Biostatistics (2007), **8**, 2, *pp.* 414–432
doi:10.1093/biostatistics/kxl019
Advance Access publication on August 23, 2006

The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments

JOHN D. STOREY*, JAMES Y. DAI, JEFFREY T. LEEK

Department of Biostatistics, University of Washington, Seattle, Washington, 98195, USA
jstorey@u.washington.edu

SUMMARY

As much of the focus of genetics and molecular biology has shifted toward the systems level, it has become increasingly important to accurately extract biologically relevant signal from thousands of related measurements. The common property among these high-dimensional biological studies is that the measured features have a rich and largely unknown underlying structure. One example of much recent interest is identifying differentially expressed genes in comparative microarray experiments. We propose a new approach aimed at optimally performing many hypothesis tests in a high-dimensional study. This approach estimates the optimal discovery procedure (ODP), which has recently been introduced and theoretically shown to optimally perform multiple significance tests. Whereas existing procedures essentially use data from only one feature at a time, the ODP approach uses the relevant information from the entire data set when testing each feature. In particular, we propose a generally applicable estimate of the ODP for identifying differentially expressed genes in microarray experiments. This microarray method consistently shows favorable performance over five highly used existing methods. For example, in testing for differential expression between two breast cancer tumor types, the ODP provides increases from 72% to 185% in the number of genes called significant at a false discovery rate of 3%. Our proposed microarray method is freely available to academic users in the open-source, point-and-click EDGE software package.

Keywords: Differential expression; Multiple hypothesis testing; q-value; Systems biology.

1. INTRODUCTION

The problem of identifying genes that are differentially expressed across varying biological conditions based on microarray data has been a problem of much recent interest (Cui and Churchill, 2003). It is now possible to simultaneously measure thousands of related variables or "features" in a variety of biological studies. Many of these high-dimensional biological studies are aimed at identifying features showing a biological signal of interest, usually through the application of large-scale significance testing. For example, significance analyses are often performed in DNA microarray, comparative genomic hybridization, genome-wide comparative genomics, protein array, mass spectrometry, and genome-wide association studies (Cui and Churchill, 2003; Sebastiani *and others*, 2003; Wang *and others*, 2005).

*To whom correspondence should be addressed.

In many of these applications, the true biological signals of interest across the features are expected to be related. This motivates investigating approaches to large-scale testing that take advantage of widespread structure in high-dimensional data.

We propose a new approach for performing simultaneous significance tests on many features in a high-dimensional study. This approach is based on the "optimal discovery procedure" (ODP), recently developed from a theoretical perspective (Storey, 2005; Storey *and others*, 2005a). The ODP was shown to be optimal in that it maximizes the expected number of true positives (ETP) for each fixed level of expected false positives (EFP); this is also directly related to optimality in terms of the popular false discovery rate (FDR). Here, we introduce approaches to estimating the ODP in practice, and we propose a fully developed method for identifying differentially expressed genes in comparative microarray experiments.

In a microarray study, there is very often pervasive asymmetry in differential expression that is not due to chance. Indeed, it would seem unlikely that overall differential expression would be symmetric, unless the experiment was designed to achieve this behavior. Asymmetric differential expression is an example of the existence of an underlying structure present among thousands of features in a high-dimensional biological study. Due to the pathway structure of gene expression regulation, the expression measurements of genes are related at an even finer scale, which yields further structure in observed differential expression.

A procedure for identifying differentially expressed genes should take advantage of this structure, the same holding true for other high-dimensional biological studies, where much structure in signal is present. The ODP approach does exactly this, utilizing the relevant information from the entire data set in testing each gene for differential expression. The commonly used statistics in high-dimensional studies, such as the *t*-statistic, *F*-statistic, or the chi-square statistic, were originally designed for performing a single significance test. Whereas these statistics are formed using information from only one feature at a time, the ODP takes advantage of the structure in high-dimensional data.

There are two steps implicitly required for performing large-scale significance testing in high-dimensional biological studies: (1) order the features from those showing the most signal of interest to those showing the least and (2) assign a significance level to each feature, allowing one to draw a significance cutoff somewhere along this ordering. As an example, the significance analysis of a microarray study involves ranking the genes from most differentially expressed to least (the first step), and then drawing a significance cutoff based on, say, an estimate of the FDR (the second step). This paper is focused on the first step, namely estimating an optimal ordering of the features. The second step, which is not developed in this paper, has been addressed with new significance measures for high-dimensional studies, such as the FDR (Storey and Tibshirani, 2003).

Estimating the ODP in practice requires the development of a number of ideas beyond those considered in the more theoretical setting of Storey (2005), which we illustrate through the microarray application. For example, whereas a *t*-statistic automatically cancels out ancillary information in testing for differential expression, certain approaches to estimating the ODP do not. Therefore, steps must be taken so that such ancillary information has no effect on the significance results. Here, we introduce a general set of methodology that overcomes a number of these challenges.

We demonstrate the proposed ODP approach for identifying differentially expressed genes on a well-known breast cancer expression study Hedenfalk *and others* (2001), as well as on simulated data. We compare the results to those from five leading differential expression methods (Tusher *and others*, 2001; Kerr *and others*, 2000; Dudoit *and others*, 2002; Cui *and others*, 2005; Efron *and others*, 2001; Lonnstedt and Speed, 2002). Our method consistently shows substantial improvements in performance over these existing methods. For example, in testing for differential expression between *BRCA1* and *BRCA2* mutation-positive tumors, the ODP approach provides increases from 72% to 185% in the number of genes called significant at a 3% FDR. A comparison between the methods over a range of FDRs is shown in Figure 2 and Table 1.

Table 1. *Improvements of the ODP approach over existing thresholding methods. Shown are the minimum, median, and maximum percentage increases in the number of genes called significant by the proposed ODP approach relative to the existing approaches among FDR levels* 2%, 3%, . . . , 10%. *The exact same FDR methodology (Storey, 2002; Storey and Tibshirani, 2003) was applied to each gene-ranking method in order to make the comparisons fair. The model-based Bayesian method (Lonnstedt and Speed, 2002) is not defined for a three-sample analysis, so that case is omitted*

Thresholding method	% Increase by ODP—two-sample			% Increase by ODP—three-sample		
	Minimum	Median	Maximum	Minimum	Median	Maximum
SAM (Tusher *and others*, 2001)	29	43	72	76	92	211
t/F-test (Dudoit *and others* 2002, Kerr *and others*, 2000)	52	86	185	63	82	407
Shrunken t/F-test (Cui *and others*, 2005)	34	52	77	61	69	154
Bayesian local FDR (Efron *and others*, 2001)	58	87	117	76	92	211
Posterior probability (Lonnstedt & Speed 2002)	44	60	113	—	—	—

2. THE ODP

2.1 *Optimality goals*

The typical goal when identifying differentially expressed genes is to find as many true positives as possible, without incurring too many false positives (Storey and Tibshirani, 2003). Sometimes genes found to be significantly differentially expressed are subsequently studied on a case-by-case basis in order to determine their role in the differing biological conditions. It is also now possible to discover functional relationships among significant genes based on a number of ontological databases, making this an attractive and more frequently used follow-up investigation technique (Zhong *and others*, 2004).

Because of these goals in microarray experiments and a variety of other high-dimensional biological applications, the FDR has emerged as a popular criterion for assessing significance in high-dimensional biological studies (Storey and Tibshirani, 2003). The FDR is defined to be the proportion of false positives among all features called significant (Soric, 1989; Benjamini and Hochberg, 1995). For example, if 100 genes are called significant at the 5% FDR level, then one expects 5 out of these 100 to be false positives. When investigating the functional relationships of a set of significant genes, the FDR has the nice interpretation that it represents the level of "noise" present in the genes used to draw conclusions about the functional relationships.

Instead of working directly with FDRs, the ODP is based on two more fundamental quantities: ETP and EFP. Specifically, the ODP is defined as the testing procedure that maximizes the ETP for each fixed EFP level. Since FDR optimality can be written in terms of maximizing the ETP for each fixed EFP level (Storey, 2005), the ODP also provides optimality properties for FDR. A consequence of this optimality is that the rate of "missed discoveries" is minimized for each FDR level. In fact, the optimality properties of the ODP translate to a variety of settings, including misclassification rates (Storey, 2005). The ODP optimality can also be formulated as a multiple test extension of this Neyman–Pearson optimality (Storey, 2005).

As rigorously described in Storey (2005), the ODP optimization is carried out among all single-thresholding procedures (STPs). The ODP statistic defined next is called a "significance thresholding function" in Storey (2005) because it acts as a function applied to each test, where a numerical threshold is then applied to these in order to call the tests significant. An STP is simply a procedure where a single

significance thresholding procedure is used for all tests, or more simply, one where a common formula is employed for each test's statistic. Any method invariant to the labeling of the tests is an STP; all existing methods considered in this paper are STPs.

2.2 *ODP statistic*

The ODP is very much related to one of the fundamental ideas behind individual significance tests: the Neyman–Pearson lemma. Given a single set of observed data, the optimal single-testing procedure is based on the statistic

$$\mathcal{S}_{\mathrm{NP}}(\text{data}) = \frac{\text{probability of the data under the alternative distribution}}{\text{probability of the data under the null distribution}}.$$

The null hypothesis is then rejected if the statistic $\mathcal{S}_{\mathrm{NP}}(\text{data})$ exceeds some cutoff chosen to satisfy an acceptable Type I error rate. (Here, the larger the statistic is, the more significant the test is.) This Neyman–Pearson procedure is optimal because it is "most powerful," meaning that for each fixed Type I error rate, there does not exist another rule that exceeds this one in power. The optimality follows intuitively from the fact that the strength of the alternative versus the null is assessed by comparing their exact likelihoods.

The ODP statistic may be written similarly to the NP statistic. However, instead of considering the data evaluated at their own alternative and null probability density functions, the ODP considers the data for a single feature evaluated at all true probability density functions. Let "data_i" be the data for the ith feature being tested. The ODP statistic for feature i is calculated as

$$\mathcal{S}_{\mathrm{ODP}}(\text{data}_i) = \frac{\text{sum of probability of data}_i \text{ under each true alternative distribution}}{\text{sum of probability of data}_i \text{ under each true null distribution}}. \tag{2.1}$$

For a fixed cutoff chosen to attain an acceptable EFP level (or FDR level), each null hypothesis is rejected if its ODP statistic $\mathcal{S}_{\mathrm{ODP}}(\text{data}_i)$ exceeds the cutoff. Note that data_i has been evaluated at all true probability densities, thereby using the relevant information from the entire set of features. For each feature's data, evidence is added across the true alternatives and compared to that across the true nulls in forming the ratio.

Figure 1 gives a graphical representation of the ODP statistic and its relative behavior to the NP statistic. It can be seen there that the difference between the two is that the ODP borrows strength across all the tests, as opposed to using information from only one test at a time. This point is explored in depth in Storey (2005). In the supplementary material available at *Biostatistics* online, we provide a toy example showing how microarray data contain information shared across genes that can be utilized by the ODP. The NP procedure and ODP are theoretical procedures that must be estimated in practice. As it turns out, the estimated ODP may show favorable operating characteristics over estimated NP procedures when testing many hypotheses, as we demonstrate in this article.

2.3 *Mathematical formulation*

To make the definition of the ODP statistic more precise, suppose that m significance tests are performed on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where each significance test consists of n observations so that each $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. For the microarray application that we consider, x_{ij} is the relative expression level of gene i on array j. In this case, there are m genes tested for differential expression, based on n microarrays.

Assume that significance test i has null probability density function f_i and alternative density g_i; without loss of generality suppose that the null hypothesis is true for tests $i = 1, 2, \ldots, m_0$ and the

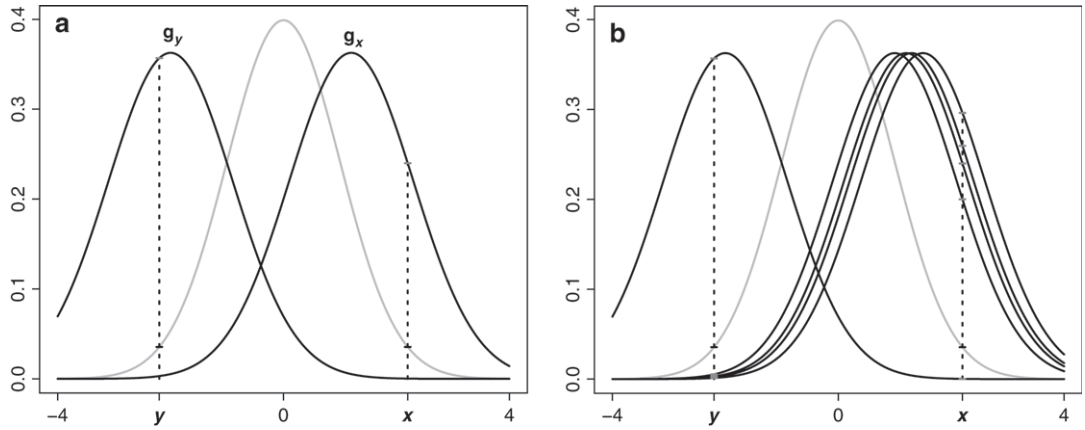

Fig. 1. Plots comparing the NP testing approach to the ODP testing approach through a simple example. (a) NP approach. The null (gray) and alternative (black) probability density functions of a single test. For observed data x and y, the statistics are calculated by taking the ratio of the alternative to the null densities at each respective point. In this NP approach, the test with data y is more significant than the test with data x. (b) ODP approach. The common null density (gray) for true null tests and the alternative densities (black) for several true alternative tests. For observed data x and y, the statistics are calculated by taking the ratio of the sum of alternative densities to the null density evaluated at each respective point. In this ODP approach, the test with data x is now more significant than the test with data y because multiple alternative densities have similar positive means even though each one is smaller than the single alternative density with negative mean. A color version of the figure is given in the supplementary material available at *Biostatistics* online, Figure 8.

alternative is true for $i = m_0 + 1, \ldots, m$. In this notation, the ODP statistic of (2.1) is written as:

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x})}. \tag{2.2}$$

Null hypothesis i is rejected if and only if $\mathcal{S}_{\text{ODP}}(\mathbf{x}_i) \geqslant \lambda$, where λ is chosen to satisfy an acceptable EFP or FDR level. In practice, the exact forms of f_i and g_i are unknown, as well as which of the tests have a true null hypothesis. Therefore, this statistic not only requires one to know the distributions associated with each test but also whether the null or alternative is true for each test.

This seemingly nonsensical requirement turns out to be tractable when estimating the ODP. However, it requires that we use a different but equivalent form of the statistic. The following equivalently defines the ODP, as shown by Storey (2005):

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x}) + g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x})}, \tag{2.3}$$

which equals 1 + equation (2.2). Since equation (2.3) = 1 + equation (2.2), these produce the exact same testing procedure (where a threshold of λ applied to the statistic defined in (2.2) is equivalent to a threshold of $1 + \lambda$ applied to the statistic defined in (2.3)). Because of this equivalence and the tractability of estimating the statistic in (2.3), we employ and estimate this statistic for the remainder of the article.

3. PROPOSED APPROACH FOR ESTIMATING THE ODP

Since the true ODP requires information not known in practice, the procedure must be estimated; here, we propose some general methodology for doing so. The goal when estimating the ODP is to be able to

reproduce the same ranking of features as the true ODP. Note that it is not necessary to reproduce the ODP statistics exactly, but rather their relative ranking. In order to estimate the ODP statistic of (2.3), one must estimate the true probability density function for each test and also address the fact that only the true null tests are represented in the denominator of the statistic. The first challenge is straightforward to address: we use the observed data for each test in order to estimate its true probability function. This is clearly justified by the fact that the data are generated from that true density function. The second challenge can be addressed in several ways, some of which we propose below.

3.1 *A canonical plug-in estimate*

A parametric approach can be taken to estimate the ODP, motivated by the generalized likelihood ratio test for single significance tests. Recall that f_i and g_i will both be defined by a set of parameters (e.g. the mean and variance of a normal distribution). For each test $i = 1, \ldots, m$, let $\widehat{f}_i$ be the version of f_i with all the unknown parameters replaced by their maximum likelihood estimates under the constraints of the null hypothesis and $\widehat{g}_i$ be the analogous estimate given by the unconstrained maximum likelihood estimates, both based only on data $\mathbf{x}_i$. In single hypothesis testing, the Neyman–Pearson procedure for test i is based on $g_i(\mathbf{x}_i)/f_i(\mathbf{x}_i)$, and it can be estimated by the generalized likelihood ratio statistic $\widehat{g}_i(\mathbf{x}_i)/\widehat{f}_i(\mathbf{x}_i)$ (Lehmann, 1986). Our proposed approach builds on this strategy.

For "true" null hypotheses $i = 1, \ldots, m_0$, the maximum likelihood parameters defining $\widehat{f}_i$ and $\widehat{g}_i$ are both consistent estimates of the actual values of f_i as the number of observations n grows to infinity. Likewise, $\widehat{g}_i$ is composed of consistent parameter estimates of g_i for false null hypotheses $i = m_0 + 1, \ldots, m$. Therefore, $\widehat{g}_1 + \cdots + \widehat{g}_m$ can be used to estimate the numerator of (2.3), where it is now unnecessary to be able to distinguish between true and false null hypotheses. This motivates the following "canonical estimate" of the ODP statistic:

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}) = \frac{\widehat{g}_1(\mathbf{x}) + \cdots + \widehat{g}_{m_0}(\mathbf{x}) + \widehat{g}_{m_0+1}(\mathbf{x}) + \cdots + \widehat{g}_m(\mathbf{x})}{\widehat{f}_1(\mathbf{x}) + \cdots + \widehat{f}_{m_0}(\mathbf{x})}. \tag{3.1}$$

We use the term "canonical" because the above is a direct plug-in estimate of the ODP thresholding function, where all unknown parameters are consistently estimated.

Consistency in the number of observations n for each test is not necessarily the best property to be concerned about in this setting, since it will usually be the case that $n \ll m$; nevertheless, many of the commonly used statistics (t, F, chi-square) can be motivated from this perspective, while also displaying good small sample properites. Other well-behaved estimates of the f_i and g_i could certainly be employed if they show favorable operating characteristics.

3.2 *Common null distribution estimate*

In general, it will not be possible to employ the canonical estimate because it requires one to be able to identify the densities of the true null hypotheses. If a common null distribution f exists and is known, then one does not need to know which of the null hypotheses are true. The canonical ODP estimate can then be simplified to

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{f(\mathbf{x})}. \tag{3.2}$$

Note that sometimes it is possible to transform the data so that the null distribution becomes known and common among all tests (e.g. by replacing the data with a pivotal statistic). However, this may remove much of the information in the data, making this approach less desirable. If there is no common and known null distribution, then the following more generally applicable estimate is proposed.

3.3 *Generally applicable estimate*

One general approach is to approximate the canonical plug-in estimate by estimating which null densities should be included in the denominator of the statistic. Let $\widehat{\omega}_i = 1$ if $\widehat{f}_i$ is to be included in the denominator, and $\widehat{\omega}_i = 0$ otherwise. The estimate of the ODP statistic is then

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \widehat{\omega}_i \, \widehat{f}_i(\mathbf{x})}. \tag{3.3}$$

More generally, the $\widehat{\omega}_i$ can be thought of as weights serving as estimates of the true status of each hypothesis. We have defined them as equaling zero or one, but they could take on a continuum of values as well.

We propose and implement a simple approach to forming the $\widehat{\omega}_i$ for the microarray application below, although many different approaches would be possible. This simple approach is based on ranking the tests by using a univariate statistic (e.g. a *t*-statistic). For all statistics exceeding some cutoff (i.e. those appearing to be significant and not likely to be true nulls), we set $\widehat{\omega}_i = 0$; for those not exceeding the cutoff, we set $\widehat{\omega}_i = 1$. The cutoff is formed so that the proportion falling below and receiving $\widehat{\omega}_i = 1$ is equal to an estimate of the proportion of true null hypotheses, based on the method in Storey (2002) and Storey and Tibshirani (2003).

Note that if the tests are consistent, then we expect the true alternative tests to rise above the cutoff with probability one. The proportion of true null tests can be estimated unbiasedly in this case (Storey, 2002), providing a reasonable method for extracting the true null densities to be employed in the denominator of the statistic. Our particular version of this procedure, based on a Kruskal–Wallis test statistic and the estimate of the proportion of true nulls by Storey (2002) and Storey and Tibshirani (2003), performs nearly as well as the canonical estimate according to our simulations.

3.4 *Nuisance parameter invariance*

In addition to estimating the ODP well, it is also necessary to consider the effect of ancillary information on the procedure. Specifically, it is desirable to obtain a "nusiance parameter invariance" property. Suppose that all significance tests have equivalently defined null and alternative hypotheses and their probability density functions all come from the same family. If the null distributions f_i are not equal, then this is due to differing nuisance parameters. However, simply changing the nuisance parameters of the true null hypotheses can produce substantial (and sometimes undesirable) alterations in the ODP (supplementary material available at *Biostatistics* online). A strong way to enforce nuisance parameter invariance is to require all f_i's to be equal. Alternatively, one may require that $\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m_0} f_i/m_0$ so that on average there is no relationship between the status of the hypotheses and the null distributions. See supplementary material available at *Biostatistics* online for a more detailed discussion on this important property.

In practice, it is sometimes possible to formulate the significance tests or transform the data so that $\sum_{i=1}^{m} f_i/m \approx \sum_{i=1}^{m_0} f_i/m_0$. When this nuisance parameter invariance property is met, $\sum_{i=1}^{m} \widehat{f}_i/m$ may serve as an estimate of $\sum_{i=1}^{m_0} f_i/m_0$, yielding the following estimate of the ODP thresholding rule:

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \widehat{f}_i(\mathbf{x})}, \tag{3.4}$$

where the unknown constant m_0/m can be omitted. However, it may also be difficult to estimate the f_i for true alternative tests since their data are in fact generated from the alternative density g_i. In other words, $\widehat{f}_i$ may be a poor estimate of f_i for $i > m_0$, making the denominator of (3.4) poorly behaved.

4. ODP FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES

For the microarray application, we found the implementation based on our general estimate of (3.3) to perform the best. This implementation requires (1) f_i and g_i to be defined, (2) estimates $\widehat{f}_i$ and $\widehat{g}_i$ to be derived, (3) an estimate of which $\widehat{f}_i$ to employ in the denominator to be derived, and (4) justification that the nuisance parameter invariance condition $\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m_0} f_i/m_0$ is approximately met.

Some notation is necessary to describe the implementation. We assume expression is measured on m genes from n arrays, where the n arrays come from one of two distinct groups. (The methodology easily extends to there being one, two, or more groups—details are given below.) Let μ_{i1} be the mean of gene i in group 1, and μ_{i2} be the mean of gene i in group 2, $i = 1, \ldots, m$. When gene i is not differentially expressed, these means are equal and we denote them by their common mean μ_{i0}. We denote x_{ij} to be the expression observation for gene i in array j, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. As before, we represent the data for a single gene by $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. Also, let $\mathbf{x}_{i1}$ be the subset of data from group 1 and $\mathbf{x}_{i2}$ the subset of data from group 2. For example, with seven arrays in group 1 and eight in group 2, we write $\mathbf{x}_{i1} = (x_{i1}, x_{i2}, \ldots, x_{i7})$ and $\mathbf{x}_{i2} = (x_{i8}, x_{i9}, \ldots, x_{i15})$.

4.1 *Probability density functions*

The model we use to estimate the ODP is that x_{ij} comes from a normal distribution with mean μ_{i1} or μ_{i2} (depending on the group that array j belongs to) and variance σ_i^2. Note that this is only an assumption insofar as claims are made about the accuracy of the estimated ODP with respect to the true ODP. We do not make any distributional assumptions when assessing the level of statistical significance for each feature. We assume that the expression measurements x_{ij} are on the log scale or whatever scale makes the use of the normal densities most reasonable.

Under this assumption, the likelihood of a set of data can be written using the normal probability density function ϕ. For example, the likelihood of data $\mathbf{x}$ with mean μ and variance σ^2 is written as

$$\phi(\mathbf{x}; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\sum_{j=1}^{n}(x_j - \mu)^2}{2\sigma^2} \right\}.$$

In the notation used to define the general ODP estimates, we therefore define

$$f_i(\mathbf{x}) = \phi(\mathbf{x}; \mu_{i0}, \sigma_i^2) \quad \text{and} \quad g_i(\mathbf{x}) = \phi(\mathbf{x}_1; \mu_{i1}, \sigma_i^2)\phi(\mathbf{x}_2; \mu_{i2}, \sigma_i^2).$$

For hypothesis i, the probability of data $\mathbf{x}$ is $f_i(\mathbf{x})$ under the null and $g_i(\mathbf{x})$ under the alternative.

4.2 *Estimates of the densities*

Ignoring nuisance parameter invariance issues, it is straightforward to define estimates of these densities. Let $(\widehat{\mu}_{i0}, \widehat{\sigma}_{i0}^2)$ be the maximum likelihood estimates under the constraints of the null hypothesis, and $(\widehat{\mu}_{i1}, \widehat{\mu}_{i2}, \widehat{\sigma}_{iA}^2)$ be the unconstrained maximum likelihood estimates. These are simply the sample means and variances under the assumptions of the null and alternative hypotheses, respectively (supplementary material available at *Biostatistics* online). The above densities can then simply be estimated by $\widehat{f}_i(\cdot) = \phi(\cdot; \widehat{\mu}_{i0}, \widehat{\sigma}_{i0}^2)$ and $\widehat{g}_i(\cdot) = \phi(\cdot; \widehat{\mu}_{i1}, \widehat{\sigma}_{iA}^2)\phi(\cdot; \widehat{\mu}_{i2}, \widehat{\sigma}_{iA}^2)$. Below, we modify these density definitions and estimates to approximately achieve nuisance parameter invariance.

4.3 *Extracting true null densities for the denominator*

We also estimate which null densities should appear in the denominator of the statistic. The ultimate goal is to recover the canonical estimate (3.3), where only $\widehat{f}_i$ corresponding to true nulls are present in the denominator. We take the approach outlined in the Section 3.3, summarized in the following algorithm.

1. Perform a Kruskal–Wallis test for differential expression on each gene and rank the genes from most differentially expressed to least according to this test.
2. Using the p-values from these tests, estimate the number of differentially expressed genes $\widehat{m}_0$ according to the methodology in Storey (2002) and Storey and Tibshirani (2003).
3. Set $\widehat{\omega}_i = 1$ for the genes falling in the bottom $\widehat{m}_0$ of the ranking; set $\widehat{\omega}_i = 0$ otherwise.

A rank-based test is used mainly because it is computationally efficient. Furthermore, if a t-statistic or F-statistic were used, then this runs the risk of preferentially selecting genes with small variances by chance, a phenomenon previously noted about such statistics (Tusher *and others*, 2001). It should be stressed that this is one of the many approaches one could take to estimating which null densities to include in the denominator. We anticipate that better strategies will be found in the future. However, the procedure proposed here does in fact show improvements over setting all $\widehat{\omega}_i = 1$. Furthermore, at this stage it is not necessarily so important to identify individual null genes well, but rather to identify a subset so that $\sum_{i=1}^{m} \widehat{\omega}_i \widehat{f}_i$ approximates $\sum_{i=1}^{m_0} f_i$ well.

4.4 *Nuisance parameter invariance*

According to our notation, the null hypothesis for gene i is that $\mu_{i1} = \mu_{i2}$ and the alternative is that $\mu_{i1} \neq \mu_{i2}$. This can be rewritten as $\mu_{i1} - \mu_{i2} = 0$ versus $\mu_{i1} - \mu_{i2} \neq 0$. Without loss of generality, the common mean when the null hypothesis is true can be defined as $\mu_{i0} = (n_1\mu_{i1} + n_2\mu_{i2})/n$, where n_1 and n_2 are the number of arrays in groups 1 and 2, respectively. The data for gene i can then be equivalently parameterized by $(\mu_{i0}, \mu_{i1} - \mu_{i2}, \sigma_i^2)$ rather than $(\mu_{i1}, \mu_{i2}, \sigma_i^2)$. It is clear that the parameters μ_{i0} and σ_i^2 are not of interest in the hypothesis test; these are the so-called nuisance parameters.

Recall that the goal is to approximately achieve the equality $\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m_0} f_i/m_0$. If (1) the distribution of the σ_i^2 is unrelated to the distribution of the $\mu_{i1} - \mu_{i2}$ and (2) each $\mu_{i0} = 0$, then we can approximately achieve the nuisance parameter invariance condition (supplementary material available at *Biostatistics* online). Standard methods make it straightforward to transform the data so that there is no apparent relationship between the σ_i^2 and the $\mu_{i1} - \mu_{i2}$ (Rocke and Durbin, 2003), so this condition can often be fulfilled in practice. Ideally, we would force $\mu_{i0} = 0$ by subtracting the true μ_{i0} from each x_{ij} for $j = 1, \ldots, n$. However, μ_{i0} are unknown, so these must be estimated. Therefore, we set $\widehat{\mu}_{i0} = \sum_{j=1}^{n} x_{ij}/n$ and define $x_{ij}^* = x_{ij} - \widehat{\mu}_{i0}$, thereby centering each gene around zero.

With the data transformed in this manner, it follows that $\mu_{i0}^* = 0$, $\mu_{i1}^* = \mu_{i1} - \mu_{i0}$ and $\mu_{i2}^* = \mu_{i2} - \mu_{i0}$, with estimates $\widehat{\mu}_{i1}^* = \widehat{\mu}_{i1} - \widehat{\mu}_{i0}$ and $\widehat{\mu}_{i2}^* = \widehat{\mu}_{i2} - \widehat{\mu}_{i0}$. The variances σ_i^2 do not change, so these can be estimated as before by taking the sample variances under the assumptions of the null and alternative hypotheses to get $\widehat{\sigma}_{i0}^2$ and $\widehat{\sigma}_{i\mathrm{A}}^2$, respectively.

4.5 *Estimated ODP thresholding function*

The ODP for identifying differentially expressed genes between two groups can then be estimated by forming the following statistic for each gene $i = 1, 2, \ldots, m$:

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i) = \frac{\sum_{g=1}^{m} \phi(\mathbf{x}_{i1}^*; \widehat{\mu}_{g1}^*, \widehat{\sigma}_{gA}^2)\phi(\mathbf{x}_{i2}^*; \widehat{\mu}_{g2}^*, \widehat{\sigma}_{gA}^2)}{\sum_{g=1}^{m} \widehat{\omega}_i \phi(\mathbf{x}_i^*; 0, \widehat{\sigma}_{g0}^2)}.$$

Note that the centered data for gene i, $\mathbf{x}_i^*$ is evaluated at the estimated likelihood functions for all genes. Therefore, if gene g has a similar signal to gene i, then its likelihood under the alternative will contribute substantially to the estimated ODP statistic of gene i. Also, the variance of a gene is taken into account in its contribution to the statistic, where the smaller the variance, the more its likelihood is allowed to contribute to gene i's statistic. The formula of the statistic also makes it clear why it is useful to use the gene-centered data $\mathbf{x}_i^*$. Strength is borrowed across genes that have a similar structure in the signal, even if they have different baseline levels of expression (which is not of interest for detecting differential gene expression).

This method is easily extended to a general K-sample analysis, where K different biological groups are compared for differential expression. For example, in a three-sample analysis the goal is to identify genes whose mean expression is different in at least one of the three groups. The estimated ODP statistic for a K-sample significance test of differential expression is a simple extension of the above two-sample statistic:

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i) = \frac{\sum_{g=1}^{m} \phi(\mathbf{x}_{i1}^*; \widehat{\mu}_{g1}^*, \widehat{\sigma}_{g\mathrm{A}}^2) \cdots \phi(\mathbf{x}_{iK}^*; \widehat{\mu}_{gK}^*, \widehat{\sigma}_{g\mathrm{A}}^2)}{\sum_{g=1}^{m} \widehat{\omega}_i \phi(\mathbf{x}_i^*; 0, \widehat{\sigma}_{g0}^2)}. \tag{4.1}$$

Analogous to the two-sample method, each gene is mean centered around zero to obtain the transformed data $\mathbf{x}_i^*$. In the one-sample case, the data do not have to be mean centered because there is no nuisance location parameter present.

4.6 *Existing methods*

Most of the existing methods for identifying differentially expressed genes implicitly make the normal distribution assumption that we have made. The statistic for gene i is then formed by $\widehat{g}_i(\mathbf{x}_i)/\widehat{f}_i(\mathbf{x}_i)$. When the estimated parameters defining $\widehat{f}_i$ and $\widehat{g}_i$ are the maximum likelihood estimates, then $\widehat{g}_i(\mathbf{x}_i)/\widehat{f}_i(\mathbf{x}_i)$ is equivalent to employing the usual t-statistic (Lehmann, 1986). When the maximum likelihood estimates are shrunken toward a common value (across genes), then the so-called significance analysis of microarrays (SAM) statistic and other similar versions emerge (Tusher *and others*, 2001; Cui *and others*, 2005; Efron *and others*, 2001). Therefore, these more intricate statistics use information across genes only in that different estimates are employed in $\widehat{g}_i(\mathbf{x}_i)/\widehat{f}_i(\mathbf{x}_i)$. Not surprisingly, these modified statistics sometimes perform worse than the traditional t-statistic and F-statistic (Section 5).

4.7 *Overall algorithm for identifying differentially expressed genes*

The following is a description of the estimated ODP for identifying differentially expressed genes. The basic approach is to form estimated versions of the ODP statistics and then assess significance using the q-value (Storey, 2002; Storey and Tibshirani, 2003). Full details of this algorithm, including exact formulas can be found in the supplementary material available at *Biostatistics* online. Note that one can also determine a useful significance threshold through estimates of the EFP and ETP, which we also outline in the supplementary material available at *Biostatistics* online.

Proposed algorithm for identifying differentially expressed genes

1. Using the formula given above in (4.1), evaluate the estimated ODP statistic for each gene.
2. For B iterations, simulate data from the null distribution for each gene by the bootstrap and recompute each statistic to get a set of null statistics. (Note: The bootstrap sampling is carried out so that for each iteration, the same resampled arrays are applied to all genes. This keeps the dependence structure of the genes intact.)

3. Using these observed and null statistics, estimate the q-value for each gene as previously described (Storey, 2002; Storey and Tibshirani, 2003).

The algorithm generates an estimated q-value for each gene and a ranking of the genes from most significant to least significant. The q-value is like the well-known p-value, but it is designed for the FDR; the q-value of a gene gives the FDR that is incurred when calling that gene and all others with larger statistics significant (Storey, 2003; Storey and Tibshirani, 2003). One may call genes significant for differential expression by forming a q-value cutoff at an appropriate level (say, 1%, 5%, or 10%), or one may simply report the q-value for every gene and let each individual researcher choose a level of desirable significance. We now apply this method to a well-known breast cancer study, and we compare the ODP approach to several highly used existing approaches.

5. RESULTS

5.1 *Analysis of breast cancer tumor tissue*

We assessed the performance of the ODP on a well-known study comparing the expression of breast cancer tumor tissues among individuals who are *BRCA1*-mutation-positive, *BRCA2*-mutation-positive, and "Sporadic" (Hedenfalk *and others*, 2001). The expression measurements used in the study consist of 3226 genes on 22 arrays; seven arrays were obtained from the *BRCA1* group, eight from the *BRCA2* group, and six from the Sporadic group. One sample was not clearly classifiable, so we eliminated it from the analysis here. Also, as previously described Storey and Tibshirani (2003), several genes have aberrantly large expression values within a single group, so we eliminated those genes from the analysis. Genes were filtered that had any absolute expression measurement greater than 20, which is well beyond several times the interquartile range from the median. These steps left measurements on 3169 genes from 21 arrays. The raw data were obtained from http://research.nhgri.nih.gov/microarray/NEJM_Supplement/ and all data were analyzed on the $\log_2$ scale. We applied our proposed procedure to identify differentially expressed genes between the *BRCA1* and *BRCA2* groups, and also between all three groups.

We compared our approach to five leading techniques, including (1) the highly used SAM software based on Tusher *and others* (2001) and Storey (2002), (2) the traditional t-tests and F-tests as previously suggested for microarray analysis (Kerr *and others*, 2000; Dudoit *and others*, 2002), (3) a recently proposed variation on these that uses "shrunken" versions of the statistics (Cui *and others*, 2005), (4) a nonparametric Bayesian method whose estimated posterior probabilities are also sometimes interpreted as estimated Bayesian local FDR estimates (Efron *and others*, 2001), and (5) a model-based empirical Bayes method giving posterior probabilities of differential expression (Lonnstedt and Speed, 2002).

The methods were compared to determine how accurately and efficiently each one extracts the relevant biological signal. Each method produces some sort of statistic for each gene, as well as a rule for thresholding these statistics. We used this ranking information to estimate q-values for each gene according to previously described methodology (Storey, 2002; Storey and Tibshirani, 2003). In order to estimate the q-values, simulated null statistics were calculated for each method. This was accomplished by simulating the same null data in order to calculate null statistics for each method.

It should be noted that several model-based Bayesian methods exist (e.g. Newton *and others* 2001, 2004; Townsend & Hartl 2002) for identifying differentially expressed genes. In particular, Newton *and others* (2004) offers a semiparametric empirical Bayes approach that provides an estimate of a Bayesian version of the FDR. The method is not included in our comparison because of its different approach to quantifying the FDR. We have only compared methods that have been proposed in or are easily amenable to the framework of calculating significance based on a resampling-based frequentist FDR.

Newton *and others* (2004) and three of the methods we include in our comparison (Tusher *and others*, 2001; Efron *and others*, 2001; Lonnstedt and Speed, 2002) are able to capture asymmetry in differential

expression signal for when comparing two groups. Tusher *and others* (2001) and Efron *and others* (2001) do not do so for three or more groups, so they are essentially equivalent to a standard F-test for three or more groups or time course studies. As we have described, the ODP captures any structure in the signal; this could be asymmetry in differential expression for two or more groups, variance structure, or structured temporal trajectories in a time course study.

5.2 *Numerical results on the breast cancer data*

The methods were compared by considering the number of genes called significant across a range of FDR cutoffs, which gives an estimate of the relative ETP levels at each given FDR (supplementary material available at *Biostatistics* online). For the methods employed here, this is equivalent to comparing the ETP for each fixed EFP level or p-value cutoff on a slightly different scale. Intuitively, the number of genes called significant quantifies the relative amount of biological information obtained at a given noise level. Figure 2 plots the number of genes called significant among the different methods across a range of estimated q-value cutoffs.

In testing for differential expression between the *BRCA1* and *BRCA2* groups, the ODP approach shows notable improvements in performance over existing methods. For example, at an FDR level of 3%, our proposed approach finds 117 significant genes, whereas existing methods only find 41–68 significant genes. The estimated ODP method therefore offers increases from 72% to 185% in the number of genes called significant. The median increase in the number of genes called significant at q-value cutoffs less than or equal to 10% ranges from 43% to 87% across all methods. In testing for three-sample differential expression among the *BRCA1*, *BRCA2*, and Sporadic groups, the ODP approach offers even

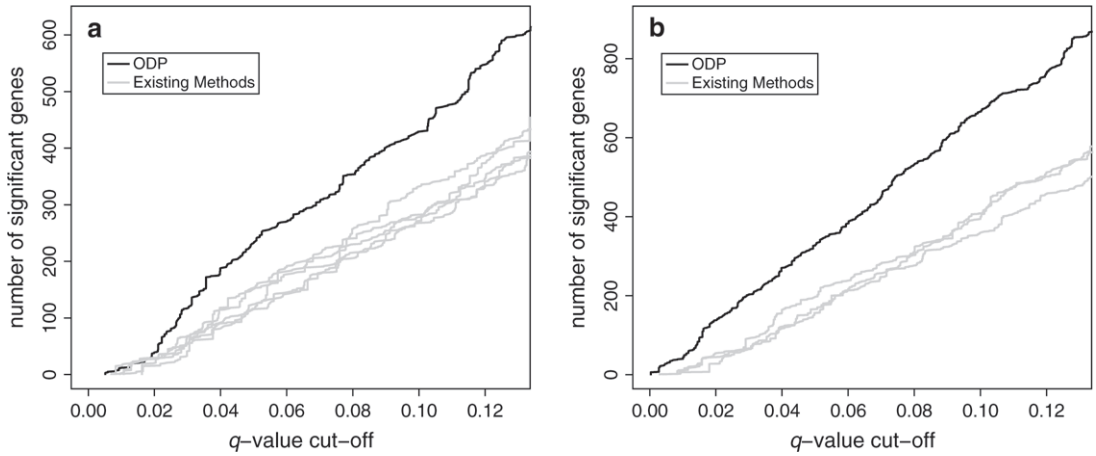

Fig. 2. A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text). The number of genes found to be significant by each method over a range of estimated q-value cutoffs is shown. The methods involved in the comparison are the proposed ODP, SAM, the traditional t-test/F-test, a shrunken t-test/F-test, a nonparametric empirical Bayes "local FDR" method, and a model-based empirical Bayes method. A color version of the figure is given in the supplementary material available at *Biostatistics* online, Figure 9. (a) Results for identifying differential expression between the *BRCA1* and *BRCA2* groups in the Hedenfalk *and others* data. (b) Results for identifying differential expression between the *BRCA1*, *BRCA2*, and Sporadic groups in the Hedenfalk *and others* data. The model-based empirical Bayes method has not been detailed for a three-sample analysis, so it is omitted in this panel.

greater improvements. For example, it provides increases from 123% to 217% in the number of genes called significant at an FDR of 3%. Table 1 shows a number of additional comparisons.

An important point is that it is not surprising that the relative performance of the ODP approach is even better in the three-sample case. The existing methods no longer take into account any asymmetry in the differential expression signal across genes, as they are mostly exactly equivalent to or variations on F-statistics. Whereas in the two-sample setting there are two possible directions for differential expression, there are now six directions in the three-sample setting. The ODP takes advantage of any systematic asymmetry of differential expression in both the two-sample and three-sample settings, whereas it is not possible to do so using any version of an F-statistic. If one were to apply the ODP approach to time course analyses (Storey *and others*, 2005b), then the gains may be even more substantial because in that setting the asymmetry is even harder to quantify using traditional statistics.

5.3 *Biological significance*

In order to determine whether the ODP leads to additional biological information, we considered our findings relative to those of the five existing methods in the context of identifying genes differentially expressed between the *BRCA1* and *BRCA2* groups. It is well known that breast tumors associated with *BRCA1* mutations and *BRCA2* mutations differ greatly from each other in their histological appearance (Lakhani *and others*, 1998). For example, whereas tumors with *BRCA1* mutations exhibit a higher mitotic index and more lymphocytic infiltration, tumors with *BRCA2* mutations are heterogeneous, are of a median or high grade, and show a reduced tubule formation (Lakhani *and others*, 1998). Concordant with these morphological differences, the gene expression profiles of these two types of tumors have also shown to be distinctive (Hedenfalk *and others*, 2001).

At a q-value cutoff of 5%, we found 232 genes to be differentially expressed. Many of the genes that we identified agree with the morphological changes mentioned above. Thirty-six of these genes are known to have functions associated with the cell cycle, including many important molecules such as *PCNA*, cyclin D1 (*CCND1*), cyclin-dependent kinase inhibitor 2C (*CDKN2C*), CDC20 cell division cycle 20 (*CDC20*), CDC28 protein kinase regulatory subunit 2 (*CKS2*), cell division cycle 25B (*CDC25B*), and CHK1 checkpoint (*CHEK1*). The majority of these cell-cycle genes are upregulated in *BRCA1*-positive tumors, except for cyclin D1, whose overexpression in *BRCA2*-associated tumors has been shown to be a useful marker for *BRCA2*-related breast cancer (Hedenfalk *and others*, 2001). Closely related to cell cycle and cell proliferation functions, many genes overexpressed in the *BRCA1* group are found to be associated with apoptosis and genome stability: *P53BP2*, *MSH2*, *PDCD5*, Myc oncogene, and others. Many of these genes have been described in an earlier analysis of this study (Hedenfalk *and others*, 2001).

At a q-value cutoff of 5%, the five existing methods found between 115 and 153 genes to be significant. Almost every gene identified by these other methods is among the 232 genes found by our ODP method. However, we find many more genes with the same error rate. Many important genes would have been missed had we not use the proposed method. Example genes include cell division cycle 25B (*CDC25B*), connective tissue growth factor (*CTGF*), growth factor receptor-bound protein 2, CCAAT/enhancer-binding protein beta (*CEBPB*), among others. In general, the gene ranking of the proposed ODP approach appears to be notably different than that of the other methods. Figure 6 of the supplementary material available at *Biostatistics* online shows the ranking of the top 200 genes from the proposed ODP approach versus each gene's ranking from the other five methods. In the two-sample comparison, genes ranked in the top 100 by the ODP approach were ranked nearly as low as 600 by other methods. In the three-sample comparison, genes ranked in the top 200 by the ODP approach were ranked lower than 400 by other methods.

5.4 *Simulation results*

Similar comparisons were made on simulated data, where one knows with certainty which genes are differentially expressed. Across a range of scenarios, our proposed method continued to perform favorably over the existing methods. It is certainly possible to find some simulation scenario where the "estimated" ODP is outperformed, but this should be distinguished from the fact that it is impossible to outperform the true ODP regardless of which STP one employs. Under certain simulation scenarios, the true ODP can be

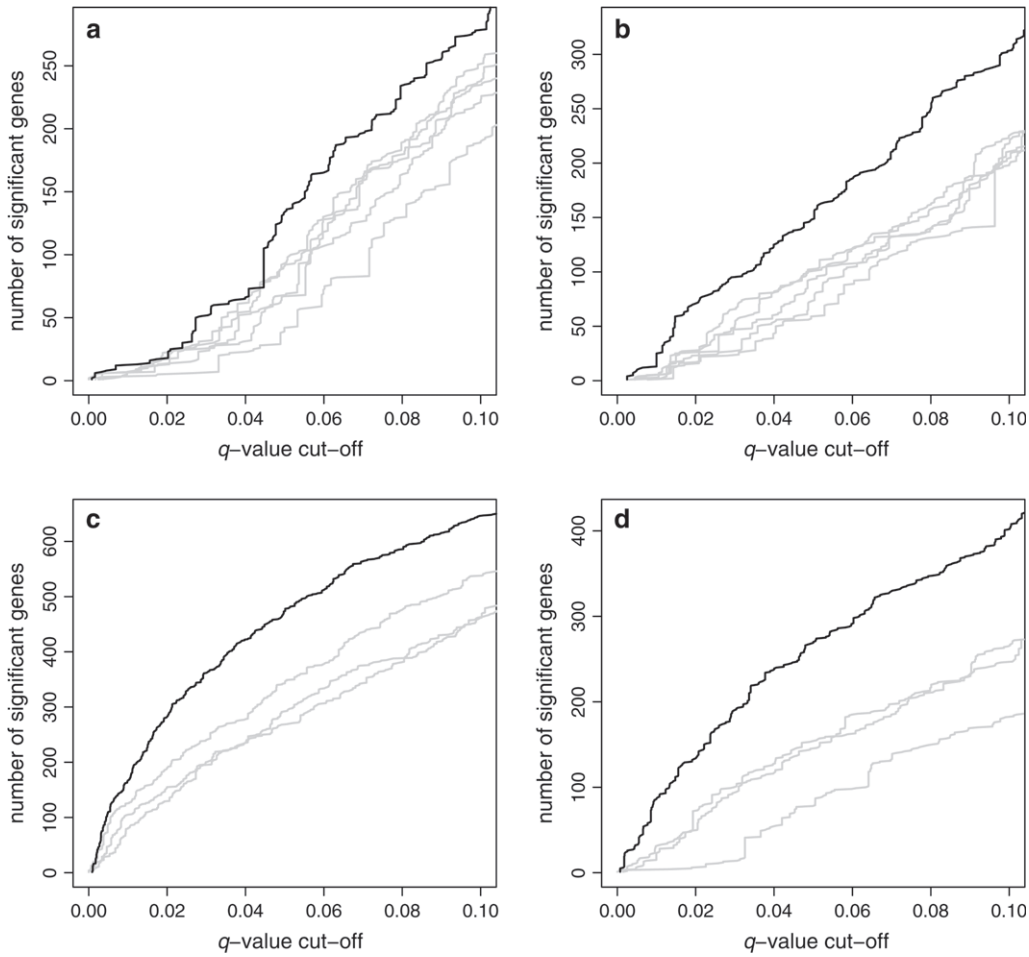

Fig. 3. A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text and Figure 2) based on simulated data. The number of genes found to be significant by each method over a range of estimated q-value cutoffs is shown for a single, representative data set from each scenario. The proposed ODP approach is in black and the other methods are in gray. In general, the data sets increase in complexity from panels (a) to (d). (a) In this scenario, two groups are compared, there is perfectly symmetric differential expression, and the variances are simulated from a unimodal, well-behaved distribution. (b) Two groups are compared, there is moderate asymmetry in the differential expression, and the variances are simulated from a bimodal distribution. (c) Three groups are compared, there is slight asymmetry in differential expression, and the variances are simulated from a unimodal, well-behaved distribution. (d) Three groups are compared, there is moderate asymmetry in differential expression, and the variances are simulated from a bimodal distribution.

reduced to a simple rule. As an extreme example, if data are simulated so that every gene has the same variance, and the signal is symmetric about zero (i.e. if one gene is positively differentially expressed, then there exists another gene with negative differential expression of the same magnitude), then it can be shown that the true ODP reduces to ranking genes by the absolute values of the fold change.

This fact is important to keep in mind when using simulations to evaluate the various procedures. Most of the existing procedures make specific assumptions when deriving their statistics; if these assumptions are enforced in the simulations, then clearly that particular method will be among the top. One advantage of our proposed method is that it does make fairly general assumptions. Because of this, it performed well under a range of scenarios.

We show results from four different scenarios in Figures 3 and 4 in order to give a flavor of the relative performance of the various methods. Both figures are based on the same four simulation scenarios. In moving from scenario (a) to (d), increasingly complicated structure is included in the data.

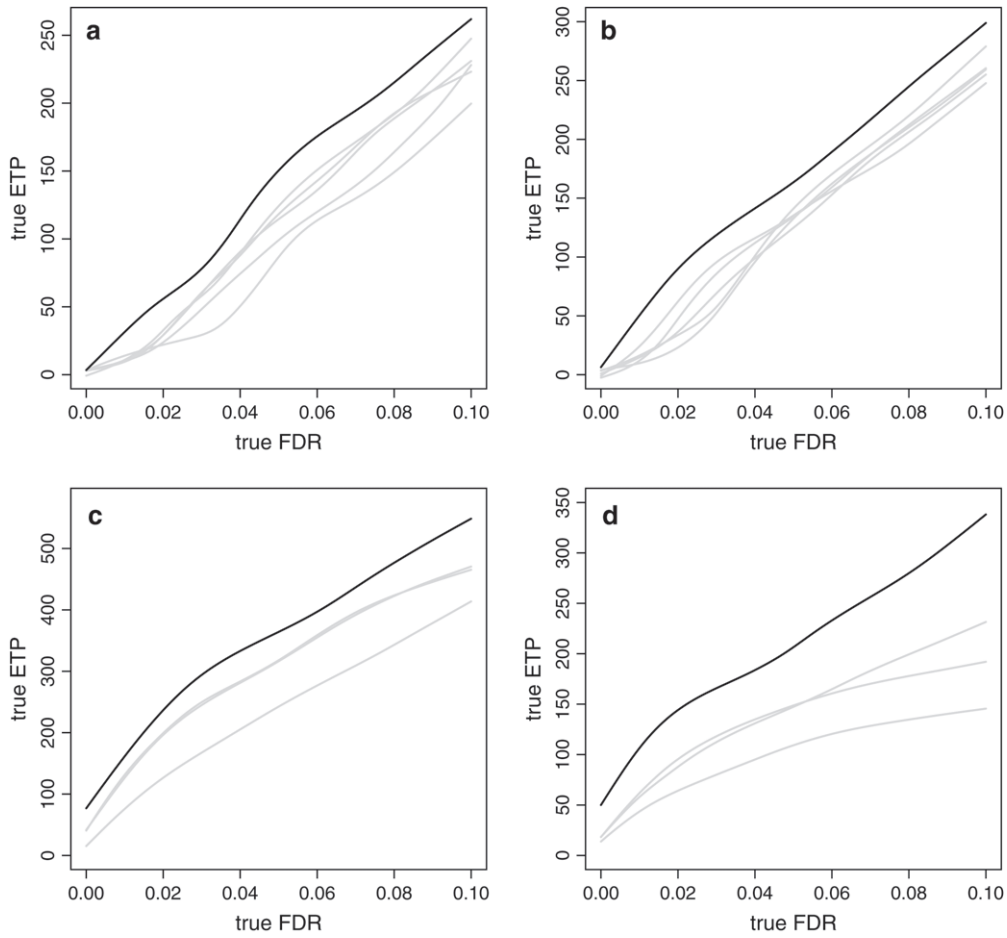

Fig. 4. A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text and Figure 2) based on simulated data. The ETP genes is shown for each true FDR level. As opposed to Figure 3, we have averaged over 100 data sets here and taken into account knowledge of which genes are true and false discoveries in order to make these exact calculations. Panels (a), (b), (c), and (d) are analogous to those in Figure 3.

Scenario (a) is based on data with no high-dimensional structure; the true ODP is more or less equivalent to ranking genes based on absolute fold change. In this scenario, two groups are compared, there is perfectly symmetric differential expression and the variances are simulated from a unimodal, well-behaved distribution.

Scenario (b) has some asymmetry in the differential expression, but the signals and the variances are simulated from distributions similar to those motivating the methods in Lonnstedt and Speed (2002) and Cui *and others* (2005). Two groups are compared, there is moderate asymmetry in the differential expression, and the variances are simulated from a bimodal distribution. In scenario (c), three groups are compared, there is slight asymmetry in differential expression, and the variances are simulated from a unimodal, well-behaved distribution. Similarly, scenario (d) also compares three groups, but there is more asymmetry in differential expression, and the variances are simulated from a bimodal distribution.

All data sets were generated using the R statistical software package; the code used to generate the data can be found in the supplementary material available at *Biostatistics* online. In each scenario, we simulated data from 3000 genes on eight samples from each biological group, where one third of the genes are differentially expressed. These commonalities were enforced and the signal to noise structure was made similar in order to more clearly demonstrate the operating characteristics of our proposed approach and the relative behavior to existing methods. The fact that one third of the genes are differentially expressed does not have a large impact on the relative performance of the various methods. We merely chose this number to closely match the overall signal in the Hedenfalk *and others* (2001) data and to provide enough signal to make the comparisons clearer.

Figure 3 is based on a single set of data from each scenario, where the number of significant genes is plotted against cutoff applied to the estimated q-values. The purpose of this figure is to show that the relative behavior of the various methods shown in Figure 2 on the Hedenfalk *and others* (2001) data can be recapitulated with simulated data. Figure 4 shows results averaged over 100 data sets each, where we have plotted true FDR versus true ETP for each method. This figure compares the relative performance of each method based on knowledge of the true status of each gene, as opposed to the empirical comparisons of Figures 2 and 3.

There are a number of reasons for the less dramatic improvements one sees in Figure 4 relative to Figure 3, including the fact that the y-axis is on a different scale. A major reason is that Figure 4 does not include the fact that in practice the q-values must be estimated for each method. The conservativeness of these estimates is greatly affected by the estimate of the proportion of true nulls (Storey, 2002; Storey *and others*, 2004), which depends on how well the method ranks the least differentially expressed genes. Our proposed approach tends to rank the genes better at both ends, showing the most dramatic improvements when one takes into account both the ranking of the most significant genes and the q-value estimation, which are both necessary in practice.

Finally, we verified that each method does in fact control the FDR. Figure 7 of the supplementary material available at *Biostatistics* online shows the estimated q-values based on Storey and Tibshirani (2003) compared to the true FDR across a relevant range of values. It can be seen that all methods we have considered here, including our proposed method, conservatively estimate the FDR at all estimated q-value cutoffs.

6. Discussion

We have presented a new approach for the significance analysis of thousands of features in a high-dimensional biological study. The approach is based on estimating the optimal procedure for applying a significance threshold to these features, called the ODP. We developed a detailed method that can be used to identify differentially expressed genes in microarray experiments. This method showed substantial

improvements over five of the leading approaches that are currently available. This method is available in the open-source, point-and-click EDGE software package (Leek *and others*, 2006).

Although the basic theoretical ODP result is straightforward to state (Storey, 2005), applying it in practice requires some care. Specifically, one must make sure to avoid overfitting or letting nuisance parameters have a strong effect on the results. We have proposed some simple guidelines here to accomplish this, although each specific application will need to be considered carefully. We used normal probability density functions in our microarray method, mainly because the data are continuous and can be shown to be approximately normal. If one were to analyze some sort of count data, such as that obtained when analyzing genome sequences, then an appropriate distribution such as the Poisson or Binomial can be used instead. Some early investigations indicate that the ODP approach may also offer substantial improvements for tests involving count data. Note that the actual significance can be calculated nonparametrically, so one does not necessarily have to use the correct parametric distribution in order to obtain a good procedure.

An important point is that characterizing the true ODP in a particular application can be a powerful tool for developing an estimated ODP. For example, if every gene's expression has the same variance, and the differential expression signal across genes is perfectly symmetric about zero, then under the normal distribution assumption it can be shown that the true ODP is equivalent to ranking the genes based on the absolute difference in gene expression (i.e. the simple log-scale fold-change criterion). Clearly, this exact situation would never occur in practice, but it stresses the fact that the approach proposed here defines a concrete goal for large-scale significance testing: to estimate the true ODP as well as possible.

In motivating the ODP approach, we described two major steps involved in large-scale significance testing: ranking the features and assigning a significance level to each one. However, for a number of genomics applications, another step may involve deciding exactly what a feature is. For example, in genome-wide tests of association or in protein mass spectrometry analysis, a feature may be a window of adjacent observations, or features may even overlap. These are questions that are also likely to play a major role in developing methods that take full advantage of the high-dimensional nature of the data. We do not claim that the exact method developed for microarrays will serve as an off-the-shelf procedure to apply to any large-scale significance testing problem. However, we do project that the basic ODP framework and some of the tactics that we employed can serve as a useful example for how one approaches these high-dimensional significance analyses.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

REFERENCES

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* **85**, 289–300.

CUI, X. AND CHURCHILL, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.

CUI, X., HWANG, J. T., QIU, J., BLADES, N. J. AND CHURCHILL, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

DUDOIT, S., YANG, Y., CALLOW, M. AND SPEED, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–39.

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**, 1151–60.

HEDENFALK, I., DUGGAN, D., CHEN, Y. D., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M., KALLIONIEMI, O. P. AND OTHERS (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539–48.

KERR, M. K., MARTIN, M. AND CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–37.

LAKHANI, S., JACQUEMIER, J., SLOANE, J., GUSTERSON, B., ANDERSON, T., VAN DE VIJVER, M., FARID, L., VENTER, D., ANTONIOU, A., STORFER-ISSER, A. AND OTHERS (1998). Multifactorial analysis of differences between sporadic breast cancers and cancers involving brca1 and brca2 mutations. *Journal of the National Cancer Institute* **90**, 1138–45.

LEEK, J. T., MONSEN, E. C., DABNEY, A. R. AND STOREY, J. D. (2006). EDGE: Extraction and analysis of differential gene expression. *Bioinformatics* **22**, 507–8.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. Menlo Park, CA: Springer.

LONNSTEDT, I. AND SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

NEWTON, M., KENDZIORSKI, C., RICHMOND, C., BLATTER, F. AND TSUI, K. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–76.

ROCKE, D. M. AND DURBIN, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19**, 966–72.

SEBASTIANI, P., GUSSONI, E., KOHANE, I. S. AND RAMONI, M. F. (2003). Statistical challenges in functional genomics. *Statistical Science* **18**, 33–70.

SORIC, B. (1989). Statistical discoveries and effect-size estimation, *Journal of the American Statistical Association* **84**, 608–10.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–98.

STOREY, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–35.

STOREY, J. D. (2005). The optimal discovery procedure: a new approach to simultaneous significance testing. *UW Biostatistics Working Paper Series, Working Paper 259*. http://www.bepress.com/uwbiostat/paper259/.

STOREY, J. D., DAI, J. Y. AND LEEK, J. T. (2005a). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *UW Biostatistics Working Paper Series, Working Paper 260*. http://www.bepress.com/uwbiostat/paper260/.

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. *Journal of the Royal Statistical Society, Series B* **66**, 187–205.

STOREY, J. D. AND TIBSHIRANI, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–5.

STOREY, J. D., XIAO, W., LEEK, J. T., TOMPKINS, R. G. AND DAVIS, R. W. (2005b). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences* **102**, 12837–42.

TOWNSEND, J. P. AND HARTL, D. L. (2002). Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biology* **3**, research0071.1–0071.16.

TUSHER, V., TIBSHIRANI, R. AND CHU, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences* **98**, 5116–21.

WANG, W. Y. S., BARRATT, B. J., CLAYTON, D. G. AND TODD, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* **6**, 109–18.

ZHONG, S., STORCH, F., LIPAN, O., KAO, M., WEITZ, C. AND WONG, W. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. *Applied Bioinformatics* **3**, 1–5.

[*Received December 20, 2005; first revision April 16, 2006; second revision August 4, 2006; accepted for publication August 22, 2006*]