

Gene expression

The optimal discovery procedure for significance analysis of general gene expression studies

Andrew J. Bass and John D. Storey  *

Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 17, 2019; revised on June 18, 2020; editorial decision on July 21, 2020; accepted on August 17, 2020

Abstract

Motivation: Analysis of biological data often involves the simultaneous testing of thousands of genes. This requires two key steps: the ranking of genes and the selection of important genes based on a significance threshold. One such testing procedure, called the optimal discovery procedure (ODP), leverages information across different tests to provide an optimal ranking of genes. This approach can lead to substantial improvements in statistical power compared to other methods. However, current applications of the ODP have only been established for simple study designs using microarray technology. Here, we extend this work to the analysis of complex study designs and RNA-sequencing studies.

Results: We apply our extended framework to a static RNA-sequencing study, a longitudinal study, an independent sampling time-series study, and an independent sampling dose–response study. Our method shows improved performance compared to other testing procedures, finding more differentially expressed genes and increasing power for enrichment analysis. Thus, the extended ODP enables a favorable significance analysis of genome-wide gene expression studies.

Availability and implementation: The algorithm is implemented in our freely available R package called `edge` and can be downloaded at <https://www.bioconductor.org/packages/release/bioc/html/edge.html>.

Contact: jstorey@princeton.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Genome-wide gene expression studies measure simultaneously the expression levels of thousands of genes using RNA-seq or DNA microarray technology. A primary objective in these studies is to discover biologically important genes by applying appropriate statistical tools to the data. One such approach is to apply a hypothesis testing procedure on a gene-by-gene basis to detect differentially expressed genes; for example, a *t*-test or *F*-test is commonly used to compare multiple biological groups. These test statistics are then ranked and a subset of genes with values above a specified threshold are deemed statistically significant. The significance threshold is chosen to control the false discovery rate (FDR), which is the expected proportion of false positives in the subset of selected genes. Thus, there are two key steps when selecting important genes: the ranking of test statistics and the selection of tests based on a significance threshold.

One commonly used class of testing procedures to rank genes is based on the likelihood ratio test (LRT). The LRT compares the goodness-of-fit between two models, namely, the alternative and null models. The test statistic is the ratio of the likelihood under the alternative model to the likelihood under the null model, where large values indicate evidence against the null model. It is popular

due to its optimality guarantees: the Neyman–Pearson (NP) lemma states that the LRT statistic is the most powerful testing procedure for a single hypothesis, where no other testing procedure can achieve more power at a fixed significance threshold (Neyman and Pearson, 1933). However, the LRT statistic may not provide the optimal ranking for multiple hypotheses (Storey, 2007). This is problematic in genomics where thousands of tests are typically performed.

When there are multiple hypotheses, such as in gene expression data, many testing procedures can improve upon the statistical power using information across genes. One such method is the optimal discovery procedure (ODP), which maximizes the number of expected true positives for a fixed number of expected false positives—a quantity related to the FDR (Storey, 2007). The ODP is a generalization of the NP lemma: while the NP lemma is optimal for a *single* hypothesis test, the ODP is optimal for *multiple* hypothesis tests. The ODP achieves the optimal ranking of test statistics by leveraging information across all tests when calculating the test statistic for each gene. Intuitively, the improvement in ranking stems from genes that follow similar patterns of expression (known as co-expression). This information is incorporated into the test statistic to quantify the evidence for differential expression. In the study by Storey *et al.* (2007), an approximation to the ODP performs favorably on DNA microarray studies compared to SAM (Tusher *et al.*,

2001), shrunken t -test or F -test (Cui *et al.*, 2005; Smyth, 2004), Bayesian local FDR (Efron *et al.*, 2001) and posterior probabilities (Lönstedt and Speed, 2002).

There are two main limitations when applying the ODP to genomic studies. First, the method was primarily developed for simple static experiments (e.g., comparing two conditions) and it has not yet been extended to more complex sampling designs. Second, the underlying assumption in the ODP is that the data are generated from a Normal distribution where the per-gene observations have the same variance (i.e., homoscedasticity). This is problematic for RNA-sequencing studies where the data are modeled using an over-dispersed Poisson distribution or a Normal distribution where the per-gene observations have different variances (i.e., heteroscedasticity) (Law *et al.*, 2014). Due to these constraints, the applicability of the ODP has been limited to static DNA microarray studies.

In this work, we extend the ODP to both complex study designs and RNA-sequencing studies. To incorporate non-linear responses commonly found in non-static studies, we utilize the regression spline methodology from Storey *et al.* (2005). An advantage of this approach is that it flexibly models gene expression responses within the least squares framework where the data are assumed to follow a Normal distribution with constant variance: this enables for a straightforward application of the ODP. For RNA-sequencing studies, we implement the same strategy in Law *et al.* (2014) to estimate the per-gene heteroscedasticity using the observed mean–variance relationship. We then use these estimated weights in a weighted least squares algorithm to adjust for unequal variances among observations. This transformation allows for the standard ODP framework to be utilized.

We apply our algorithm to three different experimental designs. The first is a ‘static sampling’ experiment, where samples are obtained at a fixed time point. For this example, we analyze a smoker study where smoking and non-smoking groups are compared to detect transcriptional differences in airway basal cells using RNA-seq technology. The second is an ‘independent sampling’ experiment, where subjects are independently sampled across time or dosage level. Here, we consider two independent sampling studies, namely, a time-series and a dose–response study. The former considers the effect of age on gene expression in the cortex region of the kidney and the latter explores breast cancer cell sensitivity in response to multiple 17β -estradiol doses. The final design is a ‘longitudinal sampling’ experiment, where subjects are sampled at multiple time points. As an example, we examine an endotoxin study which compares the leukocytes at a control group to those of an endotoxin-treated group across multiple time points.

The article is outlined as follows. Section 2 reviews background on the ODP and regression splines. We also review a computationally efficient implementation of the ODP called the modular optimal discovery procedure (mODP). Section 3 introduces our proposed algorithm, and Section 4 illustrates the results from our method on the four studies. We validate these results through comprehensive simulations.

2 Background

2.1 The optimal discovery procedure

The optimal (or ‘most powerful’) hypothesis test statistic for a single test is provided by the NP lemma (Neyman and Pearson, 1933). Given some observed data $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the NP lemma states that the ratio of the alternative likelihood $g(\mathbf{y})$ over the null likelihood $f(\mathbf{y})$ —known as the likelihood ratio—has the largest power for each false-positive rate compared to any other test statistic. Intuitively, this optimality arises because the data-generating process under each model is assumed to be known. For multiple hypotheses, the likelihood ratio could be applied on a test-by-test basis. However, potentially useful information across different hypotheses is ignored. As a consequence, the NP likelihood ratio test statistic may no longer be an optimal statistical test (Storey, 2007).

The ODP is a generalization of the NP lemma for multiple hypotheses. More specifically, consider gene expression

Algorithm 1: Clustering algorithm for the modular optimal discovery procedure (mODP)

Input: Dataset \mathbf{Y} , the alternative and null models, and the number of modules K .

Output: The parameter estimates for the alternative and null modules.

- 1 Estimate $(\hat{\boldsymbol{\mu}}_i^1, \hat{\sigma}_i^1)$ and $(\hat{\boldsymbol{\mu}}_i^0, \hat{\sigma}_i^0)$ under the alternative and null models, respectively, for genes $i = 1, 2, \dots, m$.
- 2 Initialize (\mathbf{c}_k, v_k) for modules $k = 1, 2, \dots, K$ by randomly selecting K genes. Set $c_{\text{prev}} = 0$.
- 3 **while** $|c - c_{\text{prev}}| > \epsilon$ **do**
- 4 $c_{\text{prev}} \leftarrow c$
- 5 **forall** $i \in \{1, 2, \dots, m\}$ **do**
- 6 **forall** $k \in \{1, 2, \dots, K\}$ **do**
- 7 $\delta_{ik} \leftarrow \frac{1}{2} (\mathbf{c}_k - \hat{\boldsymbol{\mu}}_i^1)^\top (\mathbf{c}_k - \hat{\boldsymbol{\mu}}_i^1) \left(\frac{1}{(v_k)^2} + \frac{1}{(\hat{\sigma}_i^1)^2} \right) + \frac{n}{2} \left(\frac{(v_k)^2}{(\hat{\sigma}_i^1)^2} + \frac{(\hat{\sigma}_i^1)^2}{(v_k)^2} \right) - n$
- 8 **end**
- 9 **end**
- 10 **forall** $k \in \{1, 2, \dots, K\}$ **do**
- 11 $R_k \leftarrow \{i \in \{1, 2, \dots, m\} : k = \arg \min_{1 \leq j \leq K} \delta_{ij}\}$
- 12 $\mathbf{c}_k \leftarrow \frac{\sum_{i \in R_k} \hat{\boldsymbol{\mu}}_i^1}{|R_k|}$
- 13 $v_k \leftarrow \sqrt{\frac{\sum_{i \in R_k} (\hat{\sigma}_i^1)^2}{|R_k|}}$
- 14 **end**
- 15 **end**
- 16 For modules $k = 1, 2, \dots, K$, the parameters under the alternative model are $\mathbf{c}_k^1 = \mathbf{c}_k$ and $v_k^1 = v_k$ and the parameters under the null model are $\mathbf{c}_k^0 = \frac{\sum_{i \in R_k} \hat{\boldsymbol{\mu}}_i^0}{|R_k|}$ and $v_k^0 = \sqrt{\frac{\sum_{i \in R_k} (\hat{\sigma}_i^0)^2}{|R_k|}}$.

measurements $(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m)$ where there are m genes and n observations. Further assume that the first m_0 and the last $m_1 = m - m_0$ hypotheses are from the null and alternative models, respectively. The ODP test statistic for gene i is

$$S_{\text{ODP}}(\mathbf{y}_i) = \frac{\sum_{k=m_0+1}^m g_k(\mathbf{y}_i)}{\sum_{k=1}^{m_0} f_k(\mathbf{y}_i)}, \quad (1)$$

where $g_k(\cdot)$ is the alternative likelihood and $f_k(\cdot)$ is the null likelihood for gene k . The numerator and denominator can be viewed as the cumulative likelihoods under the alternative and null models, respectively, across all hypotheses, where the likelihoods most related to gene i contribute the most to the above statistic. Storey (2007) shows that this test statistic maximizes the number of expected true positives (ETP) for a fixed number of expected false positives (EFP)—a quantity closely related to the FDR, in that $\text{FDR} \approx \frac{\text{EFP}}{\text{EFP} + \text{ETP}}$. Therefore, by leveraging information across different hypotheses, the ODP achieves the optimal significance ranking of test statistics. Moreover, the improvements in statistical power are sometimes substantial compared to the standard likelihood ratio test based on the NP lemma (Storey, 2007).

Evaluating equation (1) requires making assumptions on the data-generating process and the hypothesis status for each test. In this work, the alternative and null densities follow a Normal distribution with some mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$ and standard deviation σ , denoted by $(\boldsymbol{\mu}_i^1, \sigma_i^1)$ and $(\boldsymbol{\mu}_i^0, \sigma_i^0)$ for genes $i = 1, 2, \dots, m$, respectively. However, it is not known *a priori* which tests are from the alternative and null models. Instead an approximation to the true ODP statistic is estimated (Storey *et al.*, 2007), i.e.,

$\hat{S}_{\text{ODP}}(y_i) = \sum_{k=1}^m g_k(y_i; \frac{\mu_k^1, \sigma_k^1}{\sum_{k=1}^m f_k(y_i; \mu_k^0, \sigma_k^0)})$. Another complication is that the

theoretical null distribution of the ODP test statistic is unknown and so P -values cannot be analytically calculated. Therefore, a bootstrap procedure must be implemented to generate an empirical null distribution of the test statistics. This estimated ODP has been shown to provide similar power to the true ODP (Storey *et al.*, 2007). However, calculating the test statistics involves $2m^2$ computations and so it is computationally demanding for genomic datasets where m can range anywhere from 10^3 to 10^5 .

Woo *et al.* (2011) proposed a computationally efficient approximation to the ODP called the modular ODP (mODP). Similar to the estimated ODP, the mODP assumes that the data y_i are generated from a Normal distribution with parameters (μ_i^1, σ_i^1) and (μ_i^0, σ_i^0) for genes $i = 1, 2, \dots, m$ under the alternative and null models, respectively. These parameters are estimated from the data on a gene-by-gene basis using maximum likelihood, which can be calculated by a least squares fit in the Normal distribution case. A clustering algorithm then assigns genes to $k = 1, 2, \dots, K$ modules based on the symmetric Kullback–Leibler distance δ_{jk} (only the alternative model is used to determine gene-module assignments). Using the module assignments, the parameter estimates are updated and genes are reassigned to new modules; the above steps continue until a convergence criterion is met. The final module parameters are denoted by (c_k^1, v_k^1) and (c_k^0, v_k^0) under the alternative and null models, respectively (Algorithm 1).

Given the module parameters, the mODP test statistic can be expressed as

$$\hat{S}_{\text{mODP}}(y_i) = \frac{\sum_{k=1}^K g_k(y_i; c_k^1, v_k^1) |R_k|}{\sum_{k=1}^K f_k(y_i; c_k^0, v_k^0) |R_k|}, \quad (2)$$

where $g_k(\cdot)$ is the alternative likelihood and $f_k(\cdot)$ is the null likelihood under module k , and $|R_k|$ is the number of genes assigned to module k . A bootstrap algorithm is implemented to generate the empirical null distribution of the test statistics (described in Supplementary Material). The mODP reduces the number of calculations from $2m^2$ to $2Km$ where $K \ll m$. Thus, the time complexity of the mODP is linear with the number of genes. In the study by Woo *et al.* (2011), the authors demonstrate that the mODP has similar power to the estimated ODP and is robust to the number of modules when $K \geq 50$.

2.2 Regression splines

The general framework for modeling non-linear responses in complex study designs follows from the study by Storey (2005). Consider an experiment with gene expression measurements y_{ij} and explanatory variables x_j for $i = 1, 2, \dots, m$ genes and $j = 1, 2, \dots, n$ observations. In non-static (e.g., time course) studies, there can be multiple measurements of x_j for observation j , denoted by x_{jt} where $t = 1, 2, \dots, T_j$; for example, x_{jt} can be multiple time points or different dosage levels for a particular observation. The observed expression for gene i can be modeled as

$$y_{ijt} = \mu_i(x_{jt}) + \gamma_{ij} + \epsilon_{ijt}, \quad (3)$$

where $\mu_i(\cdot)$ is the population average curve, γ_{ij} is the individual-specific random deviation from the population average curve, and ϵ_{ijt} is a random error that follows a Normal distribution with mean zero and variance σ_i^2 . Here, we assume that the individual-specific random deviations follow a Normal distribution with mean zero and variance τ_i^2 .

The population average curve can be flexibly modeled using a regression spline. A regression spline is a piecewise polynomial function continuous at d specified points (or ‘knots’). We only consider natural cubic splines, which are third-order polynomial functions that are linear beyond the boundary knots. In this case, the population average curve can be parameterized by a d -dimensional basis:

$\mu_i(x) = \alpha_i \mathbf{1} + s(x) \beta_i$ where $\mathbf{1}$ is a vector of 1’s, $x = (x_{jt})$ is these of explanatory variables, $s(x) = (s_1(x), s_2(x), \dots, s_d(x))$ is a matrix of the explanatory variables evaluated on a prespecified d -dimensional basis, and the parameters α_i and $\beta_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{id})^T$ are estimated by least squares on a gene-by-gene basis (described in more detail in Supplementary Material). The parametric model for $\mu_i(x)$ enables testing of parameters α_i and β_i , which do not depend on specific x_{jt} . This simplification allows for inferences of general sampling designs (Storey *et al.*, 2005). We apply this framework to a static sampling study, two independent sampling non-static studies, and a longitudinal sampling study.

In a static sampling study, subjects are independently sampled across one or more biological groups with no functional (e.g., time) component. As an example, in the smoker study, x_j is a dichotomous variable taking values 0 or 1 indicating the smoking status of individual j and y_{ij} is the RNA-seq count for gene i (Fig. 1a). Equation (3) can be simplified by modeling the population average curve as $\mu_i(x) = \alpha_i \mathbf{1} + \beta_i x$ for the dichotomous variable $x = (x_1, x_2, \dots, x_n)^T$ where the t index is dropped because implicitly $T_j = 1$. In this study, we are interested in determining whether gene expression is differentially expressed between groups (the alternative hypothesis) or remains unchanged (the null hypothesis). Therefore, the null hypothesis model (dashed line) is fit under the constraint of $\beta_i = 0$ and the alternative hypothesis model (solid line) allows this parameter to be unconstrained.

In an independent sampling study, subjects are independently sampled across a continuous variable (similar to cross-sectional sampling). The observed population average curve can be expressed as $\mu_i(x) = \alpha_i \mathbf{1} + s(x) \beta_i$ for the continuous variable $x = (x_1, x_2, \dots, x_n)^T$ where $T_j = 1$. There are two studies analyzed with independent sampling designs. The first is a kidney aging study where human subjects are independently sampled at various ages (Fig. 1b). The second is a dose–response study where 17β -estradiol is introduced to breast cancer cells at various dosage levels (Fig. 1c). In both of these studies, the objective is to determine whether gene expression is differentially expressed across time or dosage level (the alternative hypothesis) or remains unchanged (the null hypothesis). Therefore, the null hypothesis model (dashed line) is fit under the constraint of $\beta_{il} = 0$ for $l = 1, 2, \dots, d$ and the alternative hypothesis model (solid line) allows these parameters to be unconstrained.

In a longitudinal sampling study, subjects are sampled multiple times across a continuous variable. Here, the observed population average curve can be expressed as $\mu_i(x) = \alpha_i \mathbf{1} + s(x) \beta_i$ for the continuous variable $x = (x_{jt})$ where there are $j = 1, 2, \dots, n$ observations and $t = 1, 2, \dots, T_j$ measurements of the j th observation. As an example, the endotoxin study compares two different classes across

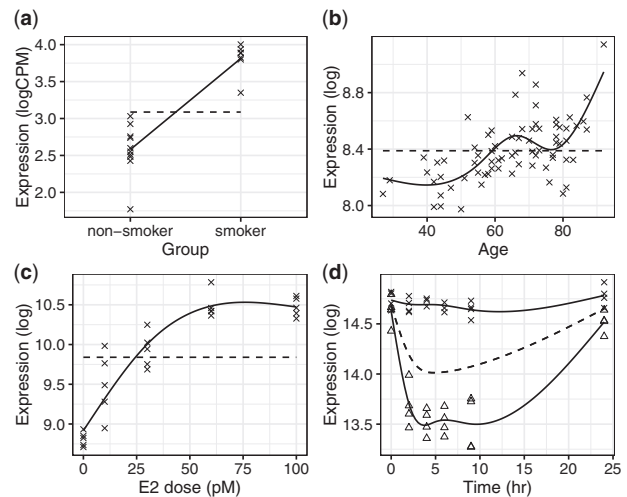


Fig. 1. Fitting regression splines to general study designs: (a) static, (b, c) independent sampling and (d) longitudinal studies. The null (dashed) and alternative (solid) models are shown for a significant gene. In (d), the endotoxin-treated and control groups are denoted by a triangle and cross, respectively

Algorithm 2: Choosing the dimensionality for the non-linear response model

Input: Dataset \mathbf{Y} , the maximum basis D , the total eigen-genes E , and explanatory variable \mathbf{x} .
Output: A \hat{d} -dimensional B -spline basis.
1 Apply singular value decomposition to \mathbf{Y} and extract $\mathbf{V} = (v_1, v_2, \dots, v_E)$ eigen-genes.
2 **forall** $i \in \{1, 2, \dots, E\}$ **do**
3 **forall** $d \in \{1, 2, \dots, D\}$ **do**
4 Determine d -dimensional basis to model population average curve $\mu_i(\mathbf{x}) = \alpha_i \mathbf{1} + \sum_{l=1}^d \beta_{il} s_l(\mathbf{x})$.
5 Regress v_i onto $\mu_i(\mathbf{x})$ to estimate parameters and calculate the cross validation error.
6 **end**
7 **end**
8 Select the \hat{d}_i that minimizes the cross validation error for eigen-genes $i = 1, 2, \dots, E$.
9 Select the maximum dimensionality across all eigen-genes, i.e., $\hat{d} = \max_i \{\hat{d}_i\}$.
10 Calculate the \hat{d} -dimensional B -spline basis: $\mathbf{s}(\mathbf{x}) = (s_1(\mathbf{x}), s_2(\mathbf{x}), \dots, s_{\hat{d}}(\mathbf{x}))$.

time, namely, endotoxin-treated versus control-treated. For this case, y_{ijt} is the gene expression measurement for gene i in individual j at time point t and x_{jt} indicates the time individual j was sampled (Fig. 1d). The alternative hypothesis is that there is differential expression between classes while the null hypothesis is that there is no difference in gene expression. Thus, the null hypothesis model fits one curve to both classes combined (dashed line) and the alternative hypothesis model fits two separate curves to each class (solid line).

Algorithm 2 summarizes the procedure for choosing the dimensionality of the regression spline. The natural cubic spline—or any polynomial spline—can be parameterized by a B -spline basis. The B -spline requires choosing the location of d knots to anchor the basis functions. We utilize the cross validation algorithm by Storey *et al.* (2005) to automatically choose the optimal d . First, we apply a singular value decomposition to extract the top $i = 1, 2, \dots, E$ eigen-genes (or right-singular vectors) which represent directions of maximal variation. The eigen-genes are used for choosing a reasonable dimensionality to model the non-linear gene expression responses. We then regress these eigen-genes onto $\mathbf{s}(\mathbf{x}) = (s_1(\mathbf{x}), s_2(\mathbf{x}), \dots, s_d(\mathbf{x}))$, where we use a $d = 1, 2, \dots, D$ dimensional B -spline bases. The knots are placed at evenly spaced quantiles, i.e., the $0, \frac{1}{d-1}, \frac{2}{d-1}, \dots, \frac{d-2}{d-1}, 1$ quantiles. Finally, the basis dimension used for model fitting is chosen by applying a cross-validation procedure to select the an estimated optimal \hat{d} across all eigen-genes. Using the selected \hat{d} , a least squares model fit is applied to estimate the population average curve and variance for all genes under the alternative and null models.

3 Materials and methods

Our algorithm introduces regression splines into the ODP framework to extend it to complex study designs. To do so, the ODP test statistic must be extended to incorporate non-linear responses. Suppose the expression values are y_{ijt} and the explanatory variables are x_{jt} where there are $i = 1, 2, \dots, m$ genes, $j = 1, 2, \dots, n$ observations and $t = 1, 2, \dots, T_j$ measurements of the j th observation. Consider two different models, namely, the null model with parameters $(\mu_i^0(\mathbf{x}), \sigma_i^0)$ and the alternative model with parameters $(\mu_i^1(\mathbf{x}), \sigma_i^1)$, where the null model is a restricted version of the alternative model (detailed in [Supplementary Material](#)). The objective is to test the null hypothesis $H_0 : \mu_i(\mathbf{x}) = \mu_i^0(\mathbf{x})$ versus the alternative hypothesis $H_1 : \mu_i(\mathbf{x}) = \mu_i^1(\mathbf{x})$. In this work, the population average curves are flexibly modeled using a regression spline. The

Algorithm 3: Determining differentially expressed genes from general study designs

Input: Dataset \mathbf{Y} , the alternative and null model, the number of bootstrap iterations B , and the number of modules K .
Output: Empirical p -values for each gene.
1 For non-linear responses, apply Algorithm 2 to choose the dimensionality of the model.
2 Estimate the population average curve $\hat{\mu}_i^0(\mathbf{x})$ under the null model using least squares.
3 Subtract the estimated null model from the observed data, $\mathbf{y}'_i = \mathbf{y}_i - \hat{\mu}_i^0(\mathbf{x})$, and then fit the alternative model $\mu_i^1(\mathbf{x})$ to \mathbf{y}'_i for $i = 1, 2, \dots, m$ genes.
4 Estimate the module parameters under the null and alternative models using Algorithm 1 and calculate $\hat{S}_{\text{mODP}}(\mathbf{y}'_i)$ for $i = 1, 2, \dots, m$ genes.
5 Simulate $b = 1, 2, \dots, B$ datasets from the null model by resampling the studentized residuals $e_i^{*(b)}$ from the alternative model fit and adding it to $\hat{\mu}_i^0(\mathbf{x})$, i.e., $\mathbf{y}_i^{*(b)} = \hat{\mu}_i^0(\mathbf{x}) + e_i^{*(b)}$.
6 Generate a set of null test statistics using the $b = 1, 2, \dots, B$ datasets.
7 Using the observed and null test statistics, calculate empirical p -values (see Supplemental Material).

parameters under both models can then be estimated by least squares, i.e., $(\hat{\mu}_i^1(\mathbf{x}), \hat{\sigma}_i^1)$ and $(\hat{\mu}_i^0(\mathbf{x}), \hat{\sigma}_i^0)$.

For non-linear responses, the estimated ODP statistic is

$$\hat{S}_{\text{ODP}}(\mathbf{y}_i) = \sum_{k=1}^m g_k(\mathbf{y}_i; \frac{\hat{\mu}_k^1(\mathbf{x}), \hat{\sigma}_k^1}{\sum_{k=1}^m f_k(\mathbf{y}_i; \hat{\mu}_k^0(\mathbf{x}), \hat{\sigma}_k^0)}) \quad (4)$$

where the likelihoods are assumed to follow a Normal distribution. It is evident that $\hat{\mu}_i^0(\mathbf{x})$ is not of interest in the testing procedure. This ancillary information can be removed by transforming the data to $\mathbf{y}'_i = \mathbf{y}_i - \hat{\mu}_i^0(\mathbf{x})$. Under this transformation, the hypotheses are $H_0 : \mu_i^1(\mathbf{x}) = \mathbf{0}$ versus $H_1 : \mu_i^1(\mathbf{x}) \neq \mathbf{0}$. This modified version of the estimated ODP statistic is

$$\hat{S}_{\text{ODP}}(\hat{\mathbf{y}}_i) = \sum_{k=1}^m g_k(\hat{\mathbf{y}}_i; \frac{\hat{\mu}_k^1(\mathbf{x}), \hat{\sigma}_k^1}{\sum_{k=1}^m f_k(\hat{\mathbf{y}}_i; \mathbf{0}, \hat{\sigma}_k^0)}) \quad (5)$$

Similar to the original implementation of the ODP, the above test statistic requires $2m^2$ calculations which makes it computationally slow for genomic datasets. Instead, we can utilize the mODP statistic according to

$$\hat{S}_{\text{mODP}}(\hat{\mathbf{y}}_i) = \frac{\sum_{k=1}^K g_k(\hat{\mathbf{y}}_i; c_k^1(\mathbf{x}), v_k^1) |R_k|}{\sum_{k=1}^K f_k(\hat{\mathbf{y}}_i; \mathbf{0}, v_k^0) |R_k|} \quad (6)$$

where there are $k = 1, 2, \dots, K$ modules, the membership size of module k is $|R_k|$ and the module parameters are estimated by applying the mODP clustering algorithm (described in Algorithm 1).

Our proposed method is summarized in Algorithm 3: The inputs are the observed dataset (\mathbf{x}, \mathbf{Y}) , the alternative and null models, the number of modules K and the number of bootstrap iterations B . First, we apply Algorithm 2 to choose the dimensionality d for the population average curves, i.e., $\mu_i(\mathbf{x}) = \alpha_i \mathbf{1} + \mathbf{s}(\mathbf{x})\beta_i$ where $\mathbf{s}(\mathbf{x}) = (s_1(\mathbf{x}), s_2(\mathbf{x}), \dots, s_d(\mathbf{x}))$. The data are then transformed by subtracting the null model fit from the observed dataset. This adjusted gene expression response variable \mathbf{y}'_i is regressed onto the alternative model to get updated parameter estimates of $(\mu_i^1(\mathbf{x}), \sigma_i^1)$.

Finally, we apply the mODP clustering algorithm to determine the parameters $(c_k^1(x), v_k^1)$ and $(0, v_k^0)$ for the $k = 1, 2, \dots, K$ modules under the alternative and null models, respectively (see Algorithm 1). Using the parameter estimates from the clustering algorithm, the mODP statistic is calculated for all genes. A bootstrap algorithm is implemented to calculate the empirical null distribution of the test statistics (described in [Supplementary Material](#)). For the datasets analyzed here, there are $B = 500$ bootstrap iterations ($B = 5000$ for the smoker study) and $K = 800$ modules.

Prior applications of ODP have focused on microarray studies where it is common to assume that the gene expression response variable is approximately Normal and homoscedastic (i.e., a single variance per gene). However, in sequencing studies, this assumption is no longer valid because the observations are heteroscedastic. To apply the ODP to sequencing studies, we implement a similar strategy detailed by [Law et al. \(2014\)](#) where RNA-seq data are log-transformed to model the observed mean–variance relationship. Using this model, weights capturing the heteroscedasticity across observations are estimated. These weights are then incorporated in a weighted least squares regression and are easily integrated into the mODP framework. Given a set of inverse variance weights w_{ijt} for $i = 1, 2, \dots, m$ genes, $j = 1, 2, \dots, n$ observations and $t = 1, 2, \dots, T_j$ measurements of the j th observation, the data are transformed as $\bar{y}_{ijt} = \sqrt{w_{ijt}}y_{ijt}$ and $\bar{s}_i(x_{it}) = \sqrt{w_{ijt}}s(x_{it})$. An ordinary least squares algorithm can then be applied to this transformed data. Thus, Algorithm 3 can be appropriately adjusted to accommodate these weights.

4 Results

The generalized mODP was applied to four different genomic experiments. The performance of mODP was compared to an F -test and a moderated F -test ([Smyth, 2004](#)) using the number of discoveries and enriched gene sets. Finally, we validated our findings through comprehensive simulations.

4.1 Datasets

Kidney study. To elucidate the transcriptional response from aging in the kidney, the kidney study collected cortex samples from 72 patients with ages ranging from 27 to 92 years ([Rodwell et al., 2004](#)). The samples were hybridized onto U133a and U133b GeneChips with 44 928 probes. Following similar filtering steps in [Storey \(2005\)](#) to control for potential confounding, only 38 833 probes were used for analysis and the expression values were log-transformed for variance stabilization.

Endotoxin study. The endotoxin study analyzed transcriptional regulation in human blood leukocytes from two experimental groups: a treatment group receiving a bacterial endotoxin (an inflammatory stimulus) and a control group ([Calvano et al., 2005](#)). There were four samples in each biological group and blood samples were collected at 2, 4, 6, 9 and 24 h intervals. One control sample had missing information at the 4 and 6 h time points. The samples were hybridized onto U133 GeneChips with 44 924 probes. The expression values were log-transformed for variance stabilization.

Smoker study. The smoker study is a two group comparison between smoking and non-smoking humans ([Ryan et al., 2014](#)). There are a total of 17 samples (10 non-smokers and 7 smokers) from human airway basal cells in the epithelium: there is one female smoker and the rest of the samples are males. The samples are sequenced (paired-end) using Illumina HiSeq 2000, and the reads are assembled using Bowtie: there are total of 65 217 genes with mapped reads. After filtering genes with fewer than 10 reads across all samples, only 26 268 genes remained for analysis. The R package `limma` is used to estimate the inverse-variance weights for the weighted least squares implementation. The expression values were transformed to log₂-counts per million (logCPM).

Dose study. The dose study is a dose–response experiment where sensitivity to 17 β -estradiol (E2) in breast cancer cells (BUS cells) was examined ([Cosier et al., 2003](#)). There are five biological replicates for each E2 concentration, where the E2 concentrations considered

were 0, 10, 30, 60 and 100 pM (25 total samples). After 48 h exposed to E2, RNA samples were hybridized onto U133a GeneChips with 22 283 probes. The expression values were log-transformed to stabilize the variance.

4.2 Determining the degrees of freedom

We implemented the cross-validation procedure detailed in Algorithm 2 to determine the appropriate dimensionality d for the B -spline basis. (The smoker study is a two-group comparison and so regression splines are not necessary.) For each study, the first four eigen-genes were determined by applying a singular value decomposition to the dataset. In the endotoxin study, the control-treated and endotoxin-treated groups were separated into two distinct datasets. Multiple regressions were fit to the eigen-genes using $d = 1, 2, 3, 4$ for the endotoxin and dose studies and $d = 1, 2, \dots, 10$ for the kidney study (an intercept term was included in the models). For each eigen-gene, the d that minimized the leave-one-out cross-validation error was selected. Finally, the maximum d across all eigen-genes was chosen as the estimated degrees of freedom. Applying the above procedure, we find $\hat{d} = 4$ for the endotoxin and kidney studies and $\hat{d} = 2$ for the dose study ([Fig. 2](#)).

4.3 Method comparisons

We compared the mODP to two other popular test statistics, namely, the F -statistic and the moderated F -statistic (described in [Supplementary Material](#)). Compared to the F -statistic, the moderated F -statistic shrinks the sample variance toward a pooled variance. This shrinkage allows for more stable inferences with low-sample size studies ([Smyth, 2004](#)). Unlike the mODP which requires an empirical null distribution, the F -statistic and moderated F -statistic have theoretical null distributions. Therefore, we also estimate an empirical null distribution for the F -test and moderated F -test using a bootstrap algorithm (described in [Supplementary Material](#)). In summary, the mODP is compared to an F -test, a moderated F -test, a bootstrap F -test and a bootstrap moderated F -test.

We applied the above testing procedures to our four chosen studies and calculated the number of differentially expressed genes at various FDRs ([Fig. 3](#)). At each FDR threshold, the mODP finds substantially more differentially expressed genes compared to the other methods. For example, when applying a FDR of 0.1, mODP detects 1481, 297, 6637 and 887 more differentially expressed genes in the kidney, smoker, endotoxin and dose studies, respectively. In addition, the mODP finds nearly all of the differentially expressed genes detected by the other methods ([Fig. 4](#)). Finally, we find that the mODP has the lowest estimated proportion of true nulls across all studies ([Supplementary Table S1](#)). This suggests that the mODP estimates a higher expected number of alternative genes.

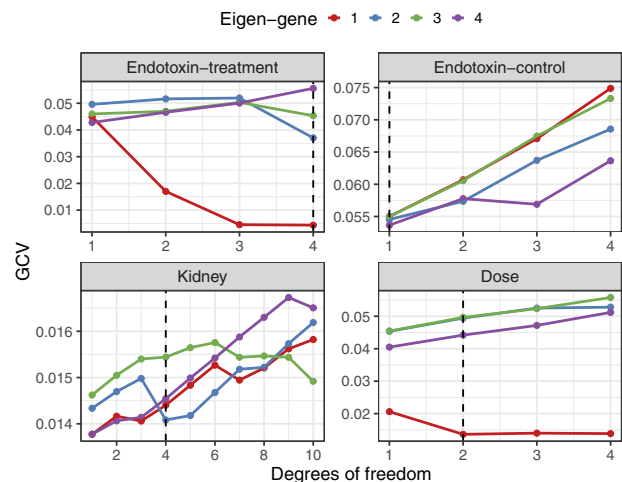


Fig. 2. Cross-validation error versus degrees of freedom for the first four eigen-genes. The dotted line indicates the chosen degrees of the freedom in the study

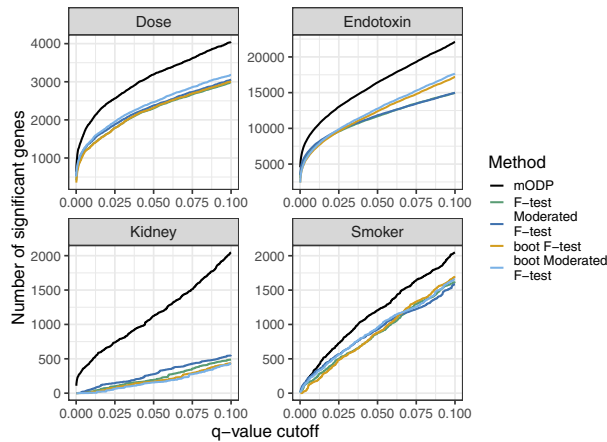


Fig. 3. Observed number of discoveries at various q -value cutoffs from the mODP (black), F -test (green), bootstrap F -test (orange), moderated F -test (blue) and bootstrap moderated F -test (light blue). These methods were applied to four different studies: endotoxin (longitudinal sampling), kidney (independent sampling), dose (independent sampling) and smoker (static sampling)

To compare the testing procedures, we also performed an enrichment analysis using the hallmark gene sets from the MSigDB database. These gene sets contain highly curated genes with clear expression for well-defined biological states or processes (Liberzon *et al.*, 2015, 2011). We developed a simple procedure to detect important gene sets by assigning the proportion of true positives to each. These values range from 0 to 1, with important gene sets having largest values (see Supplementary Material for additional details). We find that the mODP has the largest proportion of true positives across all gene sets compared to other methods (Fig. 5). Thus the mODP has more power to detect gene sets with enriched true positives.

4.4 Simulations

Comprehensive simulations were performed to verify the observed differences between mODP and other methods. We generated 500 representative datasets of the observed studies as follows. For each study, the F -testing procedure was used to separate genes into two distinct classes (alternative and null) based on a FDR threshold of 0.1. We then sampled from the population of alternative genes to get unique gene expression profiles. In total, we considered 5, 10, 50, 100 and 200 unique gene expression curves in our simulation studies. These curves defined the signal for the alternative genes. (Note the smoker study is a static experiment and so ‘unique gene expression profile’ refers to the mean differences between the two conditions.) The signals for the null genes were sampled from the null population. Random noise was added to maintain the signal-to-noise ratio and to match the observed power. Finally, the total number of alternative and null genes was chosen to keep the observed proportion of true nulls fixed. For more details, see Supplementary Material.

To compare the testing procedures in the simulated datasets, we calculated the estimated FDR and the total number of discoveries. We find that the mODP controls the FDR in all simulated studies (Fig. 6b). Furthermore, substantially more differentially expressed genes were detected relative to the other testing procedures (Fig. 6a). We also find that the moderated F -test identifies a similar number of differentially expressed genes compared to the F -test. This is unsurprising as the moderated F -test only outperforms the F -test when there are small sample sizes. When the number of unique gene expression patterns is increased, the power of the mODP decreases while the F -test and moderated F -test remain unchanged. This occurs because the number of unique gene expression curves does not change the power of the F -test and moderated F -test.

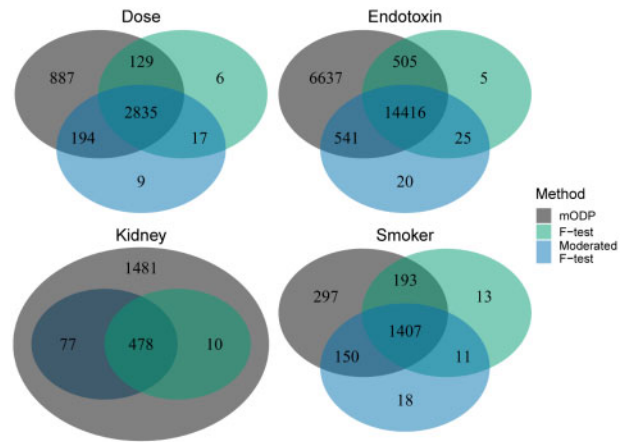


Fig. 4. Venn diagram of the total discoveries at a FDR of 0.1 for the studies in Figure 3. Only the mODP (black), F -test (green) and moderated F -test (blue) are shown

4.5 Implementation in the edge package

The mODP described above is implemented in an R package called *edge*. When applying the *edge* package to a dataset, the first step is to formulate an alternative and null hypothesis to test. Our software has a function called `build_study` to help users formulate hypotheses for static, longitudinal or independent sampling designs. Once the two hypotheses are specified, the next step is to input the design matrices into the `odp` function to apply the mODP procedure outlined here. (Note that the clustering and bootstrap algorithms are executed internally with default parameters that can be changed.) The `odp` function returns significance results such as P -values, local FDRs, q -values, and the proportion of truly null tests. These results can then be visualized using the `plot` and `hist` functions in a statistical analysis.

There are many other features included in our software: (i) the gene-module assignments are available for additional analyses (e.g. enrichment analysis) using the `kl_clust` function, (ii) the model fits under the alternative and null hypotheses can be accessed using the `fit_models` function and (iii) integrations of other popular tools such as surrogate variable analysis (SVA) (Leek and Storey, 2007, 2008) and jackstraw (Chung and Storey, 2015) are available for more comprehensive biological analyses. Additional information can be found in the vignette at <http://bioconductor.org/packages/release/bioc/html/edge.html>.

5 Discussion

The ODP is a test statistic that provides substantial improvements in statistical power compared to other testing procedures. While previous work on the ODP is limited to static microarray studies (Storey *et al.*, 2007; Storey, 2007; Woo *et al.*, 2011), here, we extend its application to complex experimental designs and sequencing studies. Our proposed algorithm is applied to two time-series studies, a dose–response study and an RNA-seq study. For each study, our method detects more differentially expressed genes and improves the statistical power for gene set enrichment analysis. These improvements in power are validated through comprehensive simulations, where data are simulated to closely resemble the observed datasets. Therefore, the ODP allows for a more thorough investigation of underlying biological mechanisms in downstream analysis.

The gained improvements in power from the ODP have important biological implications. In a genome-wide gene expression study, genes are commonly co-expressed and share similar patterns of gene expression. The ODP leverages this information across genes to strengthen the evidence for or against differential expression. To explore how the ODP depends on the ‘degree’ of co-expression, we varied the number of unique gene expression profiles with simulated data. We find that the ODP loses power as the number of unique

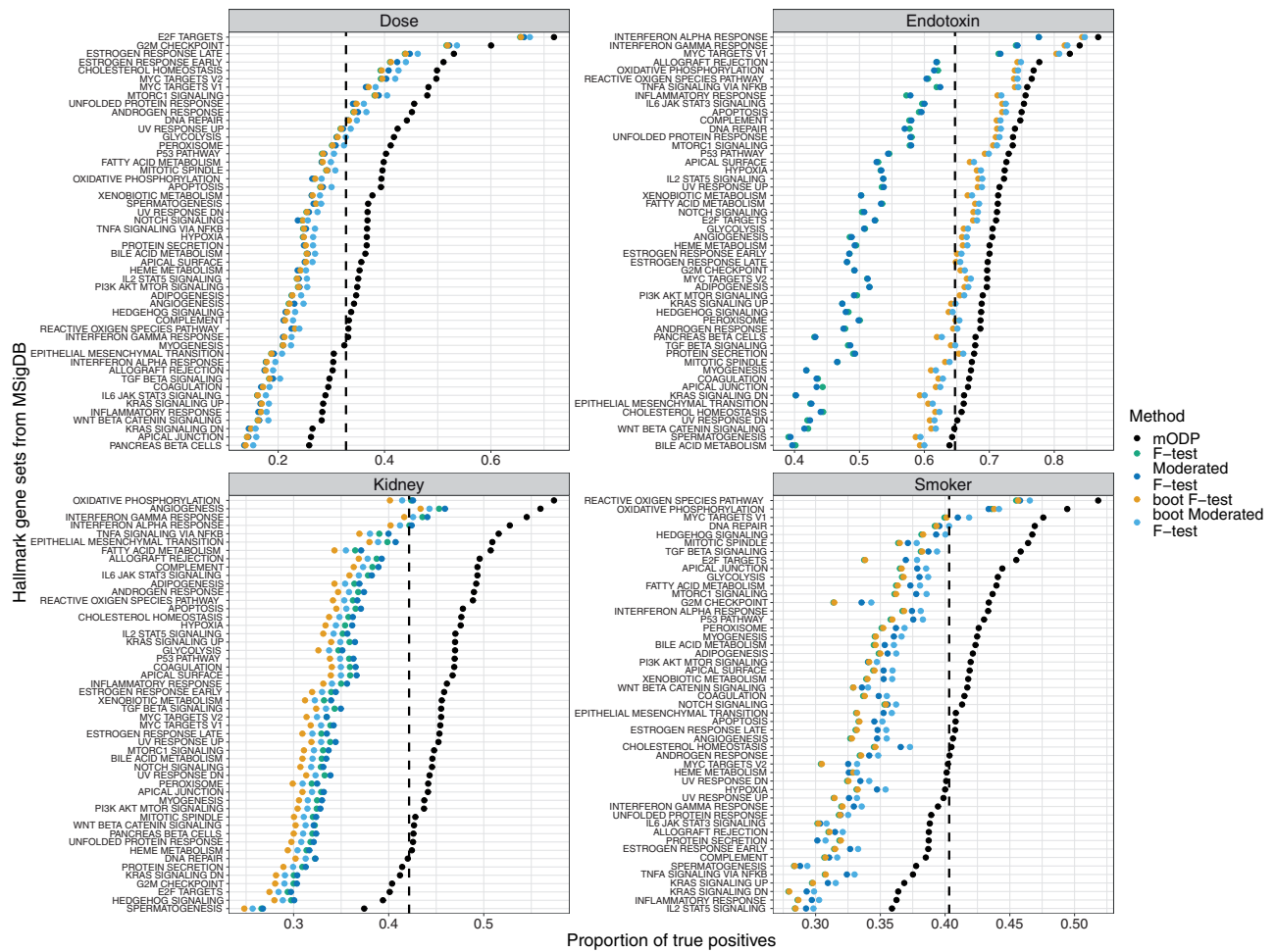


Fig. 5. Proportion of true positives for the Hallmark gene sets from MSigDB. Enrichment results for each method are shown for the dose, endotoxin, kidney and smoker studies. The dashed line indicates the proportion of true positives in the dataset using mODP

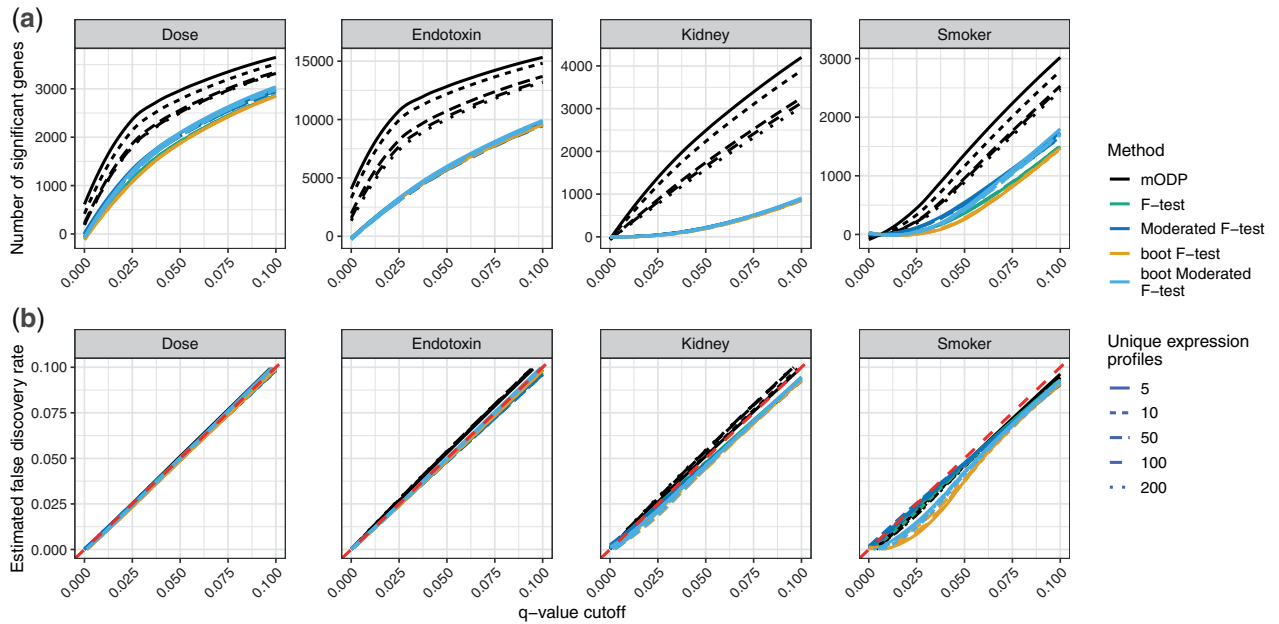


Fig. 6. Simulation results for the dose, endotoxin, kidney and smoker studies using 5, 10, 50, 100 and 200 unique gene expression profiles (linetype). The mODP (black), *F*-test (green), moderated *F*-test (blue), bootstrap *F*-test (orange) and bootstrap moderated *F*-test (light blue) are applied to the simulated studies. (a) Estimated power at multiple *q*-value cutoffs between 0.0001 and 0.1. (b) Estimated FDR. Curves represent the average value from 500 replications

gene expression profiles increases; there are fewer related genes and so there is less information that can be leveraged in the test statistic. As an extreme case, suppose genes are not co-expressed and every gene follows a unique expression pattern. In this scenario, the ODP has been shown to perform similar to the *F*-test (Storey *et al.*, 2007). While this example is unlikely in biological studies, it provides intuition for the observed power improvements compared to other testing procedures.

There are a few considerations to note when applying our extended framework to genomic datasets. First, the computationally efficient implementation of the ODP, called the modular ODP (mODP), involves specifying the number of modules. While previous work has recommended 50 modules for static microarray studies, we found choosing at least 200 modules to capture the complex relationships among genes provides the best results. Second, there needs to be an adequate number of observations in the study. This is due to the constraints of the mODP: it requires accurate estimates of the mean and variance. Furthermore, the bootstrap algorithm implemented in the procedure requires a minimum number of observations per biological condition to generate a valid empirical null. In the studies considered here, there are at least four biological replicates per condition. Finally, the appropriate degrees of freedom need to be carefully chosen to avoid overfitting the spline. To this end, we implemented a procedure from the study by Storey *et al.* (2005) that chooses the degrees of freedom based on the leave-one-out cross validation algorithm.

An interesting aspect of the mODP implementation is that a clustering algorithm assigns genes to modules, where the modules are representative of shared gene expression patterns. These modules provide valuable information that can be utilized in an exploratory data analysis. For example, we can calculate the proportion of true positives for each module and rank modules based on true-positive enrichment. Modules enriched with true positives can then be further analyzed to understand functional relationships among genes. Thus the clustering algorithm provides information of potential use in other biological analyses.

There are a number of ways the ODP can be further extended for genomic studies. One enhancement is incorporating prior weights on each hypothesis test. For example in sequencing data, higher per-gene read counts are more reliable than lower per-gene read counts. This information can be included into a weighted ODP, where weights are generated by estimating the functional proportion of true nulls based on some informative variable (Chen *et al.*, 2017). Another enhancement is to extend the ODP to generalized linear models where the response variable follows an exponential family distribution.

As the cost of generating biological samples decreases, the prevalence of complex study designs will increase. The key motivation behind these studies is to capture inherently non-linear transcriptional responses. Therefore, there is demand for statistically rigorous methodologies that can be applied to such settings. In this work, we develop a framework to model non-linear gene expression responses while optimally utilizing biological correlations among genes to improve statistical power. Our method can thus uncover important biological insights across a wide range of applications in functional, translational and clinical genomics.

6 Software and data

An implementation of the algorithm described in this article is available as an R package called *edge*. The package can be downloaded at <https://github.com/StoreyLab/edge> (most recent version) or

<https://bioconductor.org/packages/release/bioc/html/edge.html> (release). The data and code used to produce the figures in this manuscript can be found at https://github.com/StoreyLab/odp_general_studies.

Funding

This work was supported by NIH grants R01 HG002913, R01 HG006448, and 5T32HG003284.

Conflict of Interest: none declared.

References

- Calvano, S.E. *et al.*; Inflammation, and to Injury Large Scale Collaborative Research Program, H. R. (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1032–1037.
- Chen, X. *et al.* (2017) The functional false discovery rate with applications to genomics. *Biostatistics*. doi:10.1093/biostatistics/kxz010.
- Chung, N.C. and Storey, J.D. (2015) Statistical significance of variables driving systematic variation. *Bioinformatics*, **31**, 545–554.
- Coser, K.R. *et al.* (2003) Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in mcf7/bus breast cancer cells by DNA microarray. *Proc. Natl. Acad. Sci. USA*, **100**, 13994–13999.
- 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.
- Law, C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
- Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.
- Leek, J.T. and Storey, J.D. (2008) A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA*, **105**, 18718–18723.
- Liberzon, A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, **27**, 1739–1740.
- Liberzon, A. *et al.* (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
- Lönnstedt, I. and Speed, T. (2002) Replicated microarray data. *Stat. Sin.*, **12**, 31–46.
- Neyman, J. and Pearson, E.S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Charact.*, **231**, 289–337.
- Rodwell, G.E.J. *et al.* (2004) A transcriptional profile of aging in the human kidney. *PLoS Biol.*, **2**, e427.
- Ryan, D.M. *et al.* (2014) Smoking dysregulates the human airway basal cell transcriptome at COPD risk locus 19q13.2. *PLoS One*, **9**, e88051; e88051–e88051.
- Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 1–25.
- Storey, J.D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **69**, 347–368.
- 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-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, **8**, 414–432.
- Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
- Woo, S. *et al.* (2011) A computationally efficient modular optimal discovery procedure. *Bioinformatics*, **27**, 509–515.