

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

7	A Simple Motivating Example	2
8	Detailed algorithm for identifying differentially expressed genes	2
9	Comparing procedures based on the number of genes called significant	7
10	Nuisance parameter invariance	8
11	Over-fitting	10
12	Simulation Details	11

*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 $i = 1, \dots, m$ and $j = 1, \dots, n$. The data for a single gene is written as $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$. Assume there are K groups tested for differential expression, and let \mathbf{x}_{ik} be the subset of data from group k , $k = 1, \dots, K$. 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

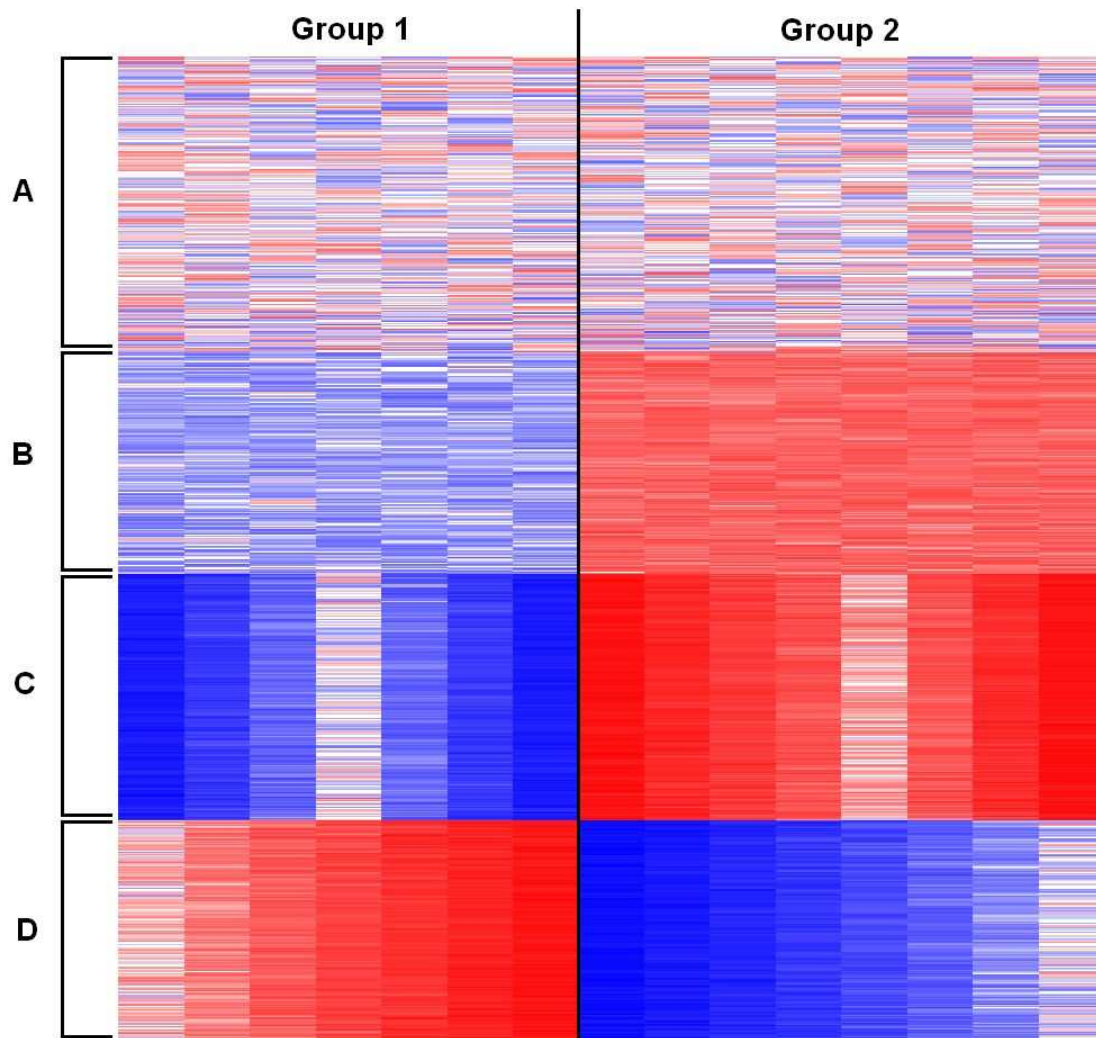


Figure 5: Simulated expression data showing structure in the biological signal of interest. A heat map of expression data on 1000 genes among 15 individuals is shown (red = high, blue = low). The first seven individuals come from Group 1 and the last eight from Group 2. The genes were hierarchically 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 $(\hat{\mu}_{i0}, \hat{\sigma}_{i0}^2)$ be estimates under the constraints of the null hypothesis, and $(\hat{\mu}_{i1}, \dots, \hat{\mu}_{iK}, \hat{\sigma}_{i1}^2)$ be unconstrained estimates. These are defined as follows:

$$\begin{aligned}\hat{\mu}_{i0} &= \sum_{j=1}^n x_{ij}/n; & \hat{\sigma}_{i0}^2 &= \sum_{j=1}^n \frac{(x_{ij} - \hat{\mu}_{i0})^2}{n-1} \\ \hat{\mu}_{ik} &= \sum_{j \in \mathcal{G}_k} x_{ij}/n_k; & \hat{\sigma}_{iA}^2 &= \sum_{k=1}^K \sum_{j \in \mathcal{G}_k} \frac{(x_{ij} - \hat{\mu}_{ik})^2}{n-K}\end{aligned}$$

Note that these are the typical estimates for Normally distributed data. The estimated weights \hat{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 ($i = 1, \dots, m$) is

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

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

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

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 ϵ_i be the alternative model residuals obtained from each gene i ’s expression data \mathbf{x}_i by setting $\epsilon_{ij} = x_{ij} - \hat{\mu}_{ik(j)}$ where $k(j)$ is the group to which array j belongs for $i = 1, \dots, m$. 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 $\hat{\mu}_{i0}$. 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 $b = 1, \dots, B$ and $i = 1, \dots, m$. 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

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

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):

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

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

rate estimate is

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

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

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

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

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

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

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

where $\hat{\pi}_0$ 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:

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

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 $\hat{w}_i = 1$, 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:

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

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

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

The overall estimate $\hat{\pi}_0$ 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:

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

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 $\widehat{\text{ETP}} = (\# \text{ significant genes})(1 - \widehat{\text{FDR}})$, as just shown above. Since each method is compared at the same value of $\widehat{\text{FDR}}$, it follows that comparing the number of significant genes is equivalent to comparing the methods based on $\widehat{\text{ETP}}$, 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 $\widehat{\text{ETP}}$ 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

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

where as before tests $1, 2, \dots, m_0$ 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, $(x_{ij} - \hat{\mu}_{i0})/\hat{\sigma}_i$. 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 $\hat{\sigma}_i$ 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:

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

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, $\hat{g}(\mathbf{x})/\hat{f}(\mathbf{x})$, 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 $b = 1, \dots, B$ iterations. The null statistics are calculated by $\hat{g}^{0b}(\mathbf{x}^{0b})/\hat{f}^{0b}(\mathbf{x}^{0b})$ where \hat{g}^{0b} and \hat{f}^{0b} are the new estimates based on \mathbf{x}^{0b} .

In our proposed procedure the null statistics are calculated by $\hat{S}_{\text{ODP}}(\mathbf{x}_i^{0b})$, 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 perfectly¹.

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

¹That 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 \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.

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]) }

```

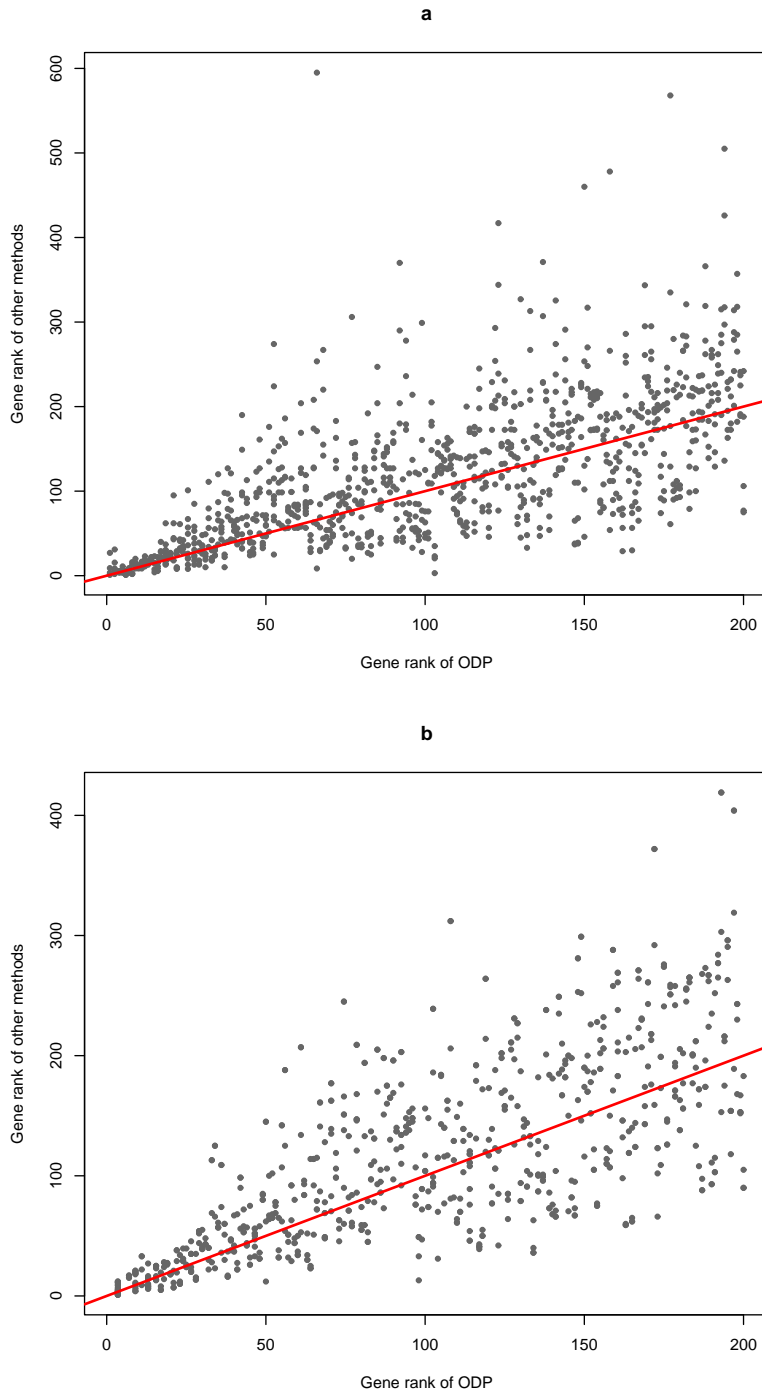


Figure 6: The gene ranking of the ODP versus the existing five methods when identifying differentially expressed genes from the Hedenfalk et al. data. For each of the top 200 ranked genes according to the ODP approach (x -axis), the ranking given by the other methods is plotted (y -axis). It can be seen 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.

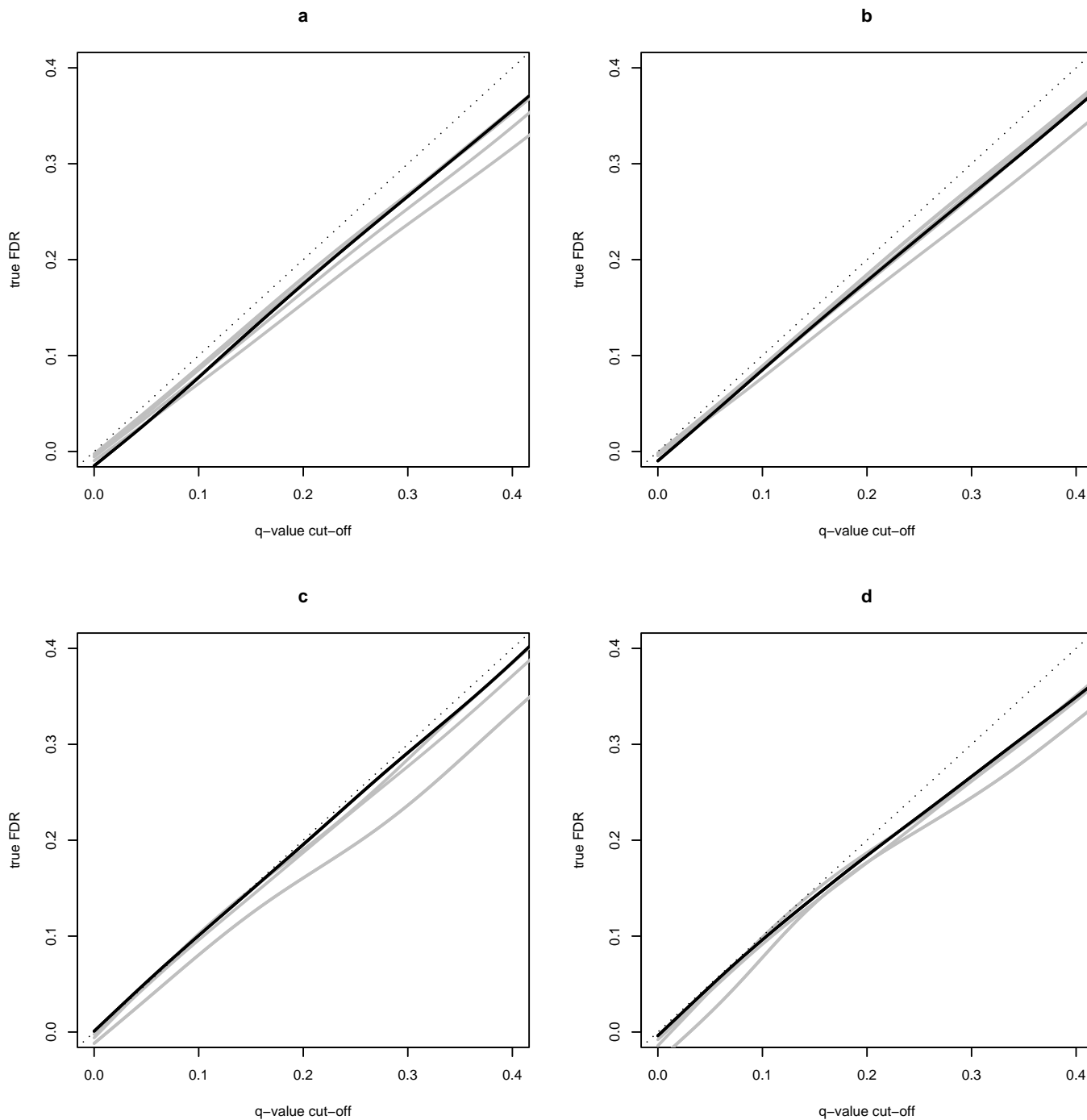


Figure 7: Plots verifying that each method considered controls the FDR by using the estimated q-value methodology of Storey (2002) and Storey & Tibshirani (2003). For each of the four simulation scenarios considered (a–d; see main text and Section 13 above), the estimated q-values versus the true FDR 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.

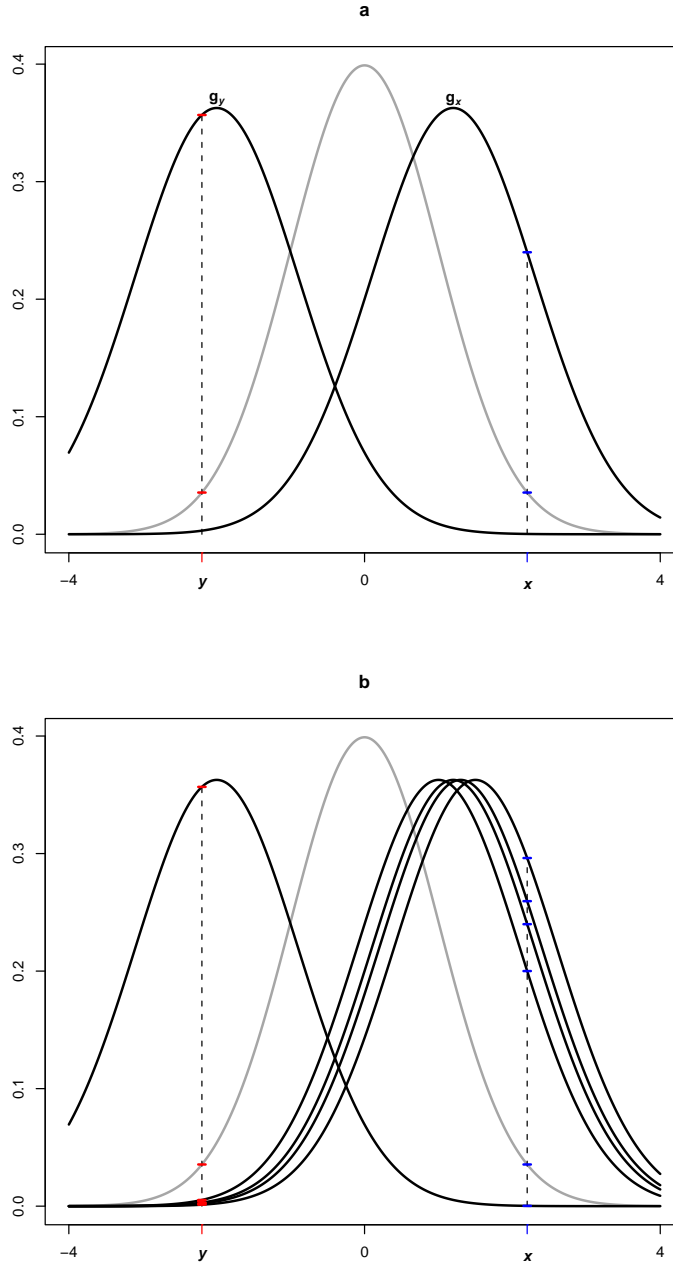


Figure 8: *Color version of Figure 1 from the main text.* Plots comparing the NP testing approach to the ODP testing approach through a simple example. **(a)** NP approach. The null (grey) and alternative (black) probability density functions of a single test. For observed data x and y , the statistics 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 negative mean.

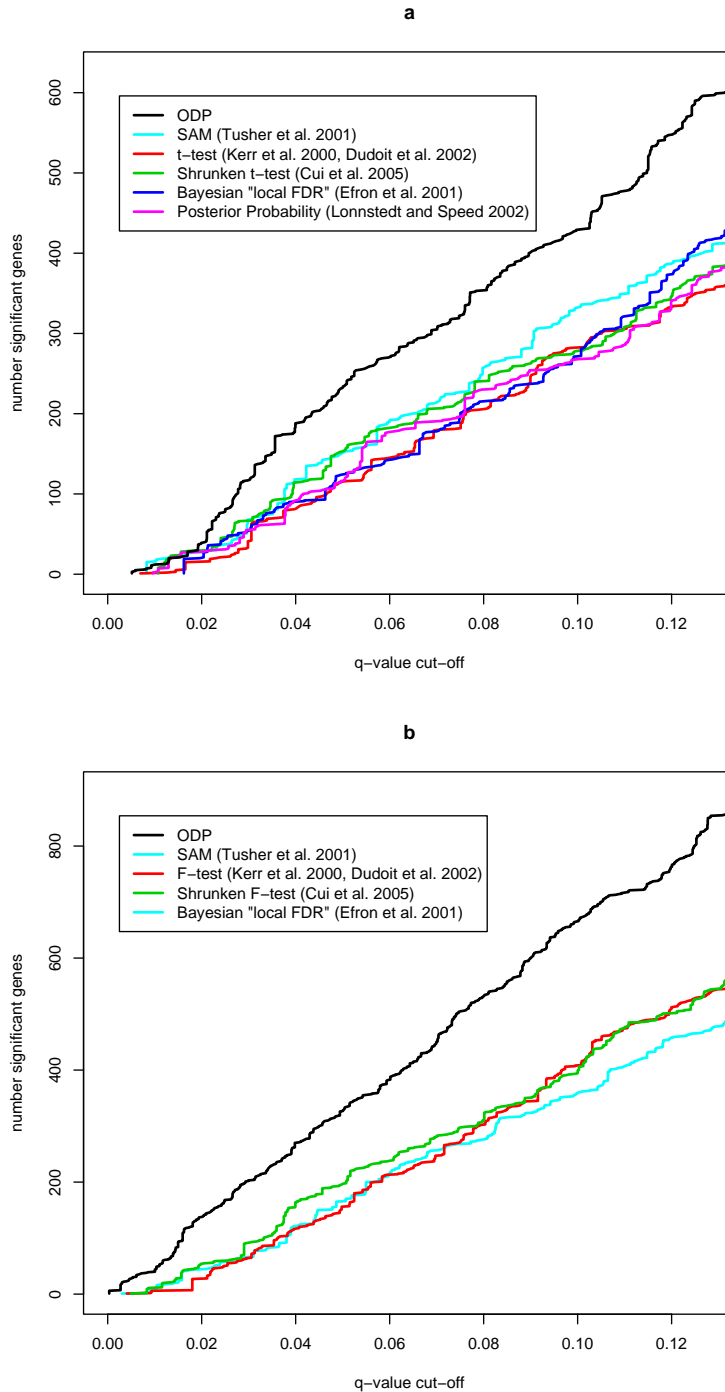


Figure 9: Color version of Figure 2 from the main text. A comparison of the ODP approach to five leading methods for identifying differentially expressed genes (described in the text). The number of genes found to be significant by each method over a range of estimated q-value cut-offs is shown. The 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.

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.