The Optimal Discovery Procedure II: Applications to Comparative Microarray Experiments

John D. Storey
University of Washington

James Y. Dai
University of Washington

Jeffrey T. Leek
University of Washington
The Optimal Discovery Procedure II: Applications to Comparative Microarray Experiments

Abstract

As much of the focus of genetics and molecular biology has shifted toward the systems level, it has become increasingly important to accurately extract biologically relevant signal from thousands of related measurements. The common property among these high-dimensional biological studies is that the measured features have a rich yet largely unknown underlying structure. An example of much current interest is the problem of identifying differentially expressed genes in comparative microarray experiments. Based on recently developed statistical theory, we propose a new approach for optimally extracting a subset of significant features from a high-dimensional data set, called the “optimal discovery procedure” (ODP). Whereas existing procedures essentially use data from only one feature at a time, the ODP uses the relevant information from the entire data set when testing each feature. We propose a generally applicable implementation of the ODP for identifying differentially expressed genes in microarray experiments. This microarray method consistently shows striking improvements in power over five highly-used existing methods. For example, in testing for differential expression between two breast cancer tumor types, the ODP provides increases from 88% to 294% in the number of genes called significant at a false discovery rate of 3%. Our proposed microarray method has been made available in the open-source, point-and-click EDGE software package.
The Optimal Discovery Procedure II: Applications to Comparative Microarray Experiments

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

Abstract: As much of the focus of genetics and molecular biology has shifted toward the systems level, it has become increasingly important to accurately extract biologically relevant signal from thousands of related measurements. The common property among these high-dimensional biological studies is that the measured features have a rich yet largely unknown underlying structure. An example of much current interest is the problem of identifying differentially expressed genes in comparative microarray experiments. Based on recently developed statistical theory, we propose a new approach for optimally extracting a subset of significant features from a high-dimensional data set, called the Optimal Discovery Procedure (ODP). Whereas existing procedures essentially use data from only one feature at a time, the ODP uses the relevant information from the entire data set when testing each feature. We propose a generally applicable implementation of the ODP for identifying differentially expressed genes in microarray experiments. This microarray method consistently shows striking improvements in power over five highly-used existing methods. For example, in testing for differential expression between two breast cancer tumor types, the ODP provides increases from 88% to 294% in the number of genes called significant at a false discovery rate of 3%. Our proposed microarray method has been made available in the open-source, point-and-click EDGE software package.

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

A microarray measures the abundance of mRNA transcripts of thousands of genes from a single biological sample (1, 2). Of much recent interest has been the problem of identifying genes that are differentially expressed across varying biological conditions based on microarray data (3). It is now possible to simultaneously measure thousands of related variables or “features” in a variety of biological studies. A rich yet largely unknown structure is usually expected to be present among these features, making their true biological signals related. 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 (3–5).

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), proven to be optimal in the companion paper (6). The ODP 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 new methodology for 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 would also yield some structure in observed differential expression. Therefore, it makes sense to capture and optimally utilize this structure when identifying differentially expressed genes. The ODP approach does exactly this, utilizing the relevant information from the entire data set in testing each gene for differential expression.

There are two steps required for performing large-scale significance testing in high-dimensional biological studies (see also the Discussion). The first step is to order the features from those showing the most signal of interest to those showing the least; the second step is to assign a significance level to each feature, allowing one to draw a significance cut-off somewhere along this ranking. For 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). Whereas the second step has been the major focus of new significance methods for high-dimensional studies (7), the ODP shows how to optimally perform the first step. This leads to the most efficient extraction of biological signal possible.

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. These statistics are formed using information from only one feature at a time. The ODP takes advantage of the structure in high-dimensional data that is ignored when using traditional test statistics, such as that mentioned above for microarrays. The ODP statistics use the relevant information from all features for assessing the relative significance of each individual one. For example, when testing whether a particular gene is differentially expressed in a microarray experiment, the statistic for that gene is formed using the expression data from all of the genes.

Even though the ODP may be simply stated in theory, estimating it in practice involves subtle detail that we illustrate in 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 many of these difficulties.

We demonstrate the proposed ODP approach for identifying differentially expressed genes on a well-known breast cancer expression study (8). We compare the results to those from five leading differential expression methods (9–14). Our method consistently shows substantial improvements in performance over these existing methods. For example, in testing for differential expression between \textit{BRCA1} and \textit{BRCA2} mutation-positive tumors, the ODP approach provides increases from 88\% to 294\% in the number of genes called significant at a 3\% FDR; at this level, it provides a 198\% increase over the highly used “SAM” software (9). A comparison between the methods over a range of FDRs is shown in Figure 1.

\section*{Methodology}

\textbf{The optimal discovery procedure.} The typical goal when identifying differentially expressed genes is to find as many true positives as possible, without incurring too many false positives (7). 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 (15).
Figure 1: A comparison of the ODP approach to five of the 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 FDR levels are 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 method has not been detailed for a 3-sample analysis, so it is omitted in this panel.
Because of these goals in microarray experiments, the FDR has emerged as a popular criterion to use when deciding how many genes to call differentially expressed, in addition to a variety of high-dimensional biological applications (7). The FDR is defined to be the proportion of false positives among all features called significant. 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 positive (EFP). Specifically, the ODP is defined as the rule 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 (6), the ODP also provides optimality properties for FDR, one consequence being that the 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 (6).

The ODP is very much related to one of the fundamental ideas behind individual significance tests, called the Neyman-Pearson lemma (16). Given a single set of observed data, the optimal single testing procedure is based on the likelihood ratio

\[
\frac{\text{likelihood of data under alternative distribution}}{\text{likelihood of data under null distribution}}.
\]

The null hypothesis is then rejected if the likelihood ratio exceeds some cut-off chosen to satisfy an acceptable Type I error rate. 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. It was shown in the companion paper that the optimality achieved by the ODP is a natural multiple test extension of this Neyman-Pearson optimality (6).

The ODP can be written similarly to the NP likelihood ratio procedure. However, instead of considering the data evaluated at its own alternative and null likelihood functions, the ODP considers the data for a single feature among all likelihood functions. The ODP statistic of each individual feature is calculated as

\[
\frac{\text{sum of data evaluated over all true alternative likelihoods}}{\text{sum of data evaluated over all true null likelihoods}}.
\]

For a fixed cut-off chosen to attain an acceptable EFP level (or FDR level), each null hypothesis
is rejected if its ODP statistic exceeds the cut-off. Note that “data” above is the data for a single feature, which has been evaluated at all true likelihoods, 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. Both intuitive and rigorous explanations of the operating characteristics of this statistic are given in the companion paper (6).

Estimating the ODP in general. The ODP requires information not generally known in practice. Here, we introduce a general approach to estimating the ODP, where the exact algorithmic details depend on the underlying assumptions. One set of assumptions turns out to be especially appropriate for identifying differentially expressed genes, allowing us to effectively implement an estimated ODP for this setting.

Suppose that \( m \) significance tests are performed on observed data sets \( x_1, x_2, \ldots, x_m \), where each significance test consists of \( n \) observations so that each \( x_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \). 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 \( i = m_0 + 1, \ldots, m \). The following statistic equivalently defines the ODP (6):

\[
S_{\text{ODP}}(x) = \frac{f_1(x) + f_2(x) + \cdots + f_{m_0}(x) + g_{m_0+1}(x) + g_{m_0+2}(x) + \cdots + g_m(x)}{f_1(x) + f_2(x) + \cdots + f_{m_0}(x)},
\]

which equals \( 1 + (a) \). Null hypothesis \( i \) is rejected if and only if \( S_{\text{ODP}}(x_i) \geq \lambda \), where \( \lambda \) is chosen to satisfy an acceptable EFP or FDR level. Note that 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.

Canonical ODP 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 \( \hat{f}_i \) be the maximum likelihood estimate\(^1\) of \( f_i \) based on data \( x_i \) under the constraints of the null hypothesis, and let \( \hat{g}_i \) be the unconstrained maximum likelihood estimate.

In single hypothesis testing, the Neyman-Pearson procedure for test \( i \) is based on \( g_i(x_i)/f_i(x_i) \), and it can be estimated by the generalized likelihood ratio statistic \( \hat{g}_i(x_i)/\hat{f}_i(x_i) \) (16). Our proposed approach builds on this strategy.

For true null hypotheses \( i = 1, \ldots, m_0 \), the maximum likelihood parameters defining \( \hat{f}_i \) and \( \hat{g}_i \) are both consistent estimates of the actual values of \( f_i \) as the number of observations \( n \) grows to infinity. Likewise, \( \hat{g}_i \) is composed of consistent parameter estimates of \( g_i \) for false null hypotheses \( i = m_0 + 1, \ldots, m \). Therefore, \( \hat{g}_1 + \cdots + \hat{g}_m \) can be used to estimate the numerator of equation (b),

\(^1\) Technically speaking, \( \hat{f}_i \) is the version of \( f_i \) defined by the unknown parameters’ maximum likelihood estimates under the constraints of the null hypothesis.
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:

\[
\hat{S}_{\text{ODP}}(x) = \frac{\hat{g}_1(x) + \cdots + \hat{g}_{m_0}(x) + \hat{g}_{m_0+1}(x) + \cdots + \hat{g}_m(x)}{\hat{f}_1(x) + \cdots + \hat{f}_{m_0}(x)}.
\]  

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 for each test \(n\) 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 properties.

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

\[
\hat{S}_{\text{ODP}}(x) = \frac{\sum_{i=1}^{m} \hat{g}_i(x)}{\hat{f}(x)}.
\]  

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

**Generally applicable estimates.** One general approach is to form weights that are consistent for the true status of each hypothesis. In other words, form data-dependent weights \(\hat{w}_i\) so that as \(n\) grows large, \(\hat{w}_i \to 0\) when the alternative hypothesis \(i\) is true and \(\hat{w}_i \to 1\) when null hypothesis \(i\) is true. If these are available, then a possible estimate of the ODP thresholding rule is

\[
\hat{S}_{\text{ODP}}(x) = \frac{\sum_{i=1}^{m} \hat{g}_i(x)}{\sum_{i=1}^{m} \hat{w}_i \hat{f}_i(x)}.
\]  

As in the canonical estimate, the unknown components of the true ODP are represented by estimates that are consistent. For example, suppose that each significance test concerns the equality of two Normal means. A threshold applied to the difference of observed means can be constructed to attain such weights; tests with a large observed mean difference obtain \(\hat{w}_i = 0\) and those with a small observed different are given \(\hat{w}_i = 1\). One simply needs to define “large” and “small” to ensure
reasonable operating characteristics of the $\hat{w}_i$.

Another more generally applicable approach relies on the concept of “nuisance parameter invariance.” 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 (Appendix).

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 (Appendix). In this case, $\sum_{i=1}^{m} \hat{f}_i/m$ serves as an estimate of $\sum_{i=1}^{m_0} f_i/m_0$, yielding the following estimate of the ODP thresholding rule:

$$\hat{\delta}_{\text{ODP}}(x) = \frac{\sum_{i=1}^{m} \hat{g}_i(x)}{\sum_{i=1}^{m} \hat{f}_i(x)},$$

where the unknown constant $m_0/m$ can be omitted. In practice, it is sometimes possible to formulate the significance tests or transform the data so that $\sum_{i=1}^{m} f_i/m \approx \sum_{i=1}^{m_0} f_i/m_0$, a strategy exploited below in the microarray application. This particular estimate can also be interpreted as an empirical Bayes procedure (Appendix).

**ODP for identifying differentially expressed genes.** For the microarray application, we found the implementation based on the estimate in equation (f) to perform the best. This implementation requires (i) $f_i$ and $g_i$ to be defined, (ii) estimates $\hat{f}_i$ and $\hat{g}_i$ to be derived, and (iii) 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 $\mu_{i1}$ be the mean of gene $i$ in group 1, and $\mu_{i2}$ 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 $\mu_{i0}$. 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 $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. Also, let $x_{i1}$ be the subset of data from group 1 and $x_{i2}$ the subset of data from group 2. For example, with seven arrays in group 1 and eight in group 2, we write $x_{i1} = (x_{i1}, x_{i2}, \ldots, x_{i7})$ and $x_{i2} = (x_{i8}, x_{i9}, \ldots, x_{i15})$.

**Probability density functions.** The model we use to estimate the ODP is that $x_{ij}$ comes from a Normal distribution with mean $\mu_{i1}$ or $\mu_{i2}$ (depending on the group that array $j$ belongs to) and variance $\sigma_i^2$. 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.

Under this assumption, the likelihood of a set of data can be written using the Normal probability density function $\phi$. For example, the likelihood of data $x$ with mean $\mu$ and variance $\sigma^2$ is written as

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

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

$$
f_i(x) = \phi(x; \mu_{i0}, \sigma_i^2) \quad \text{and} \quad g_i(x) = \phi(x_1; \mu_{i1}, \sigma_i^2) \phi(x_2; \mu_{i2}, \sigma_i^2).
$$

For hypothesis $i$, the likelihood of data $x$ is $f_i(x)$ under the null and $g_i(x)$ under the alternative.

**Estimates of the densities.** Ignoring nuisance parameter invariance issues, it is straightforward to define estimates of these densities. Let $(\hat{\mu}_{i0}, \hat{\sigma}_{i0}^2)$ be the maximum likelihood estimates under the constraints of the null hypothesis, and $(\hat{\mu}_{i1}, \hat{\sigma}_{i1}^2)$ be the unconstrained maximum likelihood estimates. These are simply the sample means and variances under the assumptions of the null and alternative hypotheses, respectively (Appendix). The above densities can then simply be estimated by $\hat{f}_i(\cdot) = \phi(\cdot; \hat{\mu}_{i0}, \hat{\sigma}_{i0}^2)$ and $\hat{g}_i(\cdot) = \phi(\cdot; \hat{\mu}_{i1}, \hat{\sigma}_{i1}^2) \phi(\cdot; \hat{\mu}_{i2}, \hat{\sigma}_{i2}^2)$. Since our implementation is based on equation (f), we need to modify these density estimates to approximately achieve nuisance parameter invariance.

**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 $(\mu_{i1}, \mu_{i2}, \sigma_i^2)$. Therefore, it is clear that the parameters $\mu_{i0}$ and $\sigma_i^2$ 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 $\sigma_i^2$ 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 (Appendix). Standard methods make it straightforward to transform the data so that there is no apparent relationship between the $\sigma_i^2$ and the $\mu_{i1} - \mu_{i2}$ (17), so this condition can usually be fulfilled in practice. Ideally, we would force $\mu_{i0} = 0$ by subtracting the true $\mu_{i0}$ from each $x_{ij}$ for $j = 1, \ldots, n$. However, $\mu_{i0}$ are unknown, so these must be estimated. Therefore, we set $\hat{\mu}_{i0} = \sum_{j=1}^{n} x_{ij}/n$ and define $x_{ij}^* = x_{ij} - \hat{\mu}_{i0}$, thereby centering each gene around zero.
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 \(\phi\). However, this turns out to be algebraically proportional to the original definition, so no modification is actually necessary. Furthermore, the slight dependence is maintained when estimating the null distribution (described below), so the null distribution may still be calculated correctly. Note that by centering each gene, we lose no information about differential expression. Furthermore, the centering also seemed to help avoid over-fitting issues (Appendix).

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 \(\hat{\mu}_{i1}^* = \hat{\mu}_{i1} - \hat{\mu}_{i0}\) and \(\hat{\mu}_{i2}^* = \hat{\mu}_{i2} - \hat{\mu}_{i0}\). The variances \(\sigma_i^2\) 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 \(\hat{\sigma}_{i0}^2\) and \(\hat{\sigma}_{i1}^2\), respectively.

**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 \(i = 1, 2, \ldots, m\):

\[
\hat{S}_{\text{ODP}}(x_i) = \frac{\sum_{g=1}^{m} \phi(x_{i1}^*; \hat{\mu}_{g1}^*, \hat{\sigma}_{g1}^2) \phi(x_{i2}^*; \hat{\mu}_{g2}^*, \hat{\sigma}_{g2}^2)}{\sum_{g=1}^{m} \phi(x_{i1}^*; 0, \hat{\sigma}_{g0}^2)}.
\]

Note that the centered data for gene \(i\), \(x_i^*\) 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 \(x_i^*\). 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:

\[
\hat{S}_{\text{ODP}}(x_i) = \frac{\sum_{g=1}^{m} \phi(x_{i1}^*; \hat{\mu}_{g1}^*, \hat{\sigma}_{g1}^2) \cdots \phi(x_{iK}^*; \hat{\mu}_{gK}^*, \hat{\sigma}_{gK}^2)}{\sum_{g=1}^{m} \phi(x_{i1}^*; 0, \hat{\sigma}_{g0}^2)}.
\]

Analogously to the two-sample method, each gene is mean centered around zero to obtain the transformed data \(x_i^*\). In the 1-sample case, the data do not have to be mean centered because there is no nuisance location parameter present.
Full algorithm for identifying differentially expressed genes. The following is a description of the full estimated ODP for identifying differentially expressed genes. Once the statistics have been formed, the remainder of the algorithm follows closely to our previous work on estimating each gene’s significance level through $q$-values and the FDR (7,18). Full details of this algorithm, including exact formulas for general $K$-sample comparisons can be found in the Appendix. Note that one can also determine a useful significance threshold through estimates of the EFP and ETP, which we also outline in the Appendix.

1. Using the formula given in equation (g), evaluate the ODP statistic $\hat{S}_{\text{ODP}}(x_i)$ for each gene $i = 1, 2, \ldots, m$.

2. For $B$ iterations, randomly permute the expression data and re-compute each statistic to get a set of null statistics $\hat{S}_{\text{ODP}}(x_i^{(b)})$ for $b = 1, \ldots, B$ and $i = 1, \ldots, m$. (Note: The permutations are carried out so that for each iteration, the same permutation is 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 (7,18).

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 (7,19). 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. Below, we apply this method to a well known breast cancer study, and we compare the ODP approach to several highly used existing approaches.

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 $\hat{g}_i(x_i)/\hat{f}_i(x_i)$. When the estimated parameters defining $\hat{f}_i$ and $\hat{g}_i$ are the maximum likelihood estimates, then $\hat{g}_i(x_i)/\hat{f}_i(x_i)$ is equivalent to employing the usual t-statistic (16). When the maximum likelihood estimates are shrunken towards a common value (across genes), then the so-called SAM statistic and other similar versions emerge (9,12,13). Therefore, these more intricate statistics use information across genes only in that different estimates are employed in $\hat{g}_i(x_i)/\hat{f}_i(x_i)$. Not surprisingly, these modified statistics sometimes perform worse than the traditional t-statistic and F-statistic (Results).
Results

We applied our proposed approach to a well-known study comparing the expression of breast cancer tumor tissues among individuals who are BRCA1-mutation-positive, BRCA2-mutation-positive, and “Sporadic” (8). The expression measurements consist of 3169 genes from 21 arrays; seven from BRCA1, eight from BRCA2, and six from Sporadic (Appendix). 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 refs. 9 and 18, (ii) the traditional t-tests and F-tests as previously suggested for microarray analysis (10, 11), (iii) a recently proposed variation on these that uses “shrunken” versions of the statistics (12), (iv) a non-parametric Bayesian method whose estimated posterior probabilities are also sometimes interpreted as estimated Bayesian “local FDR” estimates (13), and (v) a model-based empirical Bayesian method giving posterior probabilities of differential expression (14). Several other Bayesian methods exist (20–22), but these require exact specification of prior distributions, making it difficult to perform an objective comparison with these.

The methods were compared to determine how accurately and efficiently each one extracts the relevant biological signal. We took special care to make sure that the comparisons were performed fairly. 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 (7, 18). In order to estimate the $q$-values, simulated null statistics were calculated for each method. This was accomplished by performing the same set of permutations and calculating null statistics for each method. All computer code used in the comparisons is available from the authors upon request.

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 (Appendix). 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.

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 122 significant genes, whereas existing methods only find 31–65 significant genes. The estimated ODP method therefore offers increases from 88% to 294% in the number of genes called significant. In particular, it provides a 198% increase over
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 1%, 2%, ..., 10%. The exact same FDR methodology (7,18) was applied to each thresholding method in order to make the comparisons fair. The model-based Bayesian method (14) is not defined for a 3-sample analysis, so that case is omitted.

<table>
<thead>
<tr>
<th>Thresholding Method</th>
<th>% Increase by ODP – 2-sample</th>
<th>% Increase by ODP – 3-sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
<td>Median</td>
</tr>
<tr>
<td>SAM (9)</td>
<td>27</td>
<td>50</td>
</tr>
<tr>
<td>t/F-test (10,11)</td>
<td>43</td>
<td>82</td>
</tr>
<tr>
<td>Shrunken t/F-test (12)</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>Bayesian “local FDR” (13)</td>
<td>60</td>
<td>112</td>
</tr>
<tr>
<td>Posterior probability (14)</td>
<td>40</td>
<td>70</td>
</tr>
</tbody>
</table>

* Four of these methods called two or less genes significant at FDR=1%, while the ODP called 21 significant, so the maximum increase could be reported as much larger. These instances were omitted to avoid reporting possibly unstable figures.

the highly used SAM software. The median increase in the number of genes called significant at q-value cut-offs less than or equal to 10% ranges from 45–112% across all five methods. In testing for 3-sample differential expression among the \textit{BRCA1}, \textit{BRCA2}, and Sporadic groups, the ODP approach offers even greater improvements. For example, it provides increases from 111% to 200% in the number of genes called significant at a false discovery rate of 3%; at this level, it provides an 191% increase over SAM.

Figure 1 plots the number of genes called significant among the different methods across a range of q-value cut-offs for both analyses. It can be seen that while the five existing methods perform similarly, our proposed method is substantially better than all of them. This remains true across the entire range of q-value cut-offs, even those not shown in Figure 1. Table 1 shows the minimum, median, and maximum percentage increase in number of genes called significant by the ODP approach relative to each method at FDR cut-offs of 1%, 2%, ..., 10%.

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. In the 2-sample setting the only difference between, say, SAM and a t-test is that SAM uses a slightly modified t-statistic and applies an asymmetric thresholding rule based on a quantile-quantile plot (23). This is also true for the nonparametric empirical Bayesian method of ref. 13. In the 3-sample setting, these methods are no longer any different than a traditional F-test, except for the slightly modified denominator of the statistic, which actually performs worse than a traditional F-test over a majority of FDR levels (Figure 1b). 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 (24), then the gains may be even more substantial because here the asymmetry is even harder to quantify using traditional statistics.

**Biological significance.** In order to determine whether the ODP leads to additional biological insight, we considered our findings relative to those of the five existing methods in the context of identifying genes differentially expressed between the BRCA1 and BRCA2 groups. It is well known that breast tumors associated with BRCA1 mutations and BRCA2 mutations differ greatly from each other in their histological appearance (25). 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 (25). Concordant with these morphological differences, the gene expression profiles of these two types of tumors have also shown to be distinctive (8).

At a $q$-value cut-off of 5%, we found 215 genes that are 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 A2 (CCNA2), 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 (8). 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, BIRC2, Myc oncogene, and others. Many of these genes have been described in an earlier analysis of this study (8).

At a $q$-value cut-off of 5%, the five existing methods found between 99–143 genes to be significant. Almost every gene identified by these other methods is among the 215 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), cell division cycle 25B (CDC25B), connective tissue growth factor (CTGF), growth factor receptor-bound protein 2, CCAAT/enhancer binding protein beta (CEBPB), among others.

**Additional numerical results.** Similar comparisons were made on simulated data, where one knows with certainty which genes are differentially expressed. These simulated data sets had the common characteristics that there was (i) enough structure present in the data (e.g., asymmetric differential
expression) that it could be seen through straightforward plots, and (ii) a substantial percentage
(≥ 5%) of genes differentially expressed. Among these scenarios, the ODP approach was not
outperformed by any of the existing methods, and the relative gains were similar to those made on
the breast cancer data set. Since the relative performance of each method did not change across a
variety of simulations and real data sets, we do not show those results here.

It is certainly possible to find some simulation scenario where the estimated ODP is outper-
formed, 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. In simulation scenarios where
a small percentage of genes were differentially expressed, the magnitude of differential expression
was weak, and the distribution of differential expression was perfectly symmetric about zero, it was
sometimes the case that the proposed ODP estimate was outperformed. In these cases, a standard
t-test captured the signal the best among all existing methods. However, it should be noted that
in these scenarios no method was able to produce any q-values less than, say, 30–40%, so it is not
clear that such examples are relevant.

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, 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 necessarily think 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.

Appendix

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.

We denote $x_{ij}$ to be the expression observation for gene $i$ in array $j$, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The data for a single gene is written as $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. Assume there are $K$ groups tested for differential expression, and let $x_{ik}$ be the subset of data from group $k$, $k = 1, \ldots, K$. Finally, let $G_k$ be the set of arrays corresponding to group $k$ so that $x_{ik} = (x_{ij})_{j \in G_k}$. Gene $i$ in group $k$ has mean gene expression $\mu_{ik}$, and variance $\sigma_{i}^2$. Without loss of generality, we define the mean when the null hypothesis of no differential expression is true to be $\mu_{i0} = \frac{\sum_{k=1}^{K} n_k \mu_{ik}}{n}$, where $n_k$ is the number of arrays in group $k$.

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

\[
\hat{\mu}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}; \quad \hat{\sigma}^2_i = \frac{\sum_{j=1}^{n} (x_{ij} - \hat{\mu}_i)^2}{n-1}
\]

\[
\hat{\mu}_{ik} = \frac{\sum_{j \in \bar{G}_k} x_{ij}}{n_k}; \quad \hat{\sigma}^2_{ik} = \frac{\sum_{k=1}^{K} \sum_{j \in \bar{G}_k} (x_{ij} - \hat{\mu}_{ik})^2}{n-K}
\]

Note that these are the typical estimates for Normally distributed data.

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

\[
\hat{S}_{ODP}(x_i) = \frac{\sum_{g=1}^{m} \phi(x_i; \hat{\mu}_{g1}, \hat{\sigma}^2_{g1})}{\sum_{g=1}^{m} \phi(x_i; 0, \hat{\sigma}^2_{g0})}.
\]

When \(K > 1\) each gene is centered as a step in approximately achieving “nuisance parameter invariance” as described in the main text. Define \(x^*_{ij} = x_{ij} - \hat{\mu}_i\) and \(\hat{\mu}^*_{ik} = \hat{\mu}_{ik} - \hat{\mu}_i\). The estimated ODP statistic for each gene \(i\) is then

\[
\hat{S}_{ODP}(x_i) = \frac{\sum_{g=1}^{m} \phi(x^*_{i1}; \hat{\mu}^*_{g1}, \hat{\sigma}^2_{g1}) \cdots \phi(x^*_{iK}; \hat{\mu}^*_{gK}, \hat{\sigma}^2_{g1})}{\sum_{g=1}^{m} \phi(x^*_{i1}; 0, \hat{\sigma}^2_{g0})}.
\]

**Step 2: Simulating null statistics.** For \(B\) iterations, randomly permute the expression data and re-compute each statistic to get a set of null statistics \(\hat{S}_{ODP}(x^{0b}_i)\) for \(b = 1, \ldots, B\) and \(i = 1, \ldots, m\). The permutations are carried out so that for each iteration the same permutation is applied to all genes, which keeps the dependence structure of the genes intact. In other words, let \(r_{1b}, r_{2b}, \ldots, r_{nb}\) be the scrambled version of \(1, 2, \ldots, n\) for iteration \(b\). Then we define \(x^{0b}_i = (x_{ir_{1b}}, x_{ir_{2b}}, \ldots, x_{ir_{nb}})\) for each gene \(i = 1, 2, \ldots, m\). When an experimental design is in place that requires balanced permutations, then one simply carries out these restricted permutations instead.

We also experimented with using the bootstrap to simulate null statistics. This would be performed similarly to above, except the residuals obtained from the unconstrained mean parameter estimates are sampled with replacement and added back to the null constrained estimated mean parameter. This approach usually produced very similar results to the permutation approach, although we noticed that it tended to be more sensitive to pivotality issues than the permutation null.

**Step 3: Estimating q-values.** One can think of forming a significance cut-off by calling all genes significant with \(\hat{S}_{ODP}(x_i) \geq c\) for some cut-point \(c\). Given that the ODP is formulated in terms of
thresholding these statistics rather than \( p \)-values, it is conceptually easier to think of estimating \( q \)-values based on thresholding statistics directly. However, we show now how one can calculate \( q \)-values from \( p \)-values defined so that direct thresholding of the statistics implicitly takes place, while at the same time we can avoid some of the more complicated formulas involved with thresholding statistics directly.

The \( p \)-value for gene \( i \) can be calculated by

\[
p_i = \sum_{b=1}^{B} \frac{\# \{ j : \hat{S}_{\text{ODP}}(x_j) \geq \hat{S}_{\text{ODP}}(x_i) , j = 1, \ldots, m \}}{m \cdot B}.
\]

\( \text{(h)} \)

When \( p \)-values are calculated in this way, the subsequent \( q \)-value estimates are equivalent to those from directly thresholding the statistics.

In the main text, we motivated the use of \( q \)-values as a measure of significance rather than \( p \)-values. Background on estimating \( q \)-values and their use in genomics can be found in ref. 18 and ref. 7. Note that \( \pi_0 \), which is the proportion of genes that are not differentially expressed, is estimated as a part of \( q \)-value estimation. This quantity is useful in itself since \( 1 - \pi_0 \) gives the proportion of differentially expressed genes, even though all of these cannot usually be identified with certainty. Following is the general algorithm for estimating \( q \)-values from their corresponding \( p \)-values (7).

1. Let \( p_1 \leq p_2 \leq \ldots \leq p_m \) be the ordered \( p \)-values. This also denotes the ordering of the genes in terms of their evidence for differential expression.

2. For a range of \( \lambda \), say \( \mathcal{R} = \{0, 0.01, 0.02, \ldots, 0.95\} \), calculate

\[
\hat{\pi}_0(\lambda) = \frac{\# \{ p_j > \lambda ; j = 1, \ldots, m \}}{m(1 - \lambda)}.
\]

Let \( \hat{f} \) be the natural cubic spline of \( \hat{\pi}_0(\lambda) \) on \( \lambda \). (We use a smoother with 3 degrees of freedom.) Set the estimate of \( \pi_0 \) to be

\[
\hat{\pi}_0 = \hat{f}(\text{max} \mathcal{R}).
\]

3. Calculate

\[
\hat{q}(p_m) = \min_{t \geq p_m} \frac{\hat{\pi}_0 m \cdot t}{\# \{ p_j \leq t ; j = 1, \ldots, m \}} = \hat{\pi}_0 \cdot p_m.
\]

18
4. For \( i = m - 1, m - 2, \ldots, 1 \), calculate
\[
\hat{q}(p(i)) = \min_{t \geq p(i)} \frac{\hat{\pi}_0 m \cdot t}{\# \{ p_j \leq t; j = 1, \ldots, m \}} = \min \left( \frac{\hat{\pi}_0 m \cdot p(i)}{i}, \hat{q}(p(i+1)) \right).
\]

5. The estimated \( q \)-value for the \( i^{th} \) most significant gene is \( \hat{q}(p(i)) \).

The following estimate of the false discovery rate when calling all \( p \)-values \( \leq t \) significant is implicit in the algorithm:
\[
\hat{\text{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{\# \{ p_i \leq t; i = 1, \ldots, m \}}.
\]

Some articles have distinguished false discovery rate estimation based on thresholding statistics directly from that based on \( p \)-values. In particular, Tusher et al. (9) do not point out that their false discovery rate method uses the estimate implicit in the Benjamini and Hochberg (26) procedure. Also, Dudoit et al. (27) incorrectly criticize the approach of Tusher et al. (9), also not recognizing that it employs the Benjamini and Hochberg (26) estimate by way of gene non-specific \( p \)-values. It can easily be shown that false discovery rate estimation based on the original statistics is equivalent to that based on \( p \)-values as long as the \( p \)-values are calculated as we did above.

For a fixed significance cut-off \( c \) applied to the original statistics, the false discovery rate estimate is
\[
\hat{\text{FDR}}(c) = \frac{\hat{\pi}_0 \sum_{b=1}^{B} \# \{ \hat{S}_{\text{ODP}}(x_i^{(b)}) \geq c; i = 1, \ldots, m \}}{\# \{ \hat{S}_{\text{ODP}}(x_i) \geq c; i = 1, \ldots, m \}}/B,
\]
where \( \hat{\pi}_0 \) is derived from the
\[
\hat{\pi}_0(c') = \frac{\# \{ \hat{S}_{\text{ODP}}(x_i) < c'; i = 1, \ldots, m \}}{\sum_{b=1}^{B} \# \{ \hat{S}_{\text{ODP}}(x_i^{(b)}) < c'; i = 1, \ldots, m \}}/B
\]
over some range of \( c' \) exactly as in the above algorithm. As was stated in Storey and Tibshirani (7), the original FDR estimate of Storey (18) is easily shown to be equivalent to the above formula, where the only difference is that it is written in terms of thresholding the statistics directly. The key observation is that one can equivalently define the Type I error rate of a given cut-off by \( \sum_{b=1}^{B} \# \{ \hat{S}_{\text{ODP}}(x_i^{(b)}) \geq c \}/(m \cdot B) \) rather than the \( p \)-value threshold \( t \). In fact, if we define
\[
c(t) \equiv \min \{ \hat{S}_{\text{ODP}}(x_i): p_i \leq t \}
\]
then it can be shown that
\[
\frac{\hat{\pi}_0 \sum_{b=1}^{B} \# \{ \hat{S}_{\text{ODP}}(x_i^{(b)}) \geq c(t); i = 1, \ldots, m \}}{\# \{ \hat{S}_{\text{ODP}}(x_i) \geq c(t); i = 1, \ldots, m \}} = \frac{\hat{\pi}_0 m \cdot t}{\# \{ p_i \leq t; i = 1, \ldots, m \}}
\]
over some range of \( c' \) exactly as in the above algorithm.
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.

A more conservative approach to calculating \( p \)-values is to only consider the null statistics generated from that gene:

\[
p_i = \frac{\#\{b : \hat{S}_{\text{ODP}}(x_{0b}) \geq \hat{S}_{\text{ODP}}(x_i), b = 1, \ldots, B\}}{B}.
\]

However, we use the former definition given in equation (h), which is still justified by weaker conditions than assuming some parametric null distribution. Results indicating that each gene does not even have to follow the same null distribution in order for these \( p \)-values to be valid for use in \( q \)-value estimation can be found in ref. 28. Moreover, one can always test whether the conditions of ref. 28 are met or not. Basically one has to verify that the distribution of the pooled \( p \)-values does not depend on the magnitudes of the observed statistics, which can be straightforwardly accomplished by a Kolmogorov-Smirnov test. Within-gene \( p \)-values also causes one to potentially apply a different effective threshold to each ODP statistic, which is not what the ODP theory dictates.

**Estimating EFP and ETP.** One may also be interested in estimating the EFP and ETP for each ODP threshold. These estimates follow easily from the \( q \)-value estimates. In particular, for a threshold \( c \) define

\[
\hat{\text{EFP}}(c) = \hat{\pi}_0 \sum_{b=1}^{B} \frac{\#\{\hat{S}_{\text{ODP}}(x_{0b}) \geq c; i = 1, \ldots, m\}}{B}
\]

\[
\hat{\text{ETP}}(c) = \#\{\hat{S}_{\text{ODP}}(x_i) \geq c; i = 1, \ldots, m\} - \hat{\text{EFP}}(c)
\]

where \( \hat{\pi}_0 \) is estimated as above. It is straightforward to show that under the conditions making the \( q \)-values estimates valid, it also follows that these estimates are conservative in that \( E[\hat{\text{EFP}}(c)] \geq E\text{FP}(c) \) and \( E[\hat{\text{ETP}}(c)] \leq E\text{TP}(c) \). Note that these estimates are already a part of the “\( q \)-plots” defined in ref. 7.

The FDR estimate from equation (i) can be written in terms of these estimates, further showing the direct connection between the EFP, ETP, and FDR:

\[
\hat{\text{FDR}}(c) = \frac{\hat{\text{EFP}}(c)}{\text{EFP}(c) + \text{ETP}(c)}.
\]

**Comparing procedures based on the number of genes called significant.** The ODP
approach was compared to five leading procedures 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 $\hat{\text{ETP}} = (# \text{ significant genes})(1 - \hat{\text{FDR}})$, as just shown above. Since each method is compared at the same value of $\hat{\text{FDR}}$, it follows that comparing the number of significant genes is equivalent to comparing the methods based on $\hat{\text{ETP}}$, 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 $\hat{\text{ETP}}$ for each fixed $\hat{\text{EFP}}$ level.

**Nuisance parameter invariance.** The ODP is most simply defined in terms of the following rule (6):

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

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

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

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 $\phi(x; \mu_{i0}, \sigma_i^2)$ are due to differing $\mu_{i0}$ and $\sigma_i^2$, which are not used at all in defining the null and alternative hypotheses. The parameters $\mu_{i0}$ and $\sigma_i^2$ 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 (j) 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 $\mu_{10} = \cdots = \mu_{m_0,0} = 1000$ and $\mu_{m_0+1,0} = \cdots = \mu_{m0} = -1000$. In the former scenario, the ODP statistic is $\frac{1}{m_0+1}$, while in the latter it is $\frac{1}{m_0}$. This shows that the presence of nuisance parameters can significantly affect the ODP statistic.
\(\mu_{m0} = 0\). 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 \(\sigma_i^2\). 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, \((x_{ij} - \bar{\mu}_{i0})/\bar{\sigma}_i\). 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 \(\sigma_i^2\), and dividing the data by \(\bar{\sigma}_i\) 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: \(\sum_{i=1}^{m} f_i/m = \sum_{i=1}^{m0} f_i/m0\). 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 \(\sigma_i^2\).

Let \(x^*\) be the mean centered data for a single gene (thereby removing the effect of \(\mu_{i0}\)), and let \(\mu_{i1}^* = \mu_{i1} - \mu_{i0}\), \(\mu_{i2}^* = \mu_{i2} - \mu_{i0}\), \(\mu_{i0}^* = 0\). In the case that \(\sum_{i=1}^{m} \phi(\cdot;0,\sigma_i^2)/m = \sum_{i=1}^{m0} \phi(\cdot;0,\sigma_i^2)/m0\), the following statistics are all equivalent:

\[
\begin{align*}
\sum_{i=m0+1}^{m} \phi(x^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2) / \sum_{i=1}^{m} \phi(x^*; 0, \sigma_i^2) & \quad \sum_{i=m0+1}^{m} \phi(x^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2) / \sum_{i=1}^{m} \phi(x^*; 0, \sigma_i^2) \\
\sum_{i=m0+1}^{m} \phi(x^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2) / \sum_{i=1}^{m} \phi(x^*; 0, \sigma_i^2) & \quad \sum_{i=m0+1}^{m} \phi(x^*; 0, \sigma_i^2) + \sum_{i=m0+1}^{m} \phi(x^*; \mu_{i1}^*, \mu_{i2}^*, \sigma_i^2) / \sum_{i=1}^{m} \phi(x^*; 0, \sigma_i^2)
\end{align*}
\]

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 \(\sum_{i=1}^{m} \phi(\cdot;0,\sigma_i^2)/m = \sum_{i=1}^{m0} \phi(\cdot;0,\sigma_i^2)/m0\), 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 (17).

**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, \( \frac{g(x)}{f(x)} \), 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 \( x^{0b} \) for \( b = 1, \ldots, B \) iterations. The null statistics are calculated by \( \frac{\hat{g}^{0b}(x^{0b})}{\hat{f}^{0b}(x^{0b})} \) where \( \hat{g}^{0b} \) and \( \hat{f}^{0b} \) are the new estimates based on \( x^{0b} \).

In our proposed procedure the null statistics are calculated by \( \hat{S}_{ODP}(x^{0b}) \), where \( \hat{S}_{ODP} \) is the estimated thresholding function based on the original data. In other words, we do not re-estimate the densities 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 \( \hat{S}_{ODP} \), 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 \( \hat{S}_{ODP} \) 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 \( \hat{S}_{ODP} \) 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\(^2\).

\(^2\)That is, for each iteration of this simulation, a Kolmogorov-Smirnov \( p \)-value was calculated, and then these were
There seem to be two reasons why our procedure does not suffer from over-fitting. The first is
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 \( \hat{g}_i \) and \( \hat{f}_i \) with versions estimated from the resampled null data
for gene \( i \). The over-fitting of gene \( i \) is most likely to occur in \( \hat{g}_i \) and \( \hat{f}_i \), 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.

**Empirical Bayes interpretation.** The following is an empirical Bayes interpretation of our
method. Suppose there is a uniform prior on the unknown null densities \( f_1, f_2, \ldots, f_m \), as well
as on the alternative densities \( g_1, g_2, \ldots, g_m \). Further, each null hypothesis is true with some prior
probability \( \pi_0 \). It easily follows based on Lemma 3 of ref. 6 that the ODP in this scenario is defined
by the statistic

\[
\frac{g_1(x) + g_2(x) + \cdots + g_m(x)}{f_1(x) + f_2(x) + \cdots + f_m(x)},
\]

where any such test statistic exceeding some threshold \( \lambda \) is called significant.

The above rule is equivalent to thresholding tests for significance according to the quantity

\[
\frac{(1 - \pi_0) \sum_{i=1}^m g_i(x)}{\pi_0 \sum_{i=1}^m f_i(x) + (1 - \pi_0) \sum_{i=1}^m g_i(x)}
\]

which is exactly equal to the posterior probability a test with data \( x \) is a true alternative. It follows
by Theorems 1 and 6 of ref. 19 that this rule is also optimal in terms of both the FDR and the
misclassification rate.

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.
Our proposed microarray procedure is based on the statistic
\[
\frac{\hat{g}_1(x) + \hat{g}_2(x) + \cdots + \hat{g}_m(x)}{\hat{f}_1(x) + \hat{f}_2(x) + \cdots + \hat{f}_m(x)}.
\]
This statistic is simply equation \((k)\) with each \(f_i\) and the \(g_i\) replaced with an estimate based on the data, making it interpretable as an empirical Bayes estimate. Moreover, the estimate of \(\pi_0\) formed when estimating the \(q\)-values (see above) can be included to form conservative empirical Bayes estimates of the probability that a gene with data \(x\) is differentially expressed:
\[
\hat{\Pr}(\text{diff. expressed } | x) = \frac{(1 - \hat{\pi}_0) \sum_{i=1}^{m} \hat{g}_i(x)}{\hat{\pi}_0 \sum_{i=1}^{m} \hat{f}_i(x) + (1 - \hat{\pi}_0) \sum_{i=1}^{m} \hat{g}_i(x)}.
\]
Note that this probability estimate yields the same thresholding rule as the above statistic.

**Hedenfalk et al. data.** We assessed the performance of the ODP and five existing methods 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” \((8)\). 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 \((7)\), 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/](http://research.nhgri.nih.gov/microarray/NEJM_Supplement/) and all data were analyzed on the log2 scale.

**Acknowledgments**

This research was supported in part by NIH grant R01 HG002913-01. Thanks to Alan Dabney for a number of helpful comments on this manuscript, and thanks to members of the Johns Hopkins Biostatistics Department for much helpful feedback during a seminar by JDS on this work.

**References**


