BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 509–515
doi:10.1093/bioinformatics/btq701

Gene expression

Advance Access publication December 24, 2010

A computationally efficient modular optimal discovery procedure

Sangsoon Woo[1], Jeffrey T. Leek[2] and John D. Storey[3,*]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205 and [3]Lewis-Sigler Institute for Integrative Genomics and Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: It is well known that patterns of differential gene expression across biological conditions are often shared by many genes, particularly those within functional groups. Taking advantage of these patterns can lead to increased statistical power and biological clarity when testing for differential expression in a microarray experiment. The optimal discovery procedure (ODP), which maximizes the expected number of true positives for each fixed number of expected false positives, is a framework aimed at this goal. Storey *et al.* introduced an estimator of the ODP for identifying differentially expressed genes. However, their ODP estimator grows quadratically in computational time with respect to the number of genes. Reducing this computational burden is a key step in making the ODP practical for usage in a variety of high-throughput problems.

Results: Here, we propose a new estimate of the ODP called the modular ODP (mODP). The existing 'full ODP' requires that the likelihood function for each gene be evaluated according to the parameter estimates for all genes. The mODP assigns genes to modules according to a Kullback–Leibler distance, and then evaluates the statistic only at the module-averaged parameter estimates. We show that the mODP is relatively insensitive to the choice of the number of modules, but dramatically reduces the computational complexity from quadratic to linear in the number of genes. We compare the full ODP algorithm and mODP on simulated data and gene expression data from a recent study of Morrocan Amazighs. The mODP and full ODP algorithm perform very similarly across a range of comparisons.

Availability: The mODP methodology has been implemented into EDGE, a comprehensive gene expression analysis software package in R, available at http://genomine.org/edge/.

Contact: jstorey@princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 12, 2010; revised on November 10, 2010; accepted on December 15, 2010

1 INTRODUCTION

Since the development of microarrays, a large number of methods have been proposed to identify genes that are differentially expressed across biological conditions. Methods exist that borrow strength across genes to shrink variances, apply data-adaptive thresholds to traditional statistics or calculate hierarchical Bayesian posterior probabilities (Cui *et al.*, 2005; Efron *et al.*, 2001; Lonnstedt and Speed, 2002; Newton *et al.*, 2004; Smyth, 2004; Tusher *et al.*, 2001). Recently, Storey *et al.* (2007) proposed an approach called the optimal discovery procedure (ODP) that borrows strength across genes with similar expression patterns when testing them for differential expression. The ODP, which must be estimated in practice, maximizes the expected number of true discoveries for a fixed expected number of false discoveries. This optimality property makes the ODP an attractive choice for use in analyzing gene expression and other high-throughput data.

Conceptually, the ODP uses information about differential expression patterns across all genes to inform the decision about any specific gene. Figure 1 shows a simple simulated dataset that illustrates the ODP concept. The black box highlights the genes that are differentially expressed among groups. The first set of differentially expressed genes are upregulated in the first two groups and downregulated in the third. The second set of differentially expressed genes are downregulated in the second group and upregulated in the first and third. The third set is upregulated in the second group and downregulated in the others. In this example, each pattern of differential expression is shared across genes. The number of genes sharing each pattern is different, and only three of the six possible differential expression patterns are present. The ODP directly utilizes this information, stemming from the idea that if a gene shares an expression pattern with other genes that have been identified as differentially expressed, then it is more likely to be differentially expressed as well. It has been shown that the ODP is more powerful for detecting differential expression than existing methods such as the traditional t-test (or F-test), a shrunken t-test, SAM and empirical Bayes methods (Storey *et al.*, 2005, 2007).

The ODP statistic is related to the commonly used likelihood ratio (LR) test statistics, also known as the Neyman–Pearson statistic (Lehmann, 1986). Suppose we have observed an $n \times 1$ vector of expression data $\mathbf{x}_i$ for the i-th gene. The traditional LR statistic, which is optimal when testing a *single* hypothesis, evaluates the likelihood under the null hypothesis of no differential expression for that gene L_{i0} and the alternative hypothesis L_{iA}, and forms their ratio $L_{iA}(\mathbf{x}_i)/L_{i0}(\mathbf{x}_i)$ as the test statistic. If the ratio is large enough, then the gene is called differentially expressed.

The ODP has a similar structure, except the ratio is taken of all alternative likelihoods to all null likelihoods given the gene's data $\mathbf{x}_i$, where these likelihoods are evaluated across all genes. The ODP statistic is the ratio $\sum_{\text{alt}} L_{jA}(\mathbf{x}_i)/\sum_{\text{null}} L_{j0}(\mathbf{x}_i)$. As with the LR statistics, the ODP statistic for a test also captures evidence against the null hypothesis in favor of the alternative hypothesis.

*To whom correspondence should be addressed.

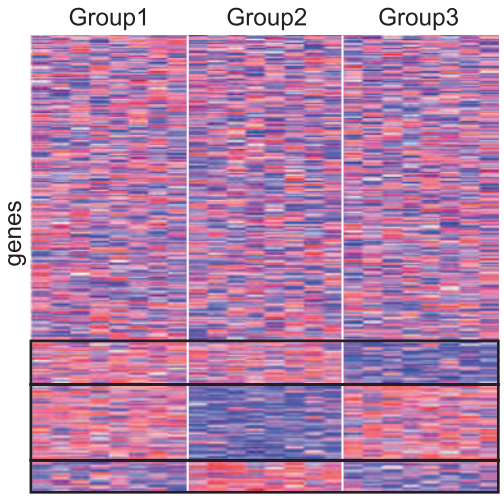

Fig. 1. A heatmap of simulated gene expression data for a study comparing three groups. The genes inside the black box show three common gene expression patterns; the first pattern is downregulated for groups 1 and 2 and upregulated for group 3. The second pattern is downregulated for groups 1 and 3 and upregulated for group 2. The third pattern is downregulated for group 2 and upregulated for groups 1 and 3. The number of genes sharing each pattern is different, and only three of the six possible differential expression patterns are present. The ODP is designed to utilize these expression patterns to improve inference of differential expression.

But in contrast to the univariate LR test, the ODP summarizes the evidence against the null taking into account the information from the other (possibly related) tests being performed. When another gene is uninformative, it contributes essentially nothing to the above ODP statistic because its likelihood is very low.

Although the ODP statistic is more powerful than traditional statistics developed for testing single hypotheses, it requires the evaluation of a large number of likelihoods for each dataset. For each gene, the number of terms to calculate in the ODP statistic is on the order of the total number of genes, resulting in a computational cost that grows quadratically in the number of genes when doing a genome-wide analysis. The original ODP estimator introduced in Storey *et al.* (2007) performs this exhaustive set of calculations. However, if groups of genes share common expression patterns, they will have similar probability distributions, and therefore their calculations can be compressed into a single computational step. Here, we propose to identify modules of genes based on similarity of probability distributions using a clustering scheme based on the Kullback–Leibler distance. The ODP statistic can then be calculated using only the distributions derived from the centroids of these modules and weighted by the number of genes in that module. Since the number of modules is much smaller than the number of genes, the ODP computation will be substantially reduced. However, the performance of and results from the modular ODP approach are very similar to the full ODP approach. In addition to a substantial reduction of computation, the unknown parameters for each centroid can be accurately estimated because multiple genes are used.

2 THE OPTIMAL DISCOVERY PROCEDURE

Gene expression and other high-throughput data can be thought of as a set of related experiments performed simultaneously. Suppose there are m genes in an experiment; then for each gene there is an $n \times 1$ vector of data $\mathbf{x}_i$ corresponding to the gene expression measurements, for each of n individuals for that gene. A usual goal in high-throughput data analysis is to test a statistical hypothesis for each gene, for example, testing the hypotheses that each gene has constant expression across some groups of interest versus the hypotheses that some genes mean expression varies by group. In other words, testing the hypotheses: $H_0 : \boldsymbol{\mu}_i = \mu_i^0 \mathbf{1}$ versus $H_1 : \boldsymbol{\mu}_i = \boldsymbol{\mu}_i^1$ where $\boldsymbol{\mu}_i^1$ parameterizes difference in means across samples.

For simplicity, in the remainder of the discussion we will assume that the data $\mathbf{x}_i$ are sample of n independent observations from a Normal distribution with mean vector $\boldsymbol{\mu}_i$ and common variance σ_i^2. However, any other data generating distribution can be substituted in the discussion and methods that follow. For Normal distributed data, the likelihood is

$$L(\boldsymbol{\mu}_i, \sigma_i | \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\sum_{j=1}^{n} \frac{-(x_{ij}-\mu_{ij})^2}{2\sigma_i^2}}.$$

The so-called generalized LR test is an approximation to the most powerful test given a fixed type one error rate. Using the Normal model, the generalized LR test statistic is given by

$$\widehat{\mathcal{S}}_{\mathrm{LR}}(\mathbf{x}_i) = \frac{L(\widehat{\boldsymbol{\mu}}_i^1, \widehat{\sigma}_i^1 | \mathbf{x}_i)}{L(\widehat{\boldsymbol{\mu}}_i^0, \widehat{\sigma}_i^0 | \mathbf{x}_i)},$$

where $(\widehat{\boldsymbol{\mu}}_i^1, \widehat{\sigma}_i^1)$ and $(\widehat{\boldsymbol{\mu}}_i^0, \widehat{\sigma}_i^0)$ are the maximum likelihood parameter estimates under the alternative and the null hypotheses, respectively (Lehmann, 1986). In the case of Normal data for a two sample comparison, the generalized LR statistic is equivalent to the standard two-sided t-test. However, the LR statistic is far more general and can be fit to a wide range of data types.

The ODP uses a concept similar to the Neyman–Pearson approach, except the ODP approach was developed for testing multiple hypotheses. Rather than maximizing the power for a fixed false positive rate of a single test, the ODP maximizes the overall number of expected true positives (ETP) for a fixed level of expected false positives (EFP) among multiple hypothesis tests. The ETP is simply the sum of the power across all truly alternative tests and the EFP is the sum of the false positive rates across all truly null tests. The ODP, like the LR test, is optimal when the null and alternative distributions are known. The ODP can be estimated by using the same principles for forming the generalized LR statistics, which is an estimate of the theoretical optimal Neyman–Pearson LR statistic.

For identifying differentially expressed genes in a microarray study, a simple estimate of the full ODP has been developed (Storey *et al.*, 2007). The first step is to calculate the maximum likelihood estimates for each gene under the alternative and null hypotheses, $(\widehat{\boldsymbol{\mu}}_i^1, \widehat{\sigma}_i^1)$ and $(\widehat{\boldsymbol{\mu}}_i^0, \widehat{\sigma}_i^0)$ for $i = 1, \ldots, m$. To calculate the estimated full ODP statistic for a given gene, the gene's likelihood function is evaluated at all of these fitted maximum likelihood estimates, summed over all tests, and the ratio between the alternative likelihood sum to the null likelihood sum is formed. In mathematical notation, the estimated ODP statistic for gene i can be written as follows:

$$\widehat{\mathcal{S}}_{\mathrm{ODP}}(\mathbf{x}_i) = \frac{\sum_{j=1}^{m} L(\widehat{\boldsymbol{\mu}}_j^1, \widehat{\sigma}_j^1 | \mathbf{x}_i)}{\sum_{j=1}^{m} L(\widehat{\boldsymbol{\mu}}_j^0, \widehat{\sigma}_j^0 | \mathbf{x}_i)}. \quad (1)$$

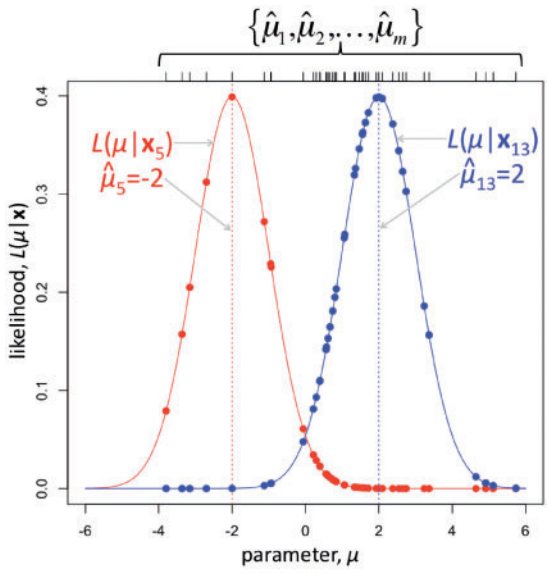

Fig. 2. A demonstration of the difference between the ODP approach and LR statistic. Suppose that hypothesis tests $H_0: \mu = 0$ versus $H_1: \mu \neq 0$ are performed on $\mu_1, \mu_2, \ldots, \mu_m$ based on respective datasets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$. Shown are the likelihood functions for test 5, $L(\mu|\mathbf{x}_5)$ in red, and test 13, $L(\mu|\mathbf{x}_{13})$ in blue. Their maximum likelihood estimates are such that $L(\widehat{\mu}_5|\mathbf{x}_5) = L(\widehat{\mu}_{13}|\mathbf{x}_{13})$, implying that they would produce equal LR statistics. The ODP utilizes information from all of the maximum likelihood estimates $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_m$, shown at the top of the plot. These tend to be more similar to $\widehat{\mu}_{13}$ than $\widehat{\mu}_5$, lending greater evidence against the null hypothesis for test 13. The ODP quantifies this evidence by calculating the likelihood functions over all maximum likelihood estimates, shown as red dots for test 5 and in blue dots for test 13. It can be seen that $\sum_{i=1}^{m} L(\widehat{\mu}_j|\mathbf{x}_{13}) \gg \sum_{i=1}^{m} L(\widehat{\mu}_j|\mathbf{x}_5)$, implying that the ODP statistic for test 13 would be larger than that for test 5. This makes sense in that there are many more positive $\widehat{\mu}_j$ than negative, so we should attribute stronger evidence against the null hypothesis to those tests with positive estimates. In more complex situations such as those encountered in gene expression studies, this aggregation of information becomes even more useful.

An intuitive understanding of the ODP statistic and how it relates to the traditional LR statistic is explained in Figure 2. Storey *et al.* (2007) presented a more general form of (1) as well as some steps one may take to remove ancillary information, which may be incorporated into our proposed method (see Supplementary Material).

Evaluating (1) requires $2m$ likelihood calculations for each of the m tests, resulting in $2m^2$ likelihood calculations. The statistical significance for this full ODP statistic is evaluated using a bootstrap procedure. Because the ODP statistic is estimated for each bootstrapped dataset, this can lead to substantial computational costs.

In most experiments, the data for many features will follow a common pattern of variation. The pattern of variation that is relevant to the inference is that captured by the probability distribution used to model each gene's data, parameterized by the (μ_i, σ_i). If genes a and b have similar relevant patterns of variation, then $(\mu_a, \sigma_a) \approx (\mu_b, \sigma_b)$, as well as their likelihood values $L(\mu_a, \sigma_a|\mathbf{x}_i) \approx L(\mu_b, \sigma_b|\mathbf{x}_i)$ for any given gene i. Thus, it is not necessary to do each calculation individually in (1), but rather approximate them with a single 'average' calculation. There may be many more than two genes with common patterns of variation, thereby allowing us to reduce the computation even further.

A key problem is how to decide if $(\mu_a, \sigma_a) \approx (\mu_b, \sigma_b)$, and how to identify larger sets of genes with this similarity. Even more so, we want $L(\mu_a, \sigma_a|\mathbf{x}) \approx L(\mu_b, \sigma_b|\mathbf{x}_i)$ for all genes $i = 1, \ldots, m$, because agglomerating genes a and b in the ODP statistic will be applied in the calculation of every gene's ODP statistic. To this end, we use a modified Kullback–Leibler divergence (Kullback and Leibler, 1961), which indeed quantifies the probabilistic distance between $L(\widehat{\mu}_a, \widehat{\sigma}_a|\mathbf{x})$ and $L(\widehat{\mu}_b, \widehat{\sigma}_b|\mathbf{x})$ over all $\mathbf{x} \in \mathbf{R}^n$. The Kullback–Leibler divergence measures the discrepancy between two distributions. Because the Kullback–Leibler divergence is asymmetric, we use a symmetric version that is sometimes called the Kullback–Leibler distance.

We assign each gene to one of K modules by utilizing a clustering algorithm based on the Kullback–Leibler distance (Nielsen and Nock, 2009). This creates K centroid estimates of the parameters for the alternative model fits and the null model fits. Then for each gene, we only evaluate its likelihood function at the centroid parameters and weight it by the number of genes in that module. Since the number of modules is much smaller than the number of genes, this approach substantially reduces the computational burden of the approach. We reduce $2m^2$ calculations to $2Km$, where $K \ll m$. If K stays approximately fixed, then the computational cost of the proposed modular ODP (mODP) algorithm grows linearly in the number of genes. At the same time, we show that the mODP gives nearly identical inference results to the original full ODP estimate.

3 METHODS

Our approach to calculating the ODP statistics has three steps: (i) cluster genes into K modules based on similarity of expression variation captured by the Kullback–Leibler distance; (ii) evaluate each gene's likelihood function at each module centroid model fit, weighted by the number of genes assigned to that module; and (iii) aggregate these weighted centroid likelihood calculations into modular ODP (mODP) statistics.

For the first step, we use a modification of the well-known Kullback–Leibler (KL) divergence as our metric for measuring how similar two genes' estimated probability distributions are. Let F_a and F_b be two continuous probability distributions with a common support and corresponding probability density functions, f_a and f_b. The KL divergence between these two distributions is given by

$$\begin{aligned} \mathrm{KL}(F_a, F_b) &= \mathrm{E}_{F_a}\left[\log\left(f_a(\mathbf{x})/f_b(\mathbf{x})\right)\right] \\ &= \int \log\left(f_a(\mathbf{x})/f_b(\mathbf{x})\right) f_a(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

The KL divergence is not a symmetric measure, so we use the following KL distance:

$$d(F_a, F_b) = \mathrm{KL}(F_a, F_b) + \mathrm{KL}(F_b, F_a).$$

The KL distance between two of the Normal distributions that we consider here is calculated to be:

$$d\left(N(\boldsymbol{\mu}_a, \sigma_a^2), N(\boldsymbol{\mu}_b, \sigma_b^2)\right) =$$

$$\frac{1}{2}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)\left(\frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}\right) + \frac{n}{2}\left(\frac{\sigma_a^2}{\sigma_b^2} + \frac{\sigma_b^2}{\sigma_a^2}\right) - n.$$

We construct modules by extending K-means clustering (Nielsen and Nock, 2009), which is typically based on Euclidean distance of a gene's expression vector to that of K cluster centroids, to a KL distance-based

approach. The distance between a gene and a module is based on the KL distance between the gene's estimated probability distribution and the 'average distribution' of the genes within a module. The mODP estimation algorithm proceeds as shown below. We represent the gene's estimated probability distribution by its maximum likelihood parameters under the alternative hypothesis, $(\widehat{\boldsymbol{\mu}}_i^1, \widehat{\sigma}_i^1)$. The 'average distribution' of the module is simply based on the average of the parameter estimates for every gene in that module. Therefore, the approach by which we construct clusters is derived from K-means clustering, except we replace Euclidean distance with KL distance, and we replace expression vectors and centroids with parameter estimates and averaged parameter estimates.

In constructing modules, we use the alternative hypothesis estimates rather than the null hypothesis estimates because the former is fit under the unconstrained model, thereby allowing for each gene to be truly null or truly alternative without having to introduce an extra estimation step. Also, when incorporating data transformations to remove ancillary information, as proposed by Storey *et al.* (2007) (see Supplementary Material), we get $\widehat{\boldsymbol{\mu}}_i^0 = \mathbf{0}$ for all i making them uninformative for module construction.

Once the module construction is completed [step (i) above], then steps (ii) and (iii) become straightforward. Our proposed method is summarized in the following algorithm.

Algorithm. The Modular Optimal Discovery Procedure (mODP)

1. For a user-chosen number of modules K, initiate the module parameter estimates by randomly selecting K of the m genes and setting their alternative hypothesis estimates to be the module parameter estimates. Specifically, for $k = 1, \ldots, K$, the module k is parameterized by $\tilde{\boldsymbol{\mu}}_k = \widehat{\boldsymbol{\mu}}_{i(k)}^1$ and $\tilde{\sigma}_k = \widehat{\sigma}_{i(k)}^1$, where $i(k)$ is a randomly chosen gene index among the m.

2. For each gene $i = 1, \ldots, m$ and each module $k = 1, \ldots, K$, calculate their KL distance using the formula:

$$d_{ik} = \frac{1}{2}(\tilde{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_i^1)^T(\tilde{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_i^1)\left(\frac{1}{\tilde{\sigma}_k^2} + \frac{1}{(\widehat{\sigma}_i^1)^2}\right) + \frac{n}{2}\left(\frac{\tilde{\sigma}_k^2}{(\widehat{\sigma}_i^1)^2} + \frac{(\widehat{\sigma}_i^1)^2}{\tilde{\sigma}_k^2}\right) - n.$$

3. Assign gene i to the closest module in terms of KL distance, calculated by $\arg\min_{1 \le k \le K} d_{ik}$.
4. For these new module assignments, calculate their updated parameters. Let R_k be the set of indices of genes assigned to module k and $|R_k|$ be the number of genes in that module.

$$\tilde{\boldsymbol{\mu}}_k = \frac{1}{|R_k|}\sum_{j \in R_k} \widehat{\boldsymbol{\mu}}_j^1$$

$$\tilde{\sigma}_k^2 = \frac{1}{|R_k|}\sum_{j \in R_k} \left(\widehat{\sigma}_j^1\right)^2.$$

5. Repeat Steps 2–4 until the centroids are fixed in the sense that $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\sigma}_k^2$ differ less than a user chosen ϵ between two consecutive iterations. Calculate the final module estimates under the alternative and null hypotheses as follows:

$$\tilde{\boldsymbol{\mu}}_k^1 = \tilde{\boldsymbol{\mu}}_k$$

$$\tilde{\sigma}_k^1 = \tilde{\sigma}_k$$

$$\tilde{\boldsymbol{\mu}}_k^0 = \frac{1}{|R_k|}\sum\nolimits_{j \in R_k} \widehat{\boldsymbol{\mu}}_j^0$$

$$\tilde{\sigma}_k^0 = \sqrt{\frac{1}{|R_k|}\sum_{j \in R_k} \left(\widehat{\sigma}_j^0\right)^2}$$

6. For each gene $i = 1, 2, \ldots, m$, calculate the mODP statistics according to the following formula:

$$\widehat{\mathcal{S}}_{\text{mODP}}(\mathbf{x}_i) = \frac{\sum_{k=1}^K L(\tilde{\boldsymbol{\mu}}_k^1, \tilde{\sigma}_k^1 | \mathbf{x}_i) \cdot |R_k|}{\sum_{k=1}^K L(\tilde{\boldsymbol{\mu}}_k^0, \tilde{\sigma}_k^0 | \mathbf{x}_i) \cdot |R_k|}. \quad (2)$$

7. From the mODP formula above, obtain P-values and FDR q-values by using the bootstrap, exactly as proposed in Storey *et al.* (2007).

The mODP algorithm requires the user to choose the number of modules K in advance. Estimating the number of clusters in data is a notoriously difficult problem, because much biological interpretation is made of the genes contained in each cluster. However, in our setting, the clustering may be used simply as a numerical tool, making this choice much less crucial in that one is not required to make any biological interpretation of the clusters. The rule of thumb we propose is to set K large enough so that K is greater than the number of distinct patterns of expression variation, but not so large that the gain in computational speed is diminished. In the numerical results below, we have observed that $K = 50$ seems to be a well behaved choice. This may be data dependent, but one may always compare the results for different values of K, as we do below, before making a final choice.

On the other hand, embedded in our mODP method is a potentially useful new clustering algorithm. Whereas clustering is typically performed in an unsupervised manner, our algorithm allows one to cluster genes based on how the model of interest fits the data. In such a case, the choice of K becomes more important and exploring a data-driven choice of the number of clusters may be more relevant. It is also possible that this clustering algorithm driven by study design could be incorporated with more sophisticated modular clustering frameworks (Zhang and Horvath, 2005). While this is potentially a very interesting direction, it is beyond the scope of this work.

4 RESULTS

The mODP estimator (2) has two advantages over the full ODP estimator (1). First, the number of modules is generally much smaller than the number genes ($K \ll m$), thereby dramatically reducing the computational burden. The second advantage is that the averaged parameter fits within each module will be more stable than individual gene's parameter estimates. We now compare the behavior of the mODP estimator to the full ODP estimator both on simulated gene expression data and on data from a study comparing gene expression levels in human leukocytes from individuals living in three different environments. We show below that the mODP algorithm offers nearly identical results as the full ODP, while indeed requiring substantially less computing time.

4.1 Simulation results

We compared the mODP to the full ODP on a range of simulated examples; the R code and details of the simulations appear in the Supplementary Material. First, we compared the computational time for the full ODP approach versus the mODP approach with $K = 50$. Figure 3 shows the relative CPU times required to calculate the mODP and full ODP statistics for a set of genes under one of the simulation scenarios, as the number of genes increases from 100 to 10 000. (The other scenarios show equivalent results.) The computational time for the mODP, which includes the time required for clustering, grows nearly linearly in the number of genes, while the full ODP is closer to quadratic growth. Given that the ODP statistics must be recomputed for all genes for each null bootstrap

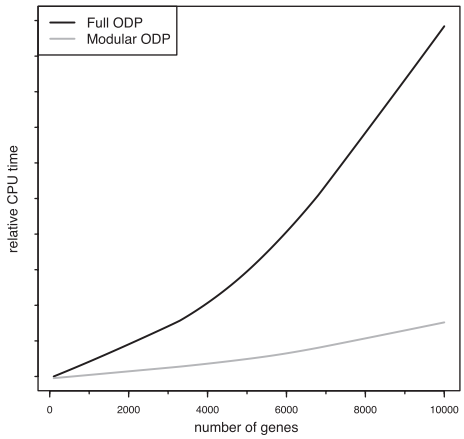

Fig. 3. A plot of relative CPU time for increasing numbers of genes for the mODP and the full ODP estimators under one of the simulation scenarios. The full ODP grows approximately quadratically in the number of genes while the mODP grows nearly linearly.

sample, the computational savings can be substantial in practice. To make this comparison, we used an Apple Mac Pro machine with OS X version 10.5.8, 2.66 GHz Intel Core 2 Duo processor and 4 GB of RAM.

Next we examined whether the mODP statistics produce similar results to the full ODP statistics. Our simulations followed the structure from the simulation study in Storey *et al.* (2007) to provide comparable results. The full ODP has already been compared to a number of other popular methods (Cui *et al.*, 2005; Efron *et al.*, 2001; Lonnstedt and Speed, 2002; Smyth, 2004; Tusher *et al.*, 2001) in Storey *et al.* (2007), as well as a Bayesian version of the full ODP in Guindani *et al.* (2009). Since our results show that the mODP provides nearly identical significance results to the full ODP, we do not repeat these comparisons here.

We simulated eight different types of gene expression studies, each corresponding to a particular set of parameters and experimental design. Four simulated studies correspond to two group comparisons and four correspond to three group comparisons. Within each of these two sets, the same signal structure is used, but we vary the variance structure. For both the two and three sample studies, the simplest case is a constant variance across all genes, followed by variances simulated from a Uniform distribution, from a Gamma distribution and from a more heterogeneous mixture of Uniform distributions. R scripts for simulating these datasets can be found in the Supplementary Material. In each case, we plot the number of significant genes across a range of estimated q-value cutoffs (Storey and Tibshirani, 2003), averaged over 100 simulated studies (Fig. 4). We also compared the estimators in terms of the true EFP and ETP (Supplementary Figs S1 and S2), showing similar results to the above comparison.

For simple variance structures, the mODP provides nearly identical performance to the more computationally intensive full ODP regardless of the number of modules K and no matter how many groups are compared. As the variance structure becomes more complex, it appears that more modules are required for the mODP to achieve the same performance as the full ODP, especially in the three group comparison. The mODP is more likely sensitive to the choice of the number of modules in a three-group comparison because the parameter structure is more complicated. However, in all simulated scenarios $K = 50$ modules or more leads to results that are nearly identical to the full ODP.

We compared the numerical values of the mODP and full ODP statistics and the gene significance rankings for $K = 50$ and $K = 200$ (Supplementary Figs S3–S5). It can be seen that the mODP and full ODP again produce similar results. We also verified that the random initial cluster centers do not heavily influence the mODP values nor the relative rankings that they produce (Supplementary Figs S7 and S8).

4.2 Environmental differential expression

Idaghdour *et al.* (2008) measured gene expression from a human population of Moroccan Amazighs composed of three different lifestyles. They collected leukocyte samples from peripheral blood to profile gene expression in 16 Bedouin, 18 Anza and 12 Sebt-Nabor individuals. The Bedouin individuals have traditional nomadic lives on the fringe of the Sahara desert near the town of Errachidia, the Anza individuals are from the coastal city of Anza near Agadir and the Sebt-Nabor individuals come from a rural mountainous region in Agadir. In total, 10 177 transcripts were expressed across the 46 samples. Details of the gene expression profiling process are described in Idaghdour *et al.* (2008). We refer to samples from Bedouin as 'Desert', samples from Sebt-Nabor as 'Village' and samples from Anza as 'Agadir'.

We conducted pairwise comparisons for all pairs of the Agadir, Village and Desert groups and also looked for differential expression across all three groups simultaneously. The plots of the number of significant genes for each q-value for both the mODP and the full ODP are shown in Figure 5. Again we considered the performance over a range of module numbers K for the mODP approach. As the plots show, the mODP and full ODP perform nearly identically when $K \geq 50$. We also compared the gene rankings produced by the full ODP to the mODP for $K = 50$ (Supplementary Fig. S6). It can be seen that the two methods produce similar gene rankings, meaning that they identify nearly the exact same genes as being differentially expressed.

5 DISCUSSION

The optimal discovery procedure (ODP) is a powerful approach for the analysis of high-throughput gene expression data (Storey *et al.*, 2007; Storey *et al.*, 2007). However, the full ODP requires the computation of a large number of likelihoods to evaluate the statistic for any specific gene. This leads to computational costs that grow quadratically in the number of genes. Since significance of these statistics is typically evaluated by a non-parametric bootstrap approach requiring many sets of ODP statistics to be calculated, there is a strong need for methods that reduce the computational cost of evaluating these statistics. Here, we have introduced a new approach for calculating ODP statistics, based on forming probabilistic gene modules using the Kullback–Leibler distance and greatly reducing the number of likelihood calculations making up each statistic. The mODP statistics are formed from a small number of likelihood calculations, making the computation grow nearly linearly in the number of genes. Even though the mODP statistics are substantially faster to calculate, we have shown that they produce nearly identical

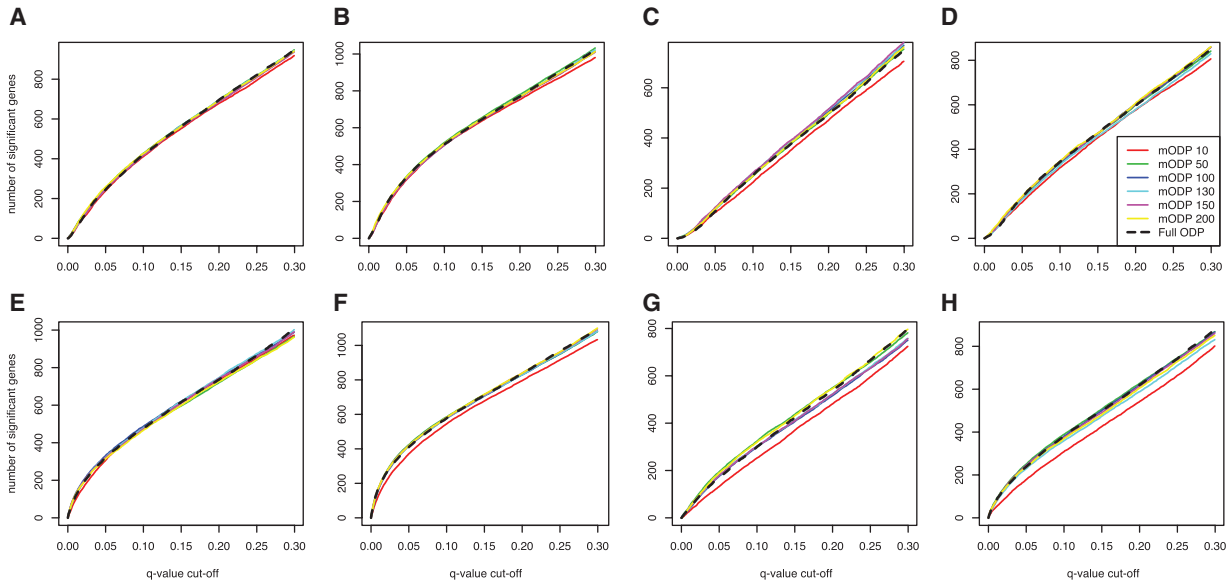

Fig. 4. A comparison of the mODP and the full ODP method based on simulated data. Each panel is the average number of genes called significant for each q-value cutoff over 100 simulated datasets. Solid colored lines are the proposed mODP method for different numbers of modules K and the black dashed line is the full ODP. The simulations correspond to (**A**) two group comparison, fixed equal variances, (**B**) two group comparison, variances Uniform sampled, (**C**) two group comparison, variances Gamma sampled and (**D**) two group comparison, variances Uniform mixture sampled, (**E**) three group comparison, fixed equal variances, (**F**) three group comparison, variances Uniform sampled, (**G**) three group comparison, variances Gamma sampled and (**H**) three group comparison, variances Uniform mixture sampled.

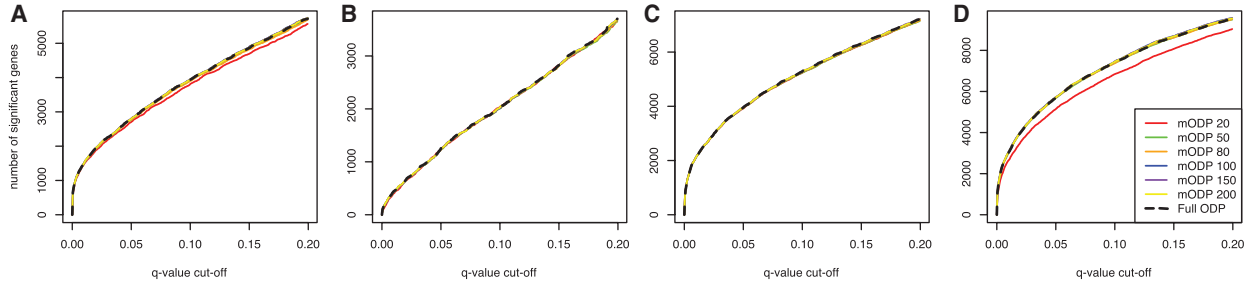

Fig. 5. A comparison of the mODP and the full ODP approaches on the Morrocan data from Idaghdour *et al.* (2008). In each plot, the number of significant genes is plotted versus the corresponding q-value cutoff. (**A**) Agadir versus Village, (**B**) Agadir versus Desert, (**C**) Desert versus Village and (**D**) Agadir versus Desert versus Village (three group comparison). The mODP performs nearly identically to the full ODP, particularly when $K \geq 50$.

results to the full ODP statistics in both simulated and real data examples.

Even though the mODP requires the user to decide the number of modules in advance, we have shown that the mODP statistics are relatively robust to the choice of the number of modules. In both the simulated and real data examples, it was observed that 50 modules or more were sufficient to match the performance of the full ODP. Also because the mODP method borrows strength across multiple genes, the averaged parameter estimates defining each module form more stable estimates of gene expression variation relevant to the study and may contribute important information beyond unsupervised clustering. Although we have used the Normal likelihood in the formulation of our method, justified because the data are continuous and can be shown to be approximately Normal, the ODP approach may be utilized with other probability distributions. The methodology presented in this article has been implemented in the EDGE software package (Leek *et al.*, 2006), freely available at http://genomine.org/edge/.

Funding: This research was supported in part by NIH grant HG002913.

Conflict of Interest: none declared.

REFERENCES

Cui,X. *et al.* (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Efron,B. *et al.* (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Guindani,M. *et al.* (2009) A bayesian discovery procedure. *J. Roy. Stat. Soc. Ser. B*, **71**, 905–925.

Idaghdour,Y. *et al.* (2008) A gemone-wide gene expression signature of environmental geography in leukocytes of moroccan amazighs. *PLoS Genet*., **4**, e1000052.

Kullback,S. and Leibler,R.A. (1961) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Leek,J.T. *et al.* (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, **22**, 507–508.

Lehmann,E.L. (1986) *Testing Statistical Hypotheses*. 2nd edn. Springer, Berlin.

Lonnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sin.*, **12**, 31–46.

Newton,M.A. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Nielsen,F. and Nock,R. (2009) Clustering multivariate normal distributions. In Nielsen,F. (ed.) *Emerging Trends in Visual Computing*. Springer, Berlin/Heidelberg, pp. 164–174.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Art. 3.

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

Storey,J.D. *et al.* (2005) Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 12837–12842.

Storey,J.D. *et al.* (2007) The optimal discovery procedure for large significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8**, 414–432.

Storey,J.D. *et al.* (2007) The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Roy. Stat. Soc. Ser. B*, **69**, 347–368.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Art. 17.