9-1-2005

The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments

John D. Storey
University of Washington, jstorey@u.washington.edu

James Y. Dai
University of Washington, yud@u.washington.edu

Jeffrey T. Leek
University of Washington, jtleek@gmail.com

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 & 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 & Churchill 2003, Sebastiani et al. 2003, Wang et al. 2005). 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). The ODP was shown to be optimal in that it maximizes the expected number of true positives for each fixed level of expected false positives; 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; (2) assign a significance level to each feature, allowing one to draw a significance cut-off 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 cut-off 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 & 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 et al. 2001), as well as on simulated data. We compare the results to those from five leading differential expression methods (Tusher et al. 2001, Kerr et al. 2000, Dudoit et al. 2002, Cui et al. 2005, Efron et al. 2001, Lonnstedt & 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.

2 The Optimal Discovery Procedure

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 & 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 et al. 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 & Tibshirani 2003). The FDR is defined to be the proportion of false positives among all features called significant (Soric 1989, Benjamini & 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: the expected number of true positives (ETP) and the expected number of false positives (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). This optimality can also be formulated as a multiple test extension of this Neyman-Pearson optimality (Storey 2005).

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

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

The null hypothesis is then rejected if the statistic $\mathcal{S}_{\text{NP}}(\text{data})$ exceeds some cut-off 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 its 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

$$\mathcal{S}_{\text{ODP}}(\text{data}_i) = \frac{\text{sum of probability of data}_i \text{ under each true alternative distribution}}{\text{sum of probability of data}_i \text{ under each true null distribution}}. \quad (1)$$

For a fixed cut-off chosen to attain an acceptable EFP level (or FDR level), each null hypothesis is rejected if its ODP statistic $\mathcal{S}_{\text{ODP}}(\text{data}_i)$ exceeds the cut-off. 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 of 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*, we provide a toy example showing how microarray data contains 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 alternative is true for $i = m_0 + 1, \ldots, m$. In this notation, the ODP statistic of equation (1) is written as:

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

Null hypothesis i is rejected if and only if $\mathcal{S}_{\text{ODP}}(\mathbf{x}_i) \geq \lambda$, where λ is chosen to satisfy an acceptable EFP or FDR level. In practice the exact forms of the 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.

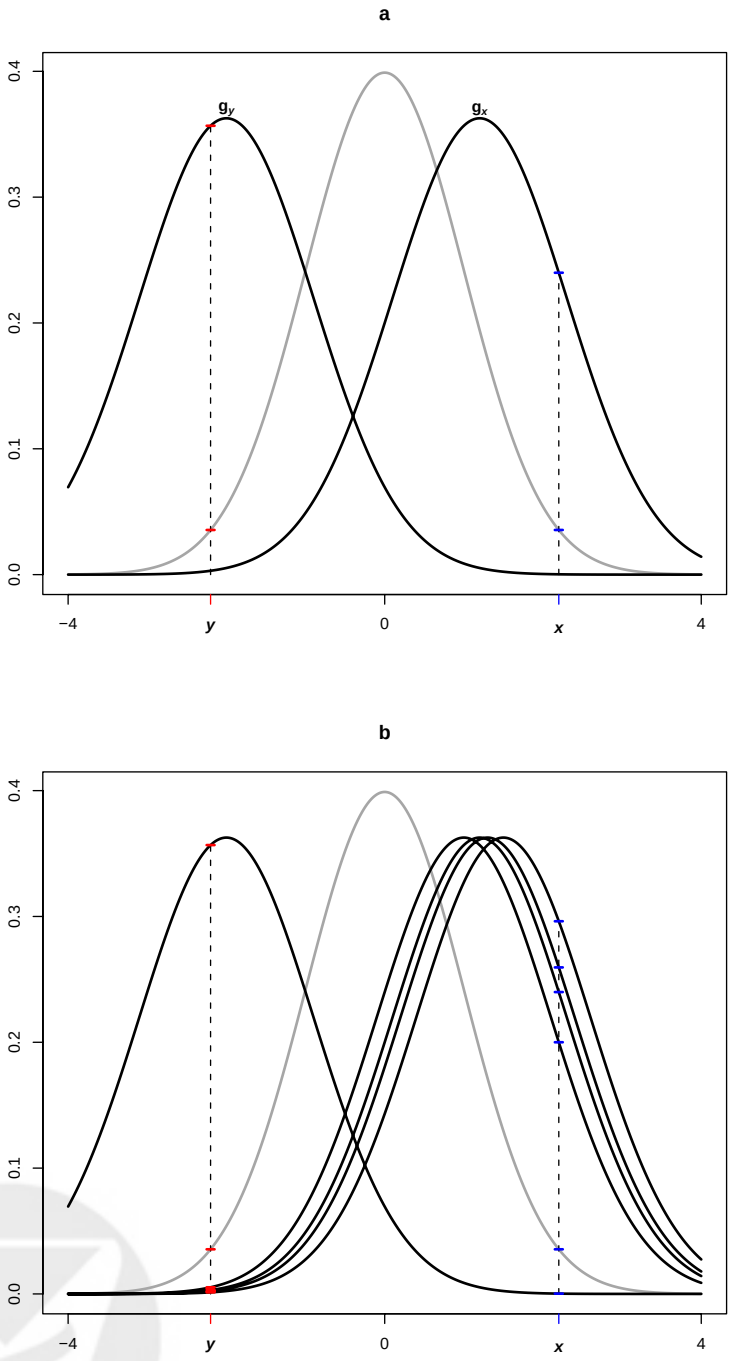

Figure 1: Plots comparing the NP testing approach to the ODP testing approach through a simple example. **(a)** NP approach. The null (grey) and alternative (black) probability density functions of a single test. For observed data x and y, the statistics are calculated by taking the ratio of the alternative 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 (grey) 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 a negative mean.

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):

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x}) + g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x})}, \tag{3}$$

which equals 1 + eq. (2). Since eq. (3) = 1 + eq. (2), these produce the exact same testing procedure [where a threshold of λ applied to the statistic defined in equation (2) is equivalent to a threshold of $1 + \lambda$ applied to the statistic defined in equation (3)]. Because of this equivalence and the tractability of estimating the statistic in equation (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 equation (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 maximum likelihood estimate[1] of f_i based on data $\mathbf{x}_i$ under the constraints of the null hypothesis, and let $\widehat{g_i}$ be the unconstrained maximum likelihood estimate. 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.

[1]Technically speaking, $\widehat{f_i}$ is the version of f_i defined by the unknown parameters' maximum likelihood estimates under the constraints of the null hypothesis.

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 equation (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:

$$\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}) = \frac{\widehat{g}_1(\mathbf{x}) + \cdots + \widehat{g}_{m_0}(\mathbf{x}) + \widehat{g}_{m_0+1}(\mathbf{x}) + \cdots + \widehat{g}_m(\mathbf{x})}{\widehat{f}_1(\mathbf{x}) + \cdots + \widehat{f}_{m_0}(\mathbf{x})}. \tag{4}$$

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

$$\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{f(\mathbf{x})}. \tag{5}$$

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 $\widehat{w}_i = 1$ if $\widehat{f}_i$ is to be included

in the denominator, and $\widehat{w}_i = 0$ otherwise. The estimate of the ODP statistic is then

$$\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \widehat{w}_i \widehat{f}_i(\mathbf{x})}. \tag{6}$$

More generally, the $\widehat{w}_i$ 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 $\widehat{w}_i$ 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 cut-off (i.e., those appearing to be significant and not likely to be true nulls), we set $\widehat{w}_i = 0$; for those not exceeding the cut-off, we set $\widehat{w}_i = 1$. The cut-off is formed so that the proportion falling below and receiving $\widehat{w}_i = 1$ is equal to an estimate of the proportion of true null hypotheses, based on the method in Storey (2002) and Storey & Tibshirani (2003).

Note that if the tests are consistent, then we expect the true alternative tests to rise above the cut-off 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 & 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*). A strong way to enforce nuisance parameter invariance is to require all f_i 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* 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:

$$\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}) = \frac{\sum_{i=1}^{m} \widehat{g}_i(\mathbf{x})}{\sum_{i=1}^{m} \widehat{f}_i(\mathbf{x})}, \tag{7}$$

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 equation (7) poorly behaved.

4 ODP for Identifying Differentially Expressed Genes

For the microarray application, we found the implementation based on our general estimate of equation (6) to perform the best. This implementation requires (i) f_i and g_i to be defined, (ii) estimates $\widehat{f}_i$ and $\widehat{g}_i$ to be derived, (iii) an estimate of which $\widehat{f}_i$ to employ in the denominator to be derived, and (iv) 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

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

For hypothesis i, the likelihood 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*). 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 (equation (6)), where only $\widehat{f}_i$ corresponding to true nulls are present in the denominator. We take the approach outlined in that 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 & Tibshirani (2003).
3. Set $\widehat{w}_i = 1$ for the genes falling in the bottom $\widehat{m}_0$ of the ranking; set $\widehat{w}_i = 0$ 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 et al. 2001). It

should be stressed that this is one of 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 $\widehat{w}_i = 1$. Furthermore, at this stage it is not necessarily so important to identify individual null genes well, but rather to identify a subset so that $\sum_{i=1}^{m} \widehat{w}_i \widehat{f}_i$ 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 re-written 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 (i) the distribution of the σ_i^2 is unrelated to the distribution of the $\mu_{i1} - \mu_{i2}$ and (ii) each $\mu_{i0} = 0$, then we can approximately achieve the nuisance parameter invariance condition (*Supplementary Material*). 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 & 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 $\widehat{\sigma}_{iA}^2$, 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 $i = 1, 2, \ldots, m$:

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

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 3-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 2-sample statistic:

$$\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) = \frac{\sum_{g=1}^{m} \phi(\mathbf{x}_{i1}^*; \widehat{\mu}_{g1}^*, \widehat{\sigma}_{gA}^2) \cdots \phi(\mathbf{x}_{iK}^*; \widehat{\mu}_{gK}^*, \widehat{\sigma}_{gA}^2)}{\sum_{g=1}^{m} \widehat{w}_i \phi(\mathbf{x}_i^*; 0, \widehat{\sigma}_{g0}^2)}. \tag{8}$$

Analogously to the two-sample method, each gene is mean centered around zero to obtain the transformed data $\mathbf{x}_i^*$. In the 1-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 towards a common value (across genes), then the so-called SAM statistic and other similar versions emerge (Tusher et al. 2001, Cui et al. 2005, Efron et al. 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 & Tibshirani 2003). Full details of this algorithm, including exact formulas can be found in the *Supplementary Material*. 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*.

Proposed Algorithm for Identifying Differentially Expressed Genes

1. Using the formula given above in equation (8), 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 re-compute 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 & 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 & Tibshirani 2003). One may call genes significant for differential expression by forming a q-value cut-off 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 et al. 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 & 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 (i) the highly-used SAM software based on Tusher et al. (2001) and Storey (2002), (ii) the traditional t-tests and F-tests as previously suggested for microarray analysis (Kerr et al. 2000, Dudoit et al. 2002), (iii) a recently proposed variation on these that uses "shrunken" versions of the statistics (Cui et al. 2005), (iv) a non-parametric Bayesian method whose estimated posterior probabilities are also sometimes interpreted as estimated Bayesian "local FDR" estimates (Efron et al. 2001), and (v) a model-based empirical Bayes method giving posterior probabilities of differential expression (Lonnstedt & 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 information to estimate q-values for each gene according to previously described methodology (Storey 2002, Storey & 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 et al. 2001, Townsend & Hartl 2002, Newton et al. 2004) for identifying differentially expressed genes. In particular, Newton et al. (2004) offers an interesting semi-parametric empirical Bayes approach that provides an estimate of a novel 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 et al. (2004) and three of the methods we include in our comparison (Tusher et al. 2001, Efron et al. 2001, Lonnstedt & Speed 2002) are able to capture asymmetry in differential expression signal for when comparing two groups. Tusher et al. (2001) and Efron et al. (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.

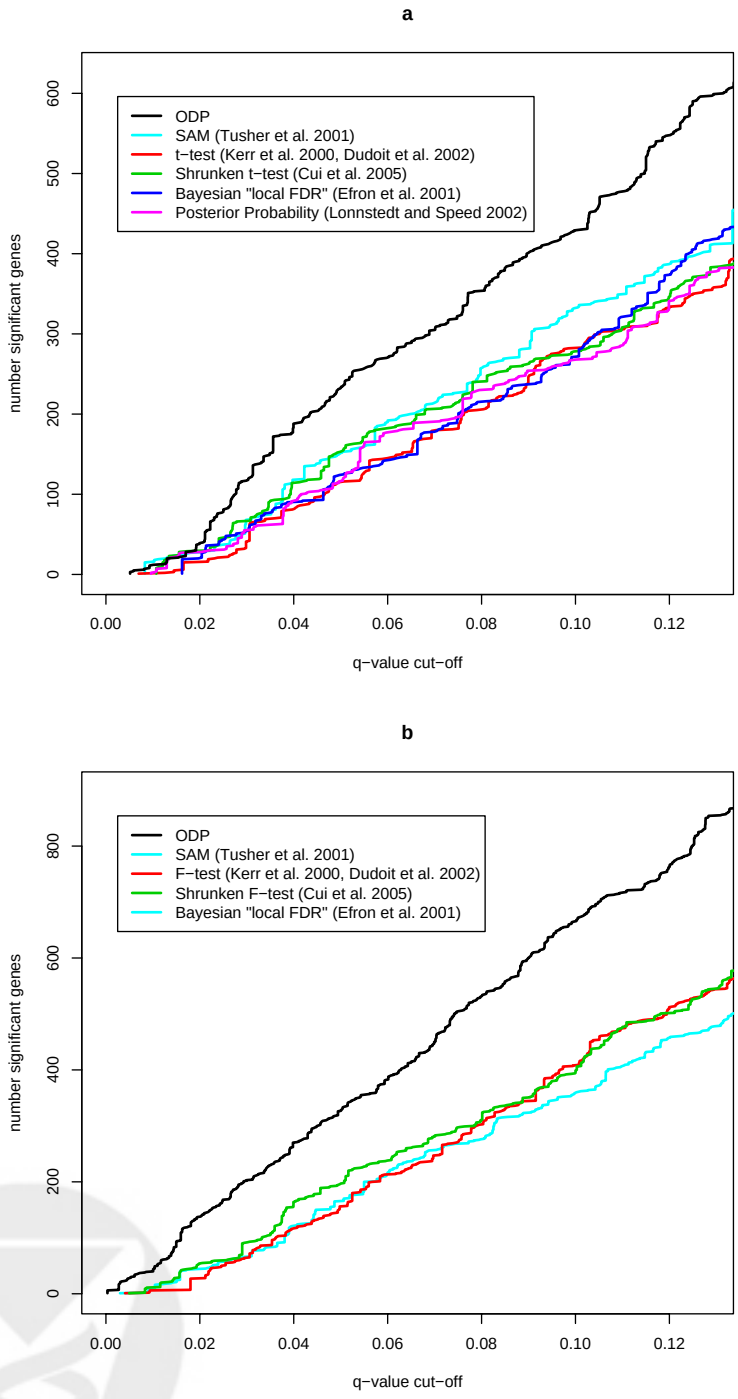

Figure 2: A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text). The number of genes found to be significant by each method over a range of estimated q-value cut-offs is shown. The methods involved in the comparison are the proposed ODP (black), SAM (turquoise), the traditional t-test/F-test (red), a shrunken t-test/F-test (green), a non-parametric empirical Bayes "local FDR" method (a: blue, b: turquoise), and a model-based empirical Bayes method (fuchsia). **(a)** Results for identifying differential expression between the *BRCA1* and *BRCA2* groups in the Hedenfalk et al. data. **(b)** Results for identifying differential expression between the *BRCA1*, *BRCA2*, and Sporadic groups in the Hedenfalk et al. data. The model-based empirical Bayes methods have not been detailed for a 3-sample analysis, so they are omitted in this panel.

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 cut-offs, which gives an estimate of the relative ETP levels at each given FDR (*Supplementary Material*). For the methods employed here, this is equivalent to comparing the ETP for each fixed EFP level or p-value cut-off 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 cut-offs.

In testing for differential expression between the *BRCA1* and *BRCA2* groups, the ODP approach shows surprisingly large improvements in performance over existing methods. For example, at a 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 cut-offs less than or equal to 10% ranges from 43–87% across all methods. In testing for 3-sample differential expression among the *BRCA1*, *BRCA2*, and Sporadic groups, the ODP approach offers even greater improvements. For example, it provides increases from 123–217% in the number of genes called significant at a false discovery rate 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 3-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 2-sample setting there are two possible directions for differential expression, there are now six directions in the 3-sample setting. The ODP takes advantage of any systematic asymmetry of differential expression in both the 2-sample and 3-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 et al. 2005), 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

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 $2\%, 3\%, \ldots, 10\%$. The exact same FDR methodology (Storey 2002, Storey & Tibshirani 2003) was applied to each thresholding method in order to make the comparisons fair. The model-based Bayesian methods (Lonnstedt & Speed 2002) is not defined for a 3-sample analysis, so that case is omitted.

	% Increase by ODP – 2-sample			% Increase by ODP – 3-sample		
Thresholding Method	Minimum	Median	Maximum	Minimum	Median	Maximum
SAM (Tusher et al. 2001)	29	43	72	76	92	211
t/F-test (Dudoit et al. 2002, Kerr et al. 2000)	52	86	185	63	82	407
Shrunken t/F-test (Cui et al. 2005)	34	52	77	61	69	154
Bayesian "local FDR" (Efron et al. 2001)	58	87	117	76	92	211
Posterior probability (Lonnstedt & Speed 2002)	44	60	113	–	–	–

appearance (Lakhani et al. 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 et al. 1998). Concordant with these morphological differences, the gene expression profiles of these two types of tumors have also shown to be distinctive (Hedenfalk et al. 2001).

At a q-value cut-off 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 up-regulated in *BRCA1* positive tumors, except for cyclin D1, whose over-expression in *BRCA2* associated tumors has been shown to be a useful marker for *BRCA2* related breast cancer (Hedenfalk et al. 2001). Closely related to cell cycle and cell proliferation functions, many genes over-expressed 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 et al. 2001).

At a q-value cut-off of 5%, the five existing methods found between 115–153 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* 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 simultaneous-thresholding rule one employs. Under certain simulation scenarios, the true ODP can be 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 (that is, 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. 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 & Speed (2002) and Cui et al. (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

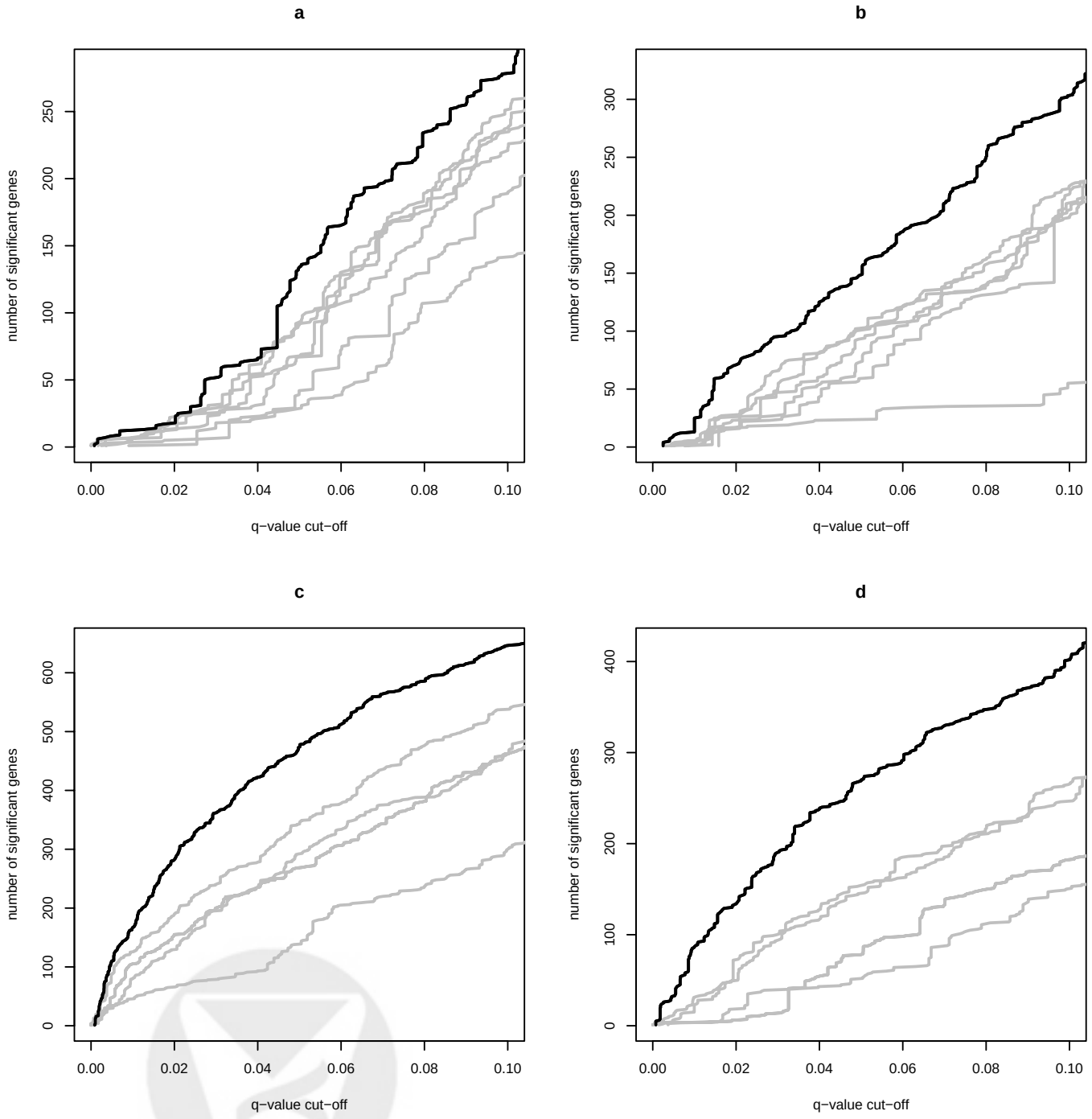

Figure 3: A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text and Figure 2) based on simulated data. The number of genes found to be significant by each method over a range of estimated q-value cut-offs is shown for a single, representative data set from each scenario. The proposed ODP approach is in black and the other methods are in grey. 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.

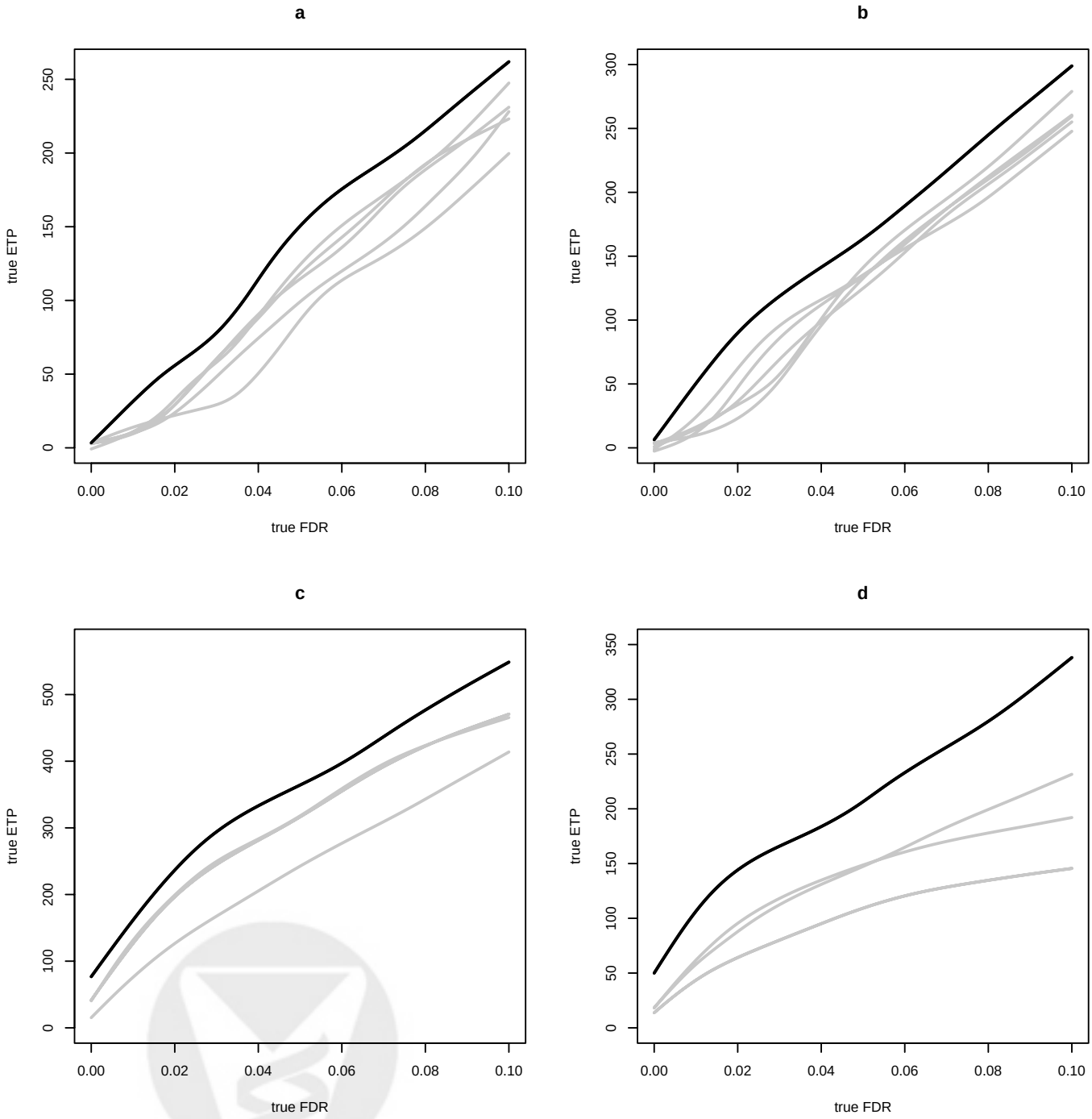

Figure 4: A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text and Figure 2) based on simulated data. The expected number of true positive (ETP) genes is shown for each true FDR level. As opposed to Figure 3, we have averaged over 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.

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*. 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 et al. (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 cut-off 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 et al. (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 et al. 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* shows the estimated q-values based on Storey & 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 cut-offs.

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 optimal discovery procedure (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 available at `http://faculty.washington.edu/jstorey/edge/`.

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 over-fitting 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.

References

Benjamini, Y. & 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. & 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. & 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. & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* **12**: 111–139.

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

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. & Trent, J. (2001). Gene-expression profiles in hereditary breast cancer, *New England Journal of Medicine* **344**: 539–548.

Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology* **7**: 819–837.

Lakhani, S., Jacquemier, J., Sloane, J., Gusterson, B., Anderson, T., van de Vijver, M., Farid, L., Venter, D., Antoniou, A., Storfer-Isser, A., Smyth, E., Steel, C., Haites, N., Scott, R., Goldgar, D., Neuhausen, S., Daly, P., Ormiston, W., McManus, R., Scherneck, S., Ponder, B., Ford, D., Peto, J., Stoppa-Lyonnet, D., Easton, D. & et al. (1998). Multifactorial analysis of differences between sporadic breast cancers and cancers involving brca1 and brca2 mutations., *J Natl Cancer Inst* **90**: 1138–45.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edn, Springer-Verlag.

Lonnstedt, I. & Speed, T. (2002). Replicated microarray data, *Statistica Sinica* **12**: 31–46.

Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method, *Biostatistics* **5**: 155–176.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. & 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.

Rocke, D. M. & Durbin, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data, *Bioinformatics* **19**: 966–972.

Sebastiani, P., Gussoni, E., Kohane, I. S. & 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**: 608–610.

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

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

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., Taylor, J. E. & 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. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. & Davis, R. W. (2005). Significance analysis of time course microarray experiments, *Proceedings of the National Academy of Sciences* **102**: 12837–12842.

Townsend, J. P. & 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. & Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Sciences* **98**: 5116–5121.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns, *Nat Rev Genet* **6**: 109–118.

Zhong, S., Storch, F., Lipan, O., Kao, M., Weitz, C. & Wong, W. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space, *Applied Bioinformatics* **3**: 1–5.

Supplementary Material:

The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments

John D. Storey*, James Y. Dai, and Jeffrey T. Leek
Department of Biostatistics
University of Washington

Contents

*Address for correspondence: jstorey@u.washington.edu

7 A Simple Motivating Example

The following toy example provides some intuition into the operating characteristics of the ODP in the context of high-dimensional biological studies. Suppose that an expression study is performed on 15 individuals, seven of which come from one group and eight from another, where the goal is to identify genes that are differentially expressed between these two groups. This design reflects the breast cancer study we consider below. Figure 5 shows a heat map of simulated expression data over 1000 genes under this study design, where the genes have been hierarchically clustered (Eisen et al. 1998). It can be seen that there is substantial structure among the differentially expressed genes. Most obviously, there is asymmetry in the differential expression: more genes are over-expressed in Group 2 than in Group 1. However, among the differentially expressed genes there are three distinct patterns. Some of these patterns make it more straightforward to detect differential gene expression than others. Moreover, the more genes with a common differential expression pattern, the more fruitful it is (in terms of the ETP to EFP trade-off) to call these genes differentially expressed.

The ODP takes this kind of structure into account, and uses it to optimally extract the differential expression signal from the noise. The distinct patterns of differential expression are denoted in Figure 5. The genes present in each cluster will have similar likelihood functions. Moreover, some types of likelihood functions will be more distinct from the null likelihoods than others. The ODP considers the data for each gene and evaluates it at the true likelihoods, forming a ratio of the sum of the data evaluated at the true alternative likelihoods to that of the true null likelihoods. The precision to which the structure can be captured depends on the level of complexity of the model for the data defining the likelihood functions.

Note that this type of structure will be present in other high-dimensional biological studies. For example, in population-based genetic tests of association, contiguous SNPs will show similar genotypic patterns. Therefore, regions showing true associations with a trait of interest will do so in a similar manner.

8 Detailed algorithm for identifying differentially expressed genes

The following is a detailed description of the full algorithm for identifying differentially expressed genes that was presented in the main text.

Let x_{ij} be the expression observation for gene i in array j, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The data for a single gene is written as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. Assume there are K groups tested for differential expression, and let $\mathbf{x}_{ik}$ be the subset of data from group k, $k = 1, \ldots, K$. Finally, let $\mathcal{G}_k$ be the set of arrays corresponding to group k so that $\mathbf{x}_{ik} = (x_{ij})_{j \in \mathcal{G}_k}$. Gene i in group k has

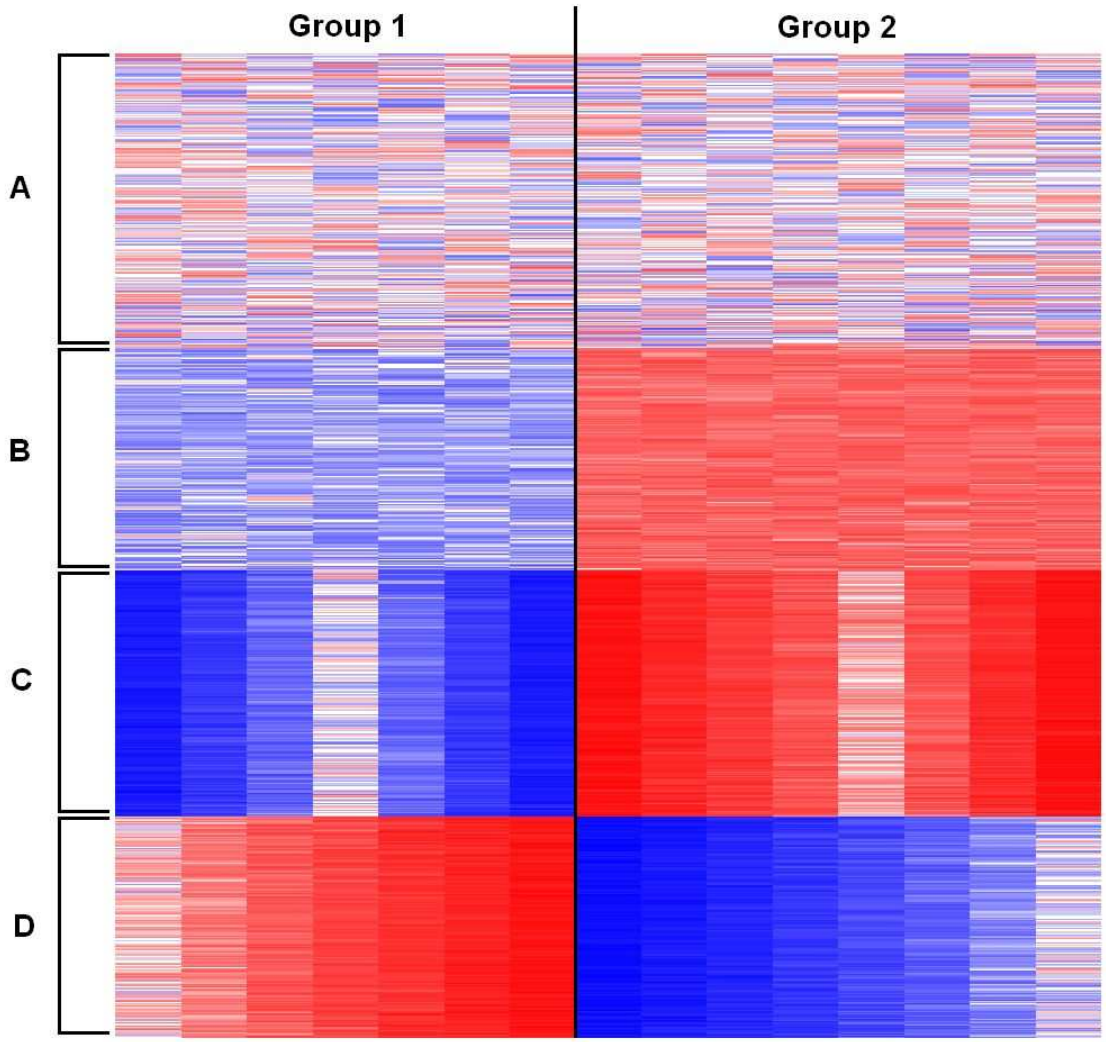

Figure 5: Simulated expression data showing structure in the biological signal of interest. A heat map of expression data on 1000 genes among 15 individuals is shown (red = high, blue = low). The first seven individuals come from Group 1 and the last eight from Group 2. The genes were hierarchically clustered, and it can be seen that four distinct clusters of genes emerge: **(A)** No differential expression; **(B)** Moderate over-expression in Group 2, low variance; **(C)** Strong over-expression in Group 2, large variance; **(D)** Strong over-expression in Group 1, large variance. Further, there is pervasive asymmetry in differential expression towards Group 2. The proposed ODP approach captures this structure and uses it to optimally separate signal from noise in identifying differentially expressed genes.

mean gene expression μ_{ik}, and variance σ_i^2. Without loss of generality, we define the mean when the null hypothesis of no differential expression is true to be $\mu_{i0} = \sum_{k=1}^{K} n_k \mu_{ik}/n$, where n_k is the number of arrays in group k.

Step 1: Calculating ODP statistics. Let $(\widehat{\mu}_{i0}, \widehat{\sigma}_{i0}^2)$ be estimates under the constraints of the null hypothesis, and $(\widehat{\mu}_{i1}, \ldots, \widehat{\mu}_{iK}, \widehat{\sigma}_{i1}^2)$ be unconstrained estimates. These are defined as follows:

$$\widehat{\mu}_{i0} = \sum_{j=1}^{n} x_{ij}/n; \qquad \widehat{\sigma}_{i0}^2 = \sum_{j=1}^{n} \frac{(x_{ij} - \widehat{\mu}_{i0})^2}{n-1}$$

$$\widehat{\mu}_{ik} = \sum_{j \in \mathcal{G}_k} x_{ij}/n_k; \qquad \widehat{\sigma}_{iA}^2 = \sum_{k=1}^{K} \sum_{j \in \mathcal{G}_k} \frac{(x_{ij} - \widehat{\mu}_{ik})^2}{n-K}$$

Note that these are the typical estimates for Normally distributed data. The estimated weights $\widehat{w}_i$ for inclusion of each null density in the denominator of the statistic are calculated as detailed in the main text.

When $K = 1$ and differential expression is defined to be average expression not equal to zero (which would be the case when examining the log ratios of expression from a direct comparison using two-channel microarrays), the estimated ODP statistic for gene i $(i = 1, \ldots, m)$ is

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

When $K > 1$ each gene is centered as a step in approximately achieving "nuisance parameter invariance" as described in the main text. Define $x_{ij}^* = x_{ij} - \widehat{\mu}_{i0}$ and $\widehat{\mu}_{ik}^* = \widehat{\mu}_{ik} - \widehat{\mu}_{i0}$. The estimated ODP statistic for each gene i is then

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

Centering each gene induces a slight dependence among the x_{ij}^* within a gene, but this can be taken into account by modifying the definition of the Normal densities ϕ. However, this turns out to be algebraically proportional to the original definition, so no modification is actually necessary. Note that by centering each gene, we lose no information about differential expression.

Step 2: Simulating null statistics. We obtain null statistics by applying the standard bootstrap procedure for generating the null distribution when testing the location parameter(s) of a distribution (Efron & Tibshirani 1993). Let $\boldsymbol{\epsilon}_i$ be the alternative model residuals obtained from each gene i's expression data $\mathbf{x}_i$ by setting $\epsilon_{ij} = x_{ij} - \widehat{\mu}_{ik(j)}$ where $k(j)$ is the group to which array j belongs for $i = 1, \ldots, m$. For each gene i, a bootstrap null set of data for gene i, $\mathbf{x}_i^0$, is obtained

by resampling n observations with replacement from among the ϵ_{ij} and adding these back to the estimated null mean $\widehat{\mu}_{i0}$. For B iterations, bootstrap from the null distribution the expression data and re-compute each statistic to get a set of null statistics $\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b})$ for $b = 1, \ldots, B$ and $i = 1, \ldots, m$. Each bootstrap sampling is applied to all genes, keeping the dependence structure of the genes intact.

It should be noted that the standard permutation null scheme (see, for example, that carried out in Storey & Tibshirani 2003) could be applied here as well. However, the ODP statistic does not carry the same pivotality properties as, say, a t-statistic. Therefore, the null permuted data from a gene with a strong differential expression signal can produce null data with a much larger variance than its true null. If many genes have a strong signal, then the null statistics from these genes are not representative of the true null distributions. Since the bootstrap null removes the signal from each gene before resampling, we have found the bootstrap approach to be more reliable in this setting. It appears more research into this issue is warranted.

Step 3: Estimating q-values. According to the ODP approach, each possible significance cut-off is formed by calling all genes significant with $\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) \geq c$ for some cut-point c. The algorithm for estimating q-values presented in Storey (2002) and Storey & Tibshirani (2003) is written in terms of p-values. We show here that if p-values are calculated for each gene in a certain fashion, then one can employ the existing p-value based q-value estimation so that direct thresholding of the statistics actually takes place. This avoids the need to re-develop q-value estimation and the theory justifying it, while allowing us to form ODP statistic thresholds (i.e., $\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) \geq c$) rather than p-value based thresholds.

Suppose that the p-value for gene i is calculated by

$$p_i = \frac{\sum_{b=1}^{B} \sum_{j=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_j^{0b}) \geq \widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i)\right)}{m \cdot B}, \tag{10}$$

where $1(\cdot)$ is standard indicator function equal to one when the argument is true and zero otherwise. When p-values are calculated in this pooled, gene non-specific way, the subsequent q-value estimation procedure as defined in Storey (2002) is equivalent to estimating q-values by directly thresholding the statistics. The following estimate of the false discovery rate when calling all p-values $\leq t$ significant is implicit in the algorithm for estimating q-values (Storey 2002, Storey & Tibshirani 2003):

$$\widehat{\text{FDR}}(t) = \frac{\widehat{\pi}_0 m \cdot t}{\sum_{i=1}^{m} 1\left(p_i \leq t\right)}.$$

For a fixed significance cut-off c applied to the original statistics, the analogous false discovery

rate estimate is

$$\widehat{\text{FDR}}(c) = \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b}) \geq c\right)/B}{\sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) \geq c\right)}, \tag{11}$$

where $\widehat{\pi}_0$ is derived from a smoother fit to the

$$\widehat{\pi}_0(c') = \frac{\sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) < c'\right)}{\sum_{b=1}^{B} \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b}) < c'\right)/B}$$

over some range of c' exactly as in the algorithm given in Storey & Tibshirani (2003).

As was stated in Storey & Tibshirani (2003), the original FDR estimate of Storey (2002) is easily shown to be equivalent to the above formula when p-values are calculated as above. The key observation is that one can equivalently define the Type I error rate of a given cut-off by $\sum_{b=1}^{B} \#\{\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b}) \geq c\}/(m \cdot B)$ rather than the p-value threshold t. In fact, if we define

$$c(t) \equiv \min\{\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) : p_i \leq t\}$$

then it can be shown that

$$\frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b}) \geq c(t)\right)/B}{\sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) \geq c(t)\right)} = \frac{\widehat{\pi}_0 m \cdot t}{\sum_{i=1}^{m} 1\left(p_i \leq t\right)}$$

making the two false discovery rate estimates equal. Therefore, q-values derived from either method are equal as long as the p-values are calculated from the gene non-specific empirical distribution of the simulated null statistics.

Out of these derivations come direct estimates for the EFP and ETP for each ODP threshold. In particular, for a threshold c define

$$\begin{aligned} \widehat{\text{EFP}}(c) &= \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b}) \geq c\right)}{B} \\ \widehat{\text{ETP}}(c) &= \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i) \geq c\right) - \widehat{\text{EFP}}(c) \end{aligned}$$

where $\widehat{\pi}_0$ is estimated as above. The FDR estimate from equation (11) can then be written in terms of these estimates, further showing the direct connection between the EFP, ETP, and FDR:

$$\widehat{\text{FDR}}(c) = \frac{\widehat{\text{EFP}}(c)}{\widehat{\text{EFP}}(c) + \widehat{\text{ETP}}(c)}. \tag{12}$$

We estimate q-values for our ODP approach in a new way. We pool simulated null statistics across genes as above, but we do not employ the null statistics from every gene. Specifically, we only use null statistics from genes with $\widehat{w}_i = 1$, i.e., those represented in the denominator of the statistic. We have found this produces more well behaved estimates of the q-values over using null statistics from every gene. In implementing this approach, the above formulas are simply replaced with the proper subset of null statistics. For a fixed significance cut-off c applied to the original statistics, the EFP and ETP estimates are:

$$\begin{aligned}\widehat{\mathrm{EFP}}(c) &= \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \sum_{i=1}^{m} \widehat{w}_i 1\left(\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i^{0b}) \geq c\right)}{B \sum_{i=1}^{m} \widehat{w}_i / m}, \\ \widehat{\mathrm{ETP}}(c) &= \sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i) \geq c\right) - \widehat{\mathrm{EFP}}(c).\end{aligned}$$

The estimates $\widehat{\pi}_0(c')$ are analogously modified to

$$\widehat{\pi}_0(c') = \frac{\sum_{i=1}^{m} 1\left(\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i) < c'\right)}{\frac{\sum_{b=1}^{B} \sum_{i=1}^{m} \widehat{w}_i 1\left(\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i^{0b}) < c'\right)}{B \sum_{i=1}^{m} \widehat{w}_i / m}}.$$

The overall estimate $\widehat{\pi}_0$ is formed by smoothing over some range of c' exactly as in the above algorithm. We then plug these estimates into equation (12) in order to estimate FDR for a given threshold c. Finally, the q-value estimate for each gene i is:

$$\widehat{q}_i = \min_{c \leq \widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i)} \widehat{\mathrm{FDR}}(c),$$

i.e., the minimum estimated FDR among all thresholds where gene i is called significant.

9 Comparing procedures based on the number of genes called significant

The ODP approach was compared to five leading procedure for identifying differentially expressed genes by comparing the number of genes called significant at each FDR level. It is straightforward to show that this is an empirical version of the comparison based on the ETP for each fixed FDR. This follows since $\widehat{\mathrm{ETP}} = (\#\text{ significant genes})(1 - \widehat{\mathrm{FDR}})$, as just shown above. Since each method is compared at the same value of $\widehat{\mathrm{FDR}}$, it follows that comparing the number of significant genes is equivalent to comparing the methods based on $\widehat{\mathrm{ETP}}$, which we showed above provides a valid estimate of the true ETP. Note that it can also be shown based on these arguments that this

comparison gives equivalent information about relative performance based on comparing $\widehat{\text{ETP}}$ for each fixed $\widehat{\text{EFP}}$ level.

10 Nuisance parameter invariance

The ODP is most simply defined in terms of the following rule (Storey 2005):

$$\frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x})}.$$

If each hypothesis test has identically defined null and alternative hypotheses then differences between the f_i would be due to nuisance parameters. For example, consider the 2-sample microarray problem where the null hypothesis for each test is that $\mu_{i1} = \mu_{i2}$ and the alternative is $\mu_{i1} \neq \mu_{i2}$. Above, we defined $\mu_{i0} = (n_1\mu_{i1} + n_2\mu_{i2})/n$, which is the common mean when the null hypothesis is true. Under the Normal distribution assumption, the ODP rule is based on

$$\frac{\sum_{i=m_0+1}^{m} \phi(\mathbf{x}; \mu_{i1}, \mu_{i2}, \sigma_i^2)}{\sum_{i=1}^{m_0} \phi(\mathbf{x}; \mu_{i0}, \sigma_i^2)}, \tag{13}$$

where as before tests $1, 2, \ldots, m_0$ have true null hypotheses and the remainder have true alternative hypotheses. Differences between the null densities $\phi(\mathbf{x}; \mu_{i0}, \sigma_i^2)$ are due to differing μ_{i0} and σ_i^2, which are not used at all in defining the null and alternative hypotheses. The parameters μ_{i0} and σ_i^2 are therefore nuisance parameters.

In the Neyman-Pearson setting, nuisance parameters are usually "canceled out" in some fashion, making them irrelevant in the hypothesis tests. In practice, "pivotal statistics" are desirable because their null distributions do not depend on any unknown nuisance parameters. In the single significance test setting, nuisance parameters are most troublesome in that they make it more difficult to calculate a null distribution. In the ODP setting, the presence of nuisance parameters is troublesome for another reason: since the ODP is defined in terms of the *true* likelihood of each test, one can manipulate the ODP quite substantially by varying the degree by which the nuisance parameters values differ between the true null and true alternative tests.

Specifically, consider the above statistic in equation (13) under the scenario where $\mu_{10} = \cdots = \mu_{m_0,0} = -1000$ and $\mu_{m_0+1,0} = \cdots = \mu_{m0} = 1000$, as opposed to the scenario where $\mu_{10} = \cdots = \mu_{m0} = 0$. Clearly these two scenarios would yield very different results. In the former case, it would be much easier to distinguish the true null hypotheses from the true alternative hypotheses. However, in practice it is not clear how much this matters since in the former case, one would not be able to estimate the ODP nearly as well. Similar examples can be constructed in terms of the

nuisance parameters σ_i^2. We have also found that certain types of nuisance parameter effects can lead to over-fitting of the data in the significance testing (see below). Therefore, it is desirable from a variety of perspectives to eliminate these effects as much as possible.

In the context of this Normal distribution example, one way to avoid effects from nuisance parameters is to transform the data so that the null distributions are all equal to the $N(0,1)$ distribution. This can be done by replacing x_{ij} with $(x_{ij} - \mu_{i0})/\sigma_i$. In practice, this could be accomplished instead with estimated values, $(x_{ij} - \widehat{\mu}_{i0})/\widehat{\sigma}_i$. The null distribution of every gene would then approximately be $N(0,1)$. This is obviously an extreme form of what we call "nuisance parameter invariance" because all nuisance parameters have been removed from the data. In our experience, this particular choice does not work well because there is relevant information in the σ_i^2, and dividing the data by $\widehat{\sigma}_i$ induces a lot of extra noise into the expression measurements.

A weaker criterion for nuisance parameter invariance involves a type of subset exchangeability across null distributions. In particular, we require that the average null likelihood among all tests is equal to that from the true null tests: $\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m_0} f_i/m_0$. This implies that the likelihoods of the true nulls cannot be pathologically different from the true alternatives simply because of nuisance parameter values. In the Normal example, one may approximately achieve this property by forcing all $\mu_{i0} = 0$ (leading to no loss of information or addition of noise) and removing any relationship between the signal $\mu_{i1} - \mu_{i2}$ and the variances σ_i^2.

Let $\mathbf{x}^*$ be the mean centered data for a single gene (thereby removing the effect of μ_{i0}), and let $\mu_{i1}^* = \mu_{i1} - \mu_{i0}$, $\mu_{i2}^* = \mu_{i2} - \mu_{i0}$, $\mu_{i0}^* = 0$. In the case that $\sum_{i=1}^{m} \phi(\cdot; 0, \sigma_i^2)/m = \sum_{i=1}^{m_0} \phi(\cdot; 0, \sigma_i^2)/m_0$, the following statistics are all equivalent:

$$\frac{\sum_{i=m_0+1}^{m} \phi(\mathbf{x}^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2)}{\sum_{i=1}^{m_0} \phi(\mathbf{x}^*; 0, \sigma_i^2)} \qquad \frac{\sum_{i=m_0+1}^{m} \phi(\mathbf{x}^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2)}{\sum_{i=m_0+1}^{m} \phi(\mathbf{x}^*; 0, \sigma_i^2)}$$

$$\frac{\sum_{i=m_0+1}^{m} \phi(\mathbf{x}^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2)}{\sum_{i=1}^{m} \phi(\mathbf{x}^*; 0, \sigma_i^2)} \qquad \frac{\sum_{i=1}^{m_0} \phi(\mathbf{x}^*; 0, \sigma_i^2) + \sum_{i=m_0+1}^{m} \phi(\mathbf{x}^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2)}{\sum_{i=1}^{m} \phi(\mathbf{x}^*; 0, \sigma_i^2)}$$

The fact that the two on the top row are equivalent is reassuring in that the true null and true alternative hypotheses do not differ in their average likelihoods due to nuisance parameters. The bottom two statistics show the transition from the original ODP (top left) to the straightforwardly estimated ODP (bottom right), which was used to motivate our proposed microarray method.

In order to approximately obtain the condition $\sum_{i=1}^{m} \phi(\cdot; 0, \sigma_i^2)/m = \sum_{i=1}^{m_0} \phi(\cdot; 0, \sigma_i^2)/m_0$, first mean center the data for each test. Then perform a proper transformation so that there is no apparent relationship between the difference in average expression between the two groups and the sample variances. This latter step has been well-studied in general and in the context of microarrays (Rocke & Durbin 2003).

11 Over-fitting

In a single test procedure, the null statistic is calculated under the assumption that the data come from the null distribution. When the statistic involves estimation of parameters, the estimation is carried out with null data when calculating null statistics. For example, suppose that a generalized likelihood ratio statistic, $\widehat{g}(\mathbf{x})/\widehat{f}(\mathbf{x})$, is formed, and a resampling based p-value is to be calculated. This involves randomly resampling the data under the null distribution to obtain null data $\mathbf{x}^{0b}$ for $b = 1, \ldots, B$ iterations. The null statistics are calculated by $\widehat{g}^{0b}(\mathbf{x}^{0b})/\widehat{f}^{0b}(\mathbf{x}^{0b})$ where $\widehat{g}^{0b}$ and $\widehat{f}^{0b}$ are the new estimates based on $\mathbf{x}^{0b}$.

In our proposed procedure the null statistics are calculated by $\widehat{\mathcal{S}}_{\text{ODP}}(\mathbf{x}_i^{0b})$, where $\widehat{\mathcal{S}}_{\text{ODP}}$ is the estimated thresholding function based on the *original* data. In other words, we do not re-estimate the densites using the null data. When calculating the null distributions of many tests, the assumption is that some subset of m_0 null hypotheses are true and the remaining $m - m_0$ are false. Therefore, the correct null distribution would be calculated by (i) resampling the m_0 true nulls from their null distributions, (ii) resampling the remaining $m - m_0$ from their alternative distributions, (iii) re-estimating $\widehat{\mathcal{S}}_{\text{ODP}}$, and (iv) calculating the EFP based on the null statistics calculated among the m_0 true nulls.

Since we cannot identify the m_0 true nulls, we resample all data from their null distributions and we use the originally estimated thresholding function. We do not re-estimate $\widehat{\mathcal{S}}_{\text{ODP}}$ for each set of resampled data because these data are *all* null, and we want to be able to control the error rate under the case where m_0 are true nulls and $m - m_0$ are true alternatives. Re-estimating $\widehat{\mathcal{S}}_{\text{ODP}}$ for each set of full null data would result in a gross inflation of significance.

The danger in calculating the null statistics as we have done is that over-fitting could cause some artificial inflation of significance. If our procedure were carried out for a *single* test, then this inflation would be very noticeable. However, we were not able to detect any evidence of over-fitting for our proposed procedure in a variety of scenarios. For example, we randomly selected 1000 genes from the Hedenfalk et al. data set and randomly permuted their data (within genes) so that we could be certain that these 1000 were true nulls. We then performed our procedure, calculating p-values for every gene exactly as described in our algorithm. The p-values corresponding to the 1000 known null genes were then tested for equality to the Uniform distribution through a Kolmogorov-Smirnov test. According to the Kolmogorov-Smirnov test carried out over many iterations of this simulation, the p-values followed the Uniform distribution nearly perfectly[1].

There seem to be two reasons why our procedure does not suffer from over-fitting. The first is

[1]That is, for each iteration of this simulation, a Kolmogorov-Smirnov p-value was calculated, and then these were again tested against the Uniform distribution, indicating that there was no evidence among the many simulations that the ODP p-values deviated from a Uniform distribution.

that the ODP thresholding function is estimated from thousands of genes, so the variance of this estimate is negligible. In other words, one can randomly select a subset of, say, 1500 genes, estimate the ODP by these, and apply it to all of the data. The results will be virtually identical to using the entire data set. This is evidence that as the number of genes grows large, the estimated ODP eventually settles down to some fixed form. The second reason why we are able to avoid over-fitting is based on the approximate nuisance parameter invariance that was achieved. Because of this, the signals of true alternatives were not allowed to affect the overall sum of null densities.

Regardless, an extra precaution one can take is the following. When calculating resampling based null statistics for gene i, replace $\widehat{g}_i$ and $\widehat{f}_i$ with versions estimated from the resampled null data for gene i. The over-fitting of gene i is most likely to occur in $\widehat{g}_i$ and $\widehat{f}_i$, so these can be re-estimated while not disturbing the status of the other significance tests. If a gene's data are very different than all the other genes, then this adjustment is crucial because the other estimated densities contribute negligible amounts to its statistic, making this gene's statistic especially susceptible to over-fitting. If this extra precaution is taken then we do not foresee over-fitting to be an issue in typical data sets. One can also always test for over-fitting in the manner that we did with the Hedenfalk et al. study.

12 Simulation Details

The following displays the R code used to generate each data set from the four simulation scenarios considered in detail here. 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.

Scenario a:

```
dat <- matrix(rnorm(3000*16), ncol=16)
y <- c(rep(1,8),rep(2,8))
sigma2 <- 0.5 + rgamma(1500, shape=2, rate=4)
sigma2 <- c(sigma2, runif(1500, min=1.7, max=2.2))
sigma2 <- sample(sigma2)
mu <- rep(0,3000)
mu[1:200] <- rnorm(200, mean=1, sd=0.3)
mu[201:333] <- 1.2
mu[334:800] <- rnorm(467, mean=-1.0, sd=0.3)
mu[801:1000] <- -0.9
```

```
mu[1:1000] <- abs(mu[1:1000])*sample(c(rep(1,500),rep(-1,500)))
for(i in 1:3000) {
  dat[i,1:8] <- dat[i,1:8]*sqrt(sigma2[i])
  dat[i,9:16] <- dat[i,9:16]*sqrt(sigma2[i]) + rep(mu[i],8)
}
```

Scenario b:

```
dat <- matrix(rnorm(3000*16), ncol=16)
y <- c(rep(1,8),rep(2,8))
sigma2 <- runif(1000, min=0.5, max=0.75)
sigma2 <- c(sigma2, runif(500, min=1.2, max=1.3))
sigma2 <- c(sigma2, runif(1500, min=1.7, max=2.2))
sigma2 <- sample(sigma2)
mu <- rep(0,3000)
mu[1:200] <- rnorm(200, mean=1, sd=0.3)
mu[201:333] <- -1.2
mu[334:800] <- rnorm(467, mean=-1.0, sd=0.3)
mu[801:1000] <- -0.9
for(i in 1:3000) {
  dat[i,1:8] <- dat[i,1:8]*sqrt(sigma2[i])
  dat[i,9:16] <- dat[i,9:16]*sqrt(sigma2[i]) + rep(mu[i],8)
}
```

Scenario c:

```
dat <- matrix(rnorm(3000*24), ncol=24)
y <- c(rep(1,8),rep(2,8),rep(3,8))
sigma2 <- runif(3000, min=0.5, max=1.25)
sigma2 <- sample(sigma2)
mu <- rep(0,3000)
mu[1:200] <- rnorm(200, mean=1, sd=0.3)
mu[201:333] <- -1.2
mu[334:800] <- rnorm(467, mean=-1.0, sd=0.3)
mu[801:1000] <- -0.9
mu[1:1000] <- abs(sample(mu[1:1000]))
for(i in 1:600) {
  dat[i,y==1] <- dat[i,y==1]*sqrt(sigma2[i])
  dat[i,y==2] <- dat[i,y==2]*sqrt(sigma2[i]) + rep(mu[i],8)
  dat[i,y==3] <- dat[i,y==3]*sqrt(sigma2[i])
}
for(i in 601:1000) {
  dat[i,y==1] <- dat[i,y==1]*sqrt(sigma2[i])
  dat[i,y==2] <- dat[i,y==2]*sqrt(sigma2[i])
  dat[i,y==3] <- dat[i,y==3]*sqrt(sigma2[i]) + rep(mu[i],8)
```

```
}
for(i in 1001:3000) {
  dat[i,] <- dat[i,]*sqrt(sigma2[i])
}
```

Scenario d:

```
dat <- matrix(rnorm(3000*24), ncol=24)
y <- c(rep(1,8),rep(2,8),rep(3,8))
sigma2 <- runif(1000, min=0.5, max=0.75)
sigma2 <- c(sigma2, runif(500, min=1.2, max=1.3))
sigma2 <- c(sigma2, runif(1500, min=1.7, max=2.2))
sigma2 <- sample(sigma2)
mu <- rep(0,3000)
mu[1:200] <- rnorm(200, mean=1, sd=0.3)
mu[201:333] <- -1.2
mu[334:800] <- rnorm(467, mean=-1.0, sd=0.3)
mu[801:1000] <- -0.9
mu[1:1000] <- abs(sample(mu[1:1000]))
for(i in 1:700) {
  dat[i,y==1] <- dat[i,y==1]*sqrt(sigma2[i])
  dat[i,y==2] <- dat[i,y==2]*sqrt(sigma2[i]) + rep(mu[i],8)
  dat[i,y==3] <- dat[i,y==3]*sqrt(sigma2[i])
}
for(i in 701:1000) {
  dat[i,y==1] <- dat[i,y==1]*sqrt(sigma2[i])
  dat[i,y==2] <- dat[i,y==2]*sqrt(sigma2[i])
  dat[i,y==3] <- dat[i,y==3]*sqrt(sigma2[i]) + rep(mu[i],8)
}
for(i in 1001:3000) {
  dat[i,] <- dat[i,]*sqrt(sigma2[i]) }
```

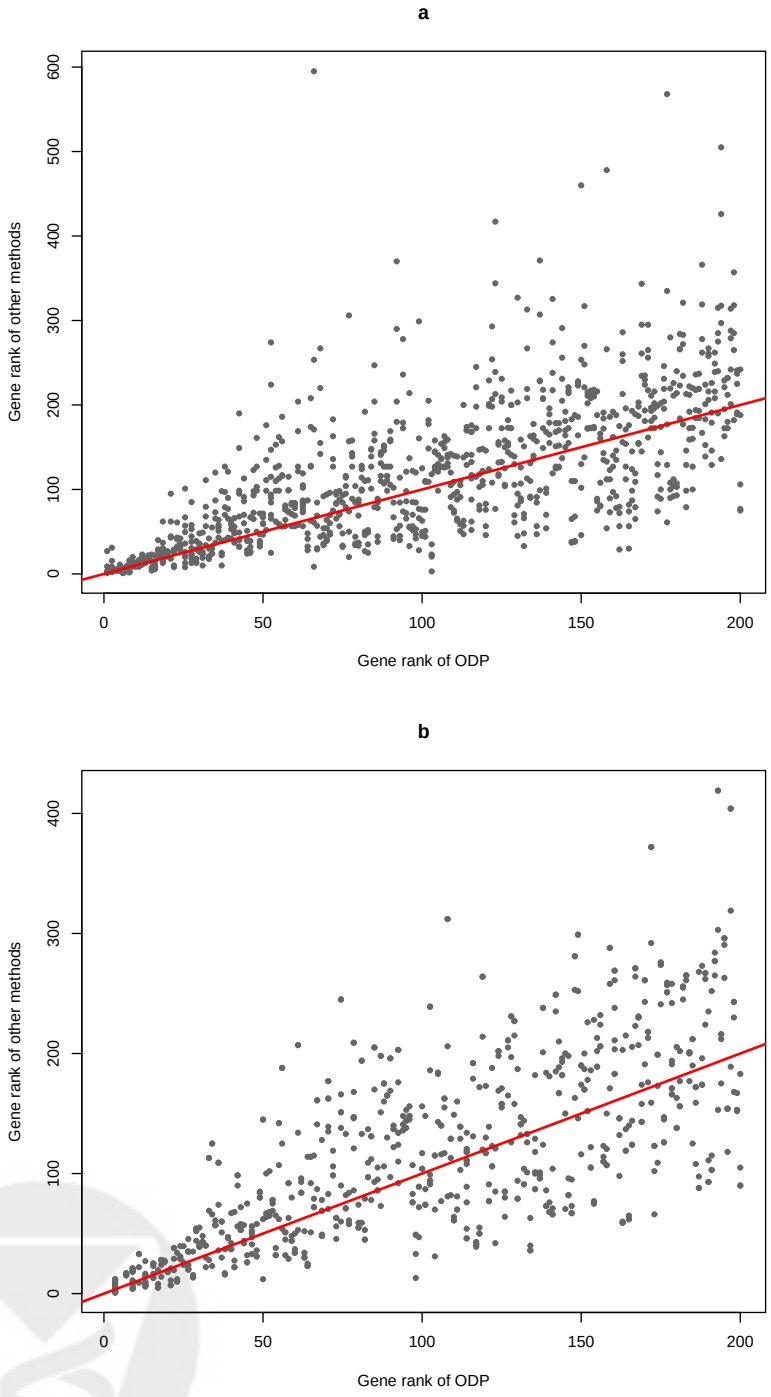

Figure 6: The gene ranking of the ODP versus the existing five methods when identifying differentially expressed genes from the Hedenfalk et al. data. For each of the top 200 ranked genes according to the ODP approach (x-axis), the ranking given by the other methods is plotted (y-axis). It can be seen that the ODP approach yields a notably different ranking of the genes. The identity line is shown in red, indicating whether the other methods produce a ranking higher or lower than the ODP approach. **(a)** Two-sample analysis identifying differentially expressed genes between *BRCA1* and *BRCA2* mutation positive tumors. **(b)** Three-sample analysis identifying differentially expressed genes between *BRCA1*, *BRCA2*, and Sporadic tumors.

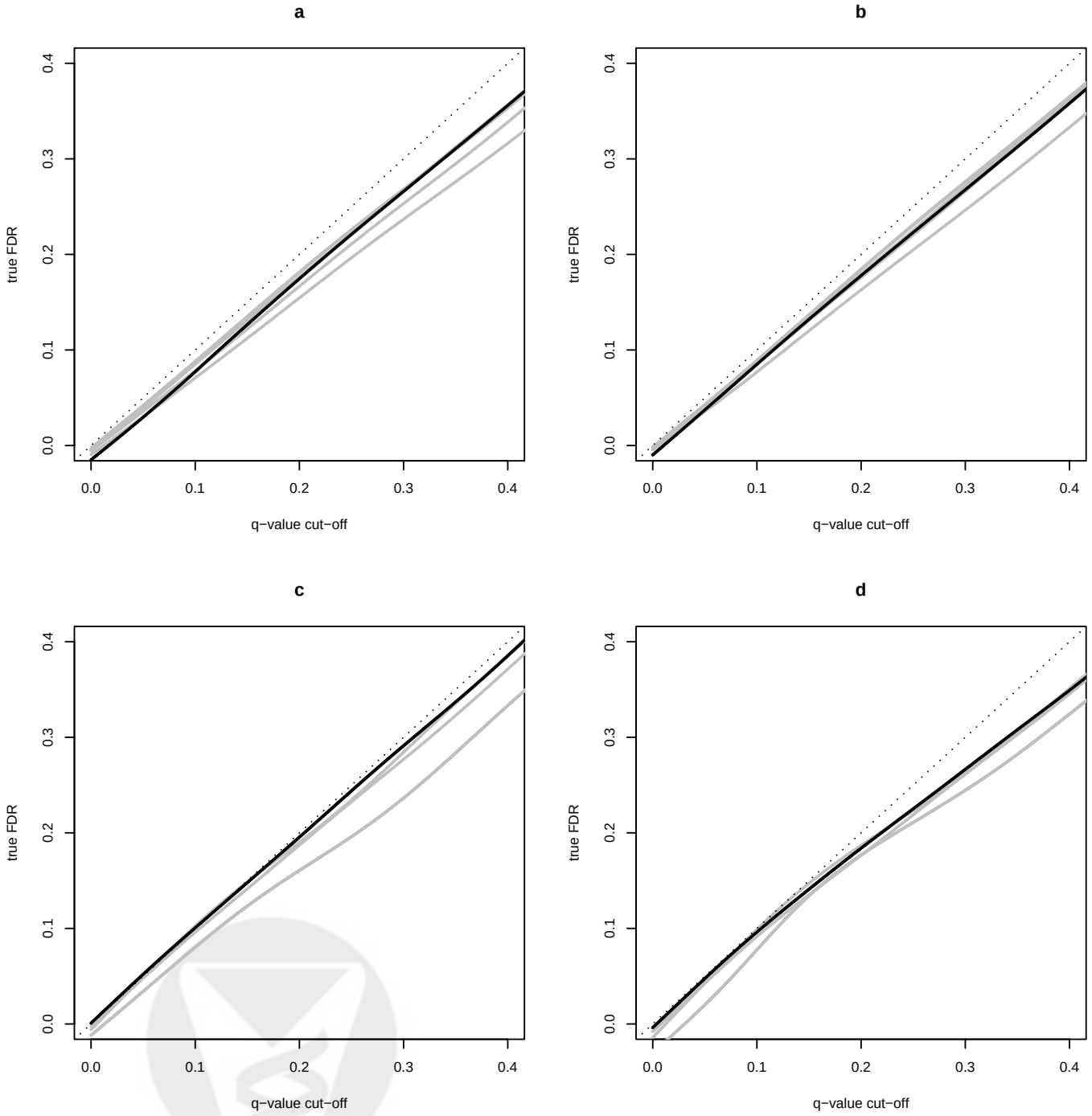

Figure 7: Plots verifying that each method considered controls the FDR by using the estimated q-value methodology of Storey (2002) and Storey & Tibshirani (2003). For each of the four simulation scenarios considered (a–d; see main text and Section 13 above), the estimated q-values versus the true FDR are plotted. The proposed ODP method is plotted in black, the other methods are plotted in grey, and the dotted line is the identity function. It can be seen that the estimated q-values conservatively estimate the FDR in all cases.

References

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

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

Rocke, D. M. & Durbin, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data, *Bioinformatics* **19**: 966–972.

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

Storey, J. D. (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., Taylor, J. E. & 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. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.