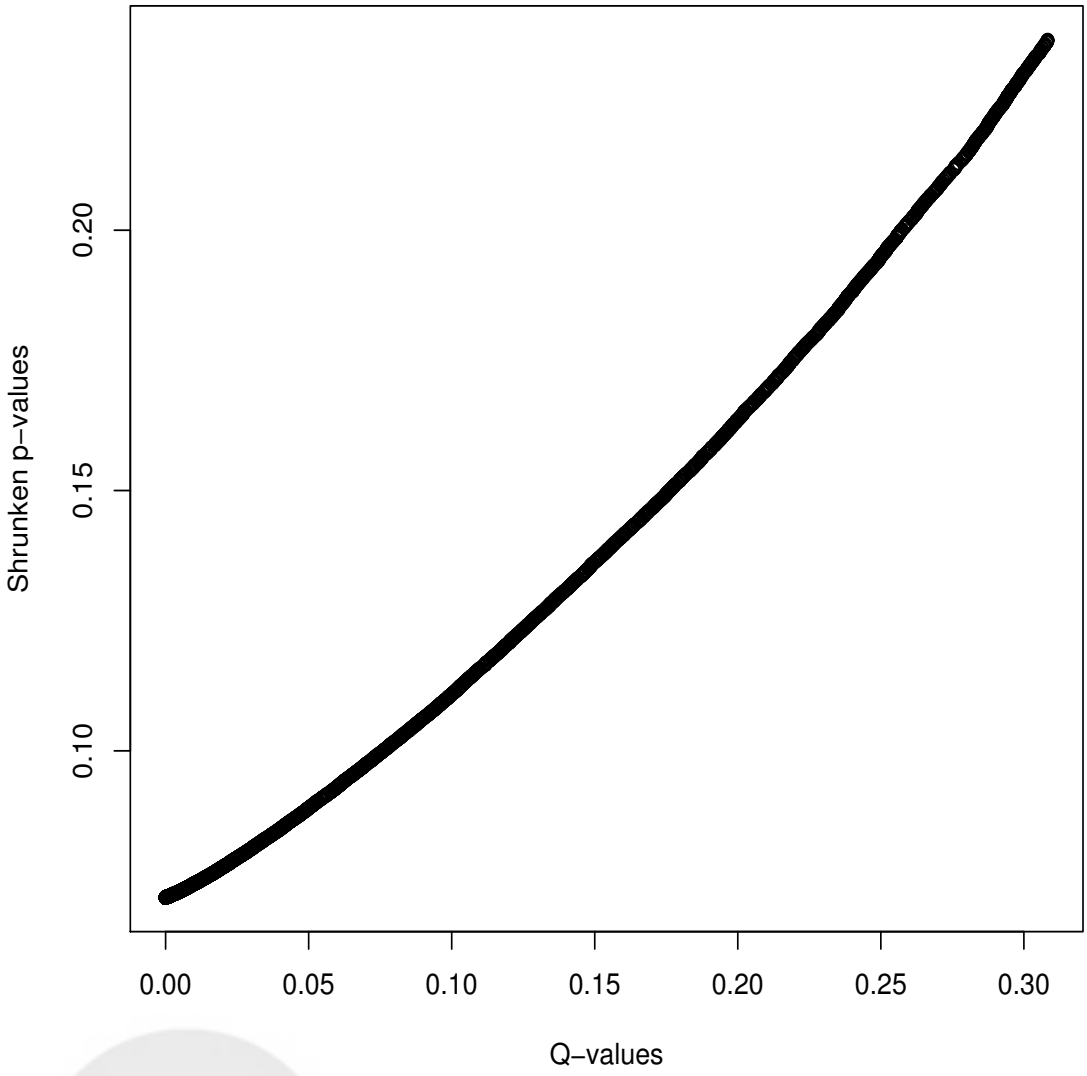

Figure 1: Plot of q-values using Storey (2002) method (horizontal axis) versus shrunken p-values from SPADE.

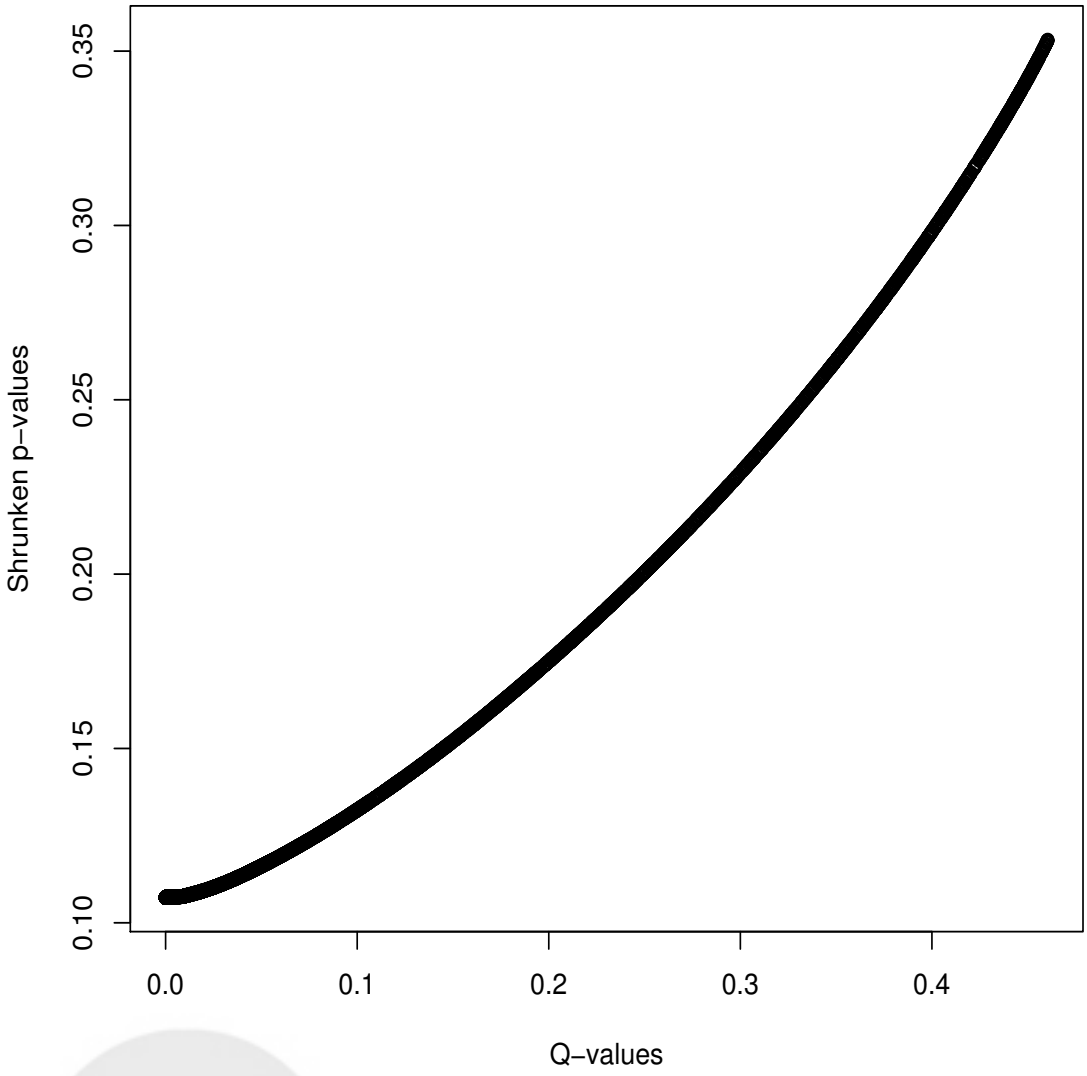

Figure 2: Plot of q-values using Pounds and Cheng (2004) method (horizontal axis) versus shrunken p-values from SPADE.

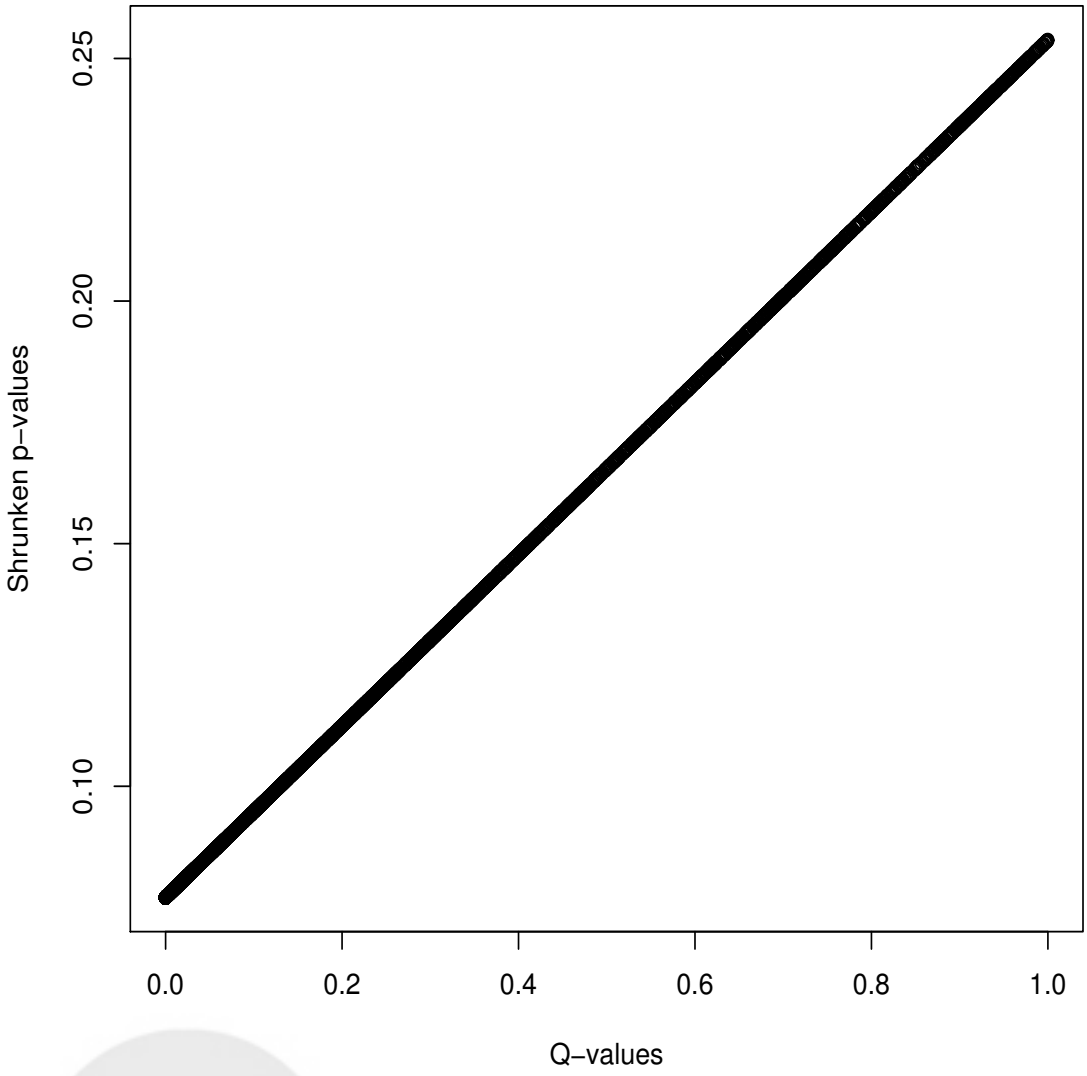

Figure 3: Plot of q-values using Dalmasso et al. (2005) method (horizontal axis) versus shrunken p-values from SPADE.