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

March 8, 2019

Abstract

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. We then apply our extended framework to a static RNA sequencing study, a longitudinal and an independent sampling time-series study, and an independent sampling dose-response study. We find that 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 superior significance analysis of genomic studies. The algorithm is implemented in our freely available R package called `edge`.

*Corresponding author: `jstorey@princeton.edu`

Contents

1 Introduction

In genomic studies, gene expression measurements for thousands of genes are obtained simultaneously 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), i.e., the 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 testing procedure to rank genes is 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, i.e., no other testing procedure can achieve more power at a fixed significance threshold [1]. However, the LRT statistic may not provide the optimal ranking for multiple hypotheses [2]. 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 by utilizing 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 [2]. 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 functionally related genes that follow similar patterns of expression. This information is incorporated into the test statistic to strengthen or weaken the evidence for differential expression. In Storey et al. (2006), an approximation to the ODP performs favorably on DNA microarray studies compared to SAM [4], shrunken t-test or F-test [5, 6], Bayesian local FDR [7], and posterior probabilities [8].

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., hetereoscedasticity) [9]. 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. In order to incorporate dynamical responses commonly found in non-static studies, we utilize the regression spline methodology from Storey et al. (2005). A benefit of this approach is that it flexibly models gene expression responses within the ordinary 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 ref. [9] to estimate the per-gene hetereoscedasticity 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 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 paper 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 testing algorithm for a *single* test is provided by the Neyman-Pearson (NP) lemma [1]. Given some observed data $\mathbf{y} = (y_1, y_2, ..., y_n)$, 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 decision rule. Intuitively, this optimality arises because the data generating process under each model is assumed to be known. For multiple hypotheses, the likelihood ratio is applied on a case-by-case basis. However, potentially useful information across different hypotheses are ignored. As a consequence, the likelihood ratio may no longer be an optimal decision rule [2].

The 'optimal discovery procedure' (ODP) is a generalization of the NP lemma for multiple hypotheses. More specifically, consider gene expression measurements $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)$ where there are m genes and n samples. Further assume that the first m_0 and the last $m_1 = m - m_0$ hypotheses are from the alternative and null models, respectively. The ODP test statistic for gene i is

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

where $g_k(\mathbf{y}_i)$ is alternative likelihood and $f_k(\mathbf{y}_i)$ is the null likelihood under gene k. The numerator and denominator can be viewed as the cumulative power under the alternative and null models, respectively, across all hypotheses: only tests related to gene i contribute 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 false discovery rate, i.e., FDR $\approx \frac{\text{EFP}}{\text{EFP+ETP}}$. Therefore, by leveraging information across different hypotheses, the ODP achieves the optimal *ordering* (or ranking) of test statistics. Moreover, the improvements in statistical power are generally substantial compared to the likelihood ratio test (see [2]).

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 finite mean and variance. 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 [3], i.e. $\hat{S}_{\text{ODP}}(\mathbf{y}_i) = \frac{\sum_{k=1}^{m} g_k(\mathbf{y}_i)}{\sum_{k=1}^{m} f_k(\mathbf{y}_i)}$. Another complication is that the ODP test statistic does not have a theoretical null distribution 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 [3]. However, calculating the test statistics involves $2m^2$ computations and so it is computationally demanding for genomic datasets where m can

Algorithm 1: KL 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.
Estimate $(\hat{\mu}_i^1, \hat{\sigma}_i^1)$ and $(\hat{\mu}_i^0, \hat{\sigma}_i^0)$ under the alternative and null models, respectively, for genes $i = 1, 2, ..., m$.
Initialize the mean and variance for modules $k = 1, 2, ...K$ by randomly selecting $i(k)$ genes, i.e. $\bar{\mu}_k = \hat{\mu}^1_{i(k)}$ and $\bar{\sigma}_k = \hat{\sigma}^1_{i(k)}$. Set $\bar{\mu}_{\text{prev}} = 0$.
while $|\bar{\mu} - \bar{\mu}_{\text{prev}}| > \epsilon$ do
    $\bar{\mu}_{\text{prev}} \leftarrow \bar{\mu}$
    forall $i \in \{1, 2, ..., n\}$ do
        forall $k \in \{1, 2, ..., K\}$ do
            $d_{ik} \leftarrow \frac{1}{2}(\bar{\mu}_k^1 - \hat{\mu}_i^1)^{\mathsf{T}}(\bar{\mu}_k^1 - \hat{\mu}_i^1)\left(\frac{1}{(\bar{\sigma}_k^1)^2} + \frac{1}{(\hat{\sigma}_i^1)^2}\right) + \frac{n}{2}\left(\frac{(\bar{\sigma}_k^1)^2}{(\hat{\sigma}_i^1)^2} + \frac{(\hat{\sigma}_i^1)^2}{(\bar{\sigma}_k^1)^2}\right) - n$
        end
    end
    forall $k \in \{1, 2, ..., K\}$ do
        $R_k \leftarrow \left\{i | \forall i \in \{1, 2, ..., n\} : k = \text{argmin}_{1 \leq j \leq K} d_{ij}\right\}$
        $\bar{\mu}_k \leftarrow \frac{\sum_{j \in R_k} \hat{\mu}_j^1}{|R_k|}$
        $\bar{\sigma}_k \leftarrow \frac{\sum_{j \in R_k} \hat{\sigma}_j^1}{|R_k|}$
    end
end
For modules $k = 1, 2, ..., K$, the parameters under the alternative model are $\bar{\mu}_k^1 = \bar{\mu}_k$ and $\bar{\sigma}_k^1 = \bar{\sigma}_k$ and the parameters under the null model are $\bar{\mu}_k^0 = \frac{\sum_{j \in R_k} \hat{\mu}_j^0}{|R_k|}$ and $\bar{\sigma}_k^0 = \sqrt{\frac{\sum_{j \in R_k} (\hat{\sigma}_j^0)^2}{|R_k|}}$.
```

range anywhere from 10^3 to 10^5.

Woo et al. (2010) proposed a computationally efficient approximation to the ODP called the 'modular optimal discovery procedure' (mODP). Similar to the estimated ODP, the mODP assumes that the data are generated from a Normal density with parameters (μ_i^1, σ_i^1) and (μ_i^0, σ_i^0) for genes $i = 1, 2, ..., n$ under the alternative and null models, respectively. These parameters are estimated from the data using an ordinary least squares algorithm. A clustering algorithm then assigns genes to $k = 1, 2, ..., K$ modules based on the Kullback-Leibler distance d_{ik} (only the alternative model is used to determine gene-module assignments). Using the module assignments, the parameters are updated and genes are reassigned to new modules: the above steps continue until a convergence criteria is met. The final module parameters are denoted by $(\bar{\mu}_k^1, \bar{\sigma}_k^1)$ and $(\bar{\mu}_k^0, \bar{\sigma}_k^0)$ under the alternative and null models, respectively (see Algorithm 1).

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

$$\hat{S}_{\text{mODP}}(\mathbf{y}_i) = \frac{\sum_{k=1}^{K} g_k(\mathbf{y}_i; \bar{\mu}_k^1, \bar{\sigma}_k^1)|R_k|}{\sum_{k=1}^{K} f_k(\mathbf{y}_i; \bar{\mu}_k^0, \bar{\sigma}_k^0)|R_k|}, \tag{2}$$

where $g_k(\mathbf{y}_i; \bar{\mu}_k^1, \bar{\sigma}_k^1)$ is the alternative likelihood and $f_k(\mathbf{y}_i; \bar{\mu}_k^0, \bar{\sigma}_k^0)$ is the null likelihood under module k, and $|R_k|$ is the number of genes belonging to module k. A bootstrap algorithm is implemented to generate the empirical null distribution of the test statistics (described in Appendix 6.2). 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 Woo et al. (2010), 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 Storey (2005): Consider an experiment with gene expression measurements y_{ij} and explanatory variable x_j for $i = 1, 2, ..., M$ genes and $j = 1, 2, ..., N$ samples. In non-static studies, there can be multiple measurements of x_j for sample j, i.e., x_{jk} where $k = 1, 2, ..., T_j$; for example, x_{jk} can be multiple time points or different dosage levels for a particular sample. The expression for gene i is modeled as

$$y_{ij} = \mu_i(x_{jk}) + \gamma_{ij} + \epsilon_{ijk}, \tag{3}$$

where $\mu_i(x)$ is the population average curve, γ_{ij} is the individual-specific random deviation from the population average curve, and ϵ_{ijk} is a random error that follows a Normal distribution with mean 0 and variance σ_i^2. Here we assume that the individual-specific random effects follow a Normal distribution with mean 0 and constant 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, i.e. $\mu_i(x) = \alpha_i + \beta_i^\top \mathbf{s}(x)$ where $\mathbf{s}(x) = (s_1(x), s_2(x), ..., s_d(x))$ is a prespecified d-dimensional natural cubic spline basis and the parameters α_i and $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{id})$ are estimated by ordinary least squares. The parametric model for $\mu_i(x)$ enables testing of parameters α_i and β_i, which do not depend on specific x: this simplification allows for inferences of general sampling designs [10]. We apply this framework to a static sampling study, two independent sampling studies, and a longitudinal sampling study.

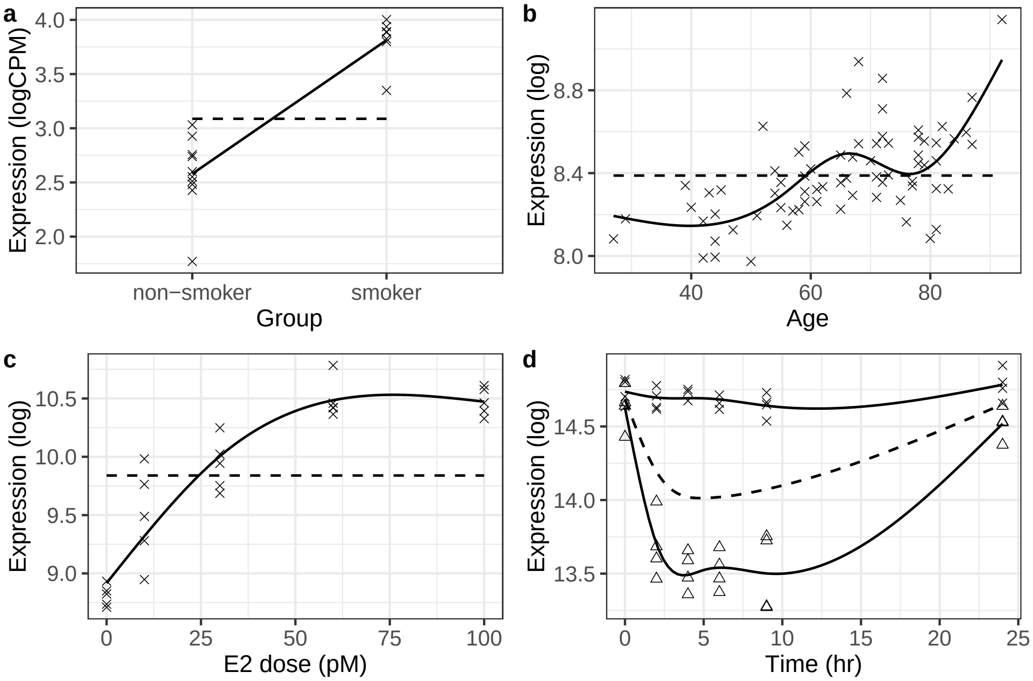

Figure 1: Fitting regression splines to general study designs: (a) static, (b-c) independent sampling, and (d) longitudinal studies. The null (dashed) and alternative (solid) models are shown for a significant gene. In (d), the endotoxin-treated and control groups are denoted by a triangle and cross,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

In a static sampling study, subjects are independently sampled across one or more biological groups at a fixed time point. As an example, in the smoker study x_j is a categorical variable indicating the smoking status of individual j and y_{ij} is the RNA-Seq count for gene i (Figure 1a). Equation (3) can be simplified by modeling the population average curve as $\mu_i(x_j) = \alpha_i + \beta_i x_j$. 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 population average curve is modeled using a d-dimensional natural cubic spline basis. 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 (Figure 1b). The second is a dose-response study where 17β-estradiol is introduced to breast cancer cells at various dosage levels (Figure 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

Algorithm 2: Algorithm for modeling non-linear responses

```
Input: Dataset $\mathbf{Y}$, the maximum basis $D$, the total eigen-genes $L$, the explanatory variable $\mathbf{x}$,
       and the alternative and null models.
Output: Estimated population average curve and standard deviation for each gene.
Apply singular value decomposition to $\mathbf{Y}$ and extract $\mathbf{V} = (v_1, v_2, ..., v_L)$ eigen-genes.
forall $i \in \{1, 2, ..., L\}$ do
    forall $j \in \{1, 2, ..., D\}$ do
        Determine $j$-dimensional basis to model population average curve
          $\mu(x) = \beta_0 + \sum_{l=1}^{j} \beta_l s_l(x)$.
        Regress $v_i$ onto $\mu(x)$ to estimate parameters and calculate the cross validation error.
    end
end
For eigen-genes $i = 1, 2, ..., L$, determine $\hat{d}_i$ that minimizes the cross validation error.
Select $\hat{d} = \max\{\hat{d}_i\}$ for $i = 1, 2, ..., L$.
Calculate the $\hat{d}$-dimensional natural cubic spline basis: $\mathbf{s}(x) = [s_1(x), s_2(x), ..., s_{\hat{d}}(x)]^\mathsf{T}$.
Apply ordinary least squares algorithm to estimate the population average curve $\hat{\mu}_i^0(x)$ and
  $\hat{\mu}_i^1(x)$ and the standard deviation $\hat{\sigma}_i^0$ and $\hat{\sigma}_i^1$ for $i = 1, 2, ..., m$ genes under the null and
  alternative models, respectively.
```

the constraint of $\beta_{il} = 0$ for $l = 1, 2, ..., d$ 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. The response variable can be modeled using equation (3) where the population average curve is a d-dimensional natural cubic spline basis. As an example, the endotoxin study compares two different classes across time, namely, endotoxin-treated versus control-treated. For this case, y_{ij} is the gene expression measurement for gene i in individual j and x_{jk} indicates the time point individuals were sampled (Figure 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 model fitting procedure to estimate the population average curve in a study. While natural cubic splines provide flexible parametric models, they require the placement of d knots. We utilize the cross validation algorithm in ref. [10] to automatically choose the optimal $\hat{d}$. First, we apply a singular value decomposition to extract the top $i = 1, 2, ..., L$ eigen-genes. We then regress these eigen-genes onto $\mathbf{s}(x) = (s_1(x), s_2(x), ..., s_j(x))$, where we use a $j = 1, 2, ..., D$ dimensional natural cubic spline basis: the knots are placed at evenly spaced quantiles, i.e. the $0, \frac{1}{j-1}, \frac{2}{j-1}, ..., 1$ quantiles. Finally, the basis dimension used for model fitting is chosen by applying a cross validation procedure to select the optimal $\hat{d}$ across all eigen-genes. Using the selected $\hat{d}$, an ordinary least squares algorithm is applied to estimate the population average curve and variance for all genes under the alternative and null models.

3 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 data are $\mathbf{y}_i$ and the explanatory variable is x_j where there are $i = 1, 2, ..., m$ genes and $j = 1, 2, ..., n$ samples. In either case, there are two different models, namely, the null model with parameters $(\mu_i^0(x), \sigma_i^0)$ and the alternative model with parameters $(\mu_i^1(x), \sigma_i^1)$. The objective is to test the null hypothesis $H_0 : \mu_i^1(x) = \mu_i^0(x)$ versus the alternative hypothesis $H_1 : \mu_i^1(x) \neq \mu_i^0(x)$. In this work, the population average curves are flexibly modeled using a d-dimensional natural cubic spline basis. The parameters under both models can then be estimated by ordinary least squares, i.e., $(\hat{\mu}_i^1(x), \hat{\sigma}_i^1)$ and $(\hat{\mu}_i^0(x), \hat{\sigma}_i^0)$.

For non-linear responses, the estimated optimal discovery procedure is

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

where the likelihoods are assumed to follow a Normal distribution. It is evident that $\hat{\mu}_0(x)$ is not of interest in the testing procedure (so-called 'nuisance parameter'): this ancillary information can be removed by transforming the data to $\mathbf{y}_i^* = \mathbf{y}_i - \hat{\mu}_i^0(x)$. Under this transformation, the hypotheses are $H_0 : \mu_i^{1*}(x) = \mathbf{0}$ versus $H_1 : \mu_i^{1*}(x) \neq \mathbf{0}$. This modified version of ODP is

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

Similar to the original implementation of the ODP, the above test statistic requires $2m^2$ calculations which makes it computationally slow for genomic data sets. Instead, we can incorporate the population average curve into the mODP as

$$\hat{S}_{\text{mODP}}(\mathbf{y}_i^*) = \frac{\sum_{k=1}^K g_k(\mathbf{y}_i^*; \bar{\mu}_k^1(x), \bar{\sigma}_k^1)|R_k|}{\sum_{k=1}^K f_k(\mathbf{y}_i^*; \mathbf{0}, \bar{\sigma}_k^0)|R_k|}, \tag{6}$$

where there are $k = 1, 2, ..., K$ 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).

```
Algorithm 3: Algorithm for analyzing complex study designs
  Input: Dataset 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 Apply Algorithm 2 to estimate the population average curves and standard deviations.
2 Subtract the null model from the observed data, $\mathbf{y}_i^* = \mathbf{y}_i - \hat{\mu}_i^0(x)$, and estimate $\hat{\mu}_i^{1*}(x)$ and
    $\hat{\sigma}_i^{1*}(x)$ for $i = 1, 2, ..., m$ genes
3 Estimate the module parameters under the null and alternative models using Algorithm 1 and
    calculate $\hat{S}_{\text{mODP}}(\mathbf{y}_i^*)$ for $i = 1, 2, ..., m$ genes.
4 Generate $k = 1, 2, ..., b$ datasets from the null model by resampling the scaled residuals $\mathbf{e}_{i,j}^*$ from
    the alternative model fit and add it to $\mathbf{y}_{i,j,k}^* = \hat{\mu}_i^0(x) + \mathbf{e}_{i,j}^*$ (described in Appendix 6.2).
5 Apply (2-3) to generate the null test statistics $\hat{S}_{\text{mODP}}^*(\mathbf{y}_{i,k}^*)$ for $k = 1, 2, ...b$ bootstrap iterations
    and $i = 1, 2, ..., m$ genes.
6 Using the observed and null test statistics, calculate empirical $p$-values.
```

Our proposed method is summarized in Algorithm 3: The inputs are the observed dataset $\mathbf{Y}$, the alternative and null models, the number of modules K and the number of bootstrap iterations b. First, we apply Algorithm 2 to explanatory variable $\mathbf{x}$ to model the non-linear gene expression responses. The data are then transformed by subtracting the null model fit from the observed dataset. This adjusted dataset $\mathbf{y}_i^*$ is regressed onto the alternative model to get updated parameter estimates, $(\hat{\mu}_i^{1*}(x), \hat{\sigma}_i^1)$. Finally, we apply the mODP clustering algorithm to determine the parameters for the $k = 1, 2, ..., K$ modules under the alternative and null models, $(\bar{\mu}_k^1(x), \bar{\sigma}_k^1)$ and $(\mathbf{0}, \bar{\sigma}_k^0)$, 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 Appendix 6.2). For the datasets in this paper, there are $b = 500$ bootstrap iterations ($b = 5000$ for the smoker study) and $K = 800$ modules.

Prior applications of ODP are focused on microarray studies where it is common to assume that the data generating process is approximately Normal and homoscedastic. However in sequencing studies, this assumption is no longer valid because the observations are heteroscedastic. In order to apply the ODP to sequencing studies, we implemented a similar strategy in ref. [9] 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 variance stabilizing weights w_{ij} and a d-dimensional natural cubic spline basis $\mathbf{s}(x_{ij})$ for $j = 1, 2, ..., n$ samples and $i = 1, 2, 3..., m$ genes, the data are transformed as $\mathbf{s}^*(x_{ij}) = w_{ij}\mathbf{s}(x_{ij})$ and $y_{ij}^* = w_{ij}y_{ij}$.

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 modular optimal discovery procedure (mODP) is applied to four different genomic experiments. The performance of mODP is compared to an F-test and a moderated F-test [6] using the number of discoveries and enriched gene sets. Finally, we validate 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 [11]. 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 [12]. 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 hr time points. The samples were hybridized onto U133 GeneChips with 44,924 probes. The expression values were log-normalized for variance stabilization.

Smoker study. The smoker study is a two group comparison between smoking and non-smoking humans [13]. 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 17β-estradiol (E2) in breast cancer cells (BUS cells) was examined [14]. There are five biological replicates for each

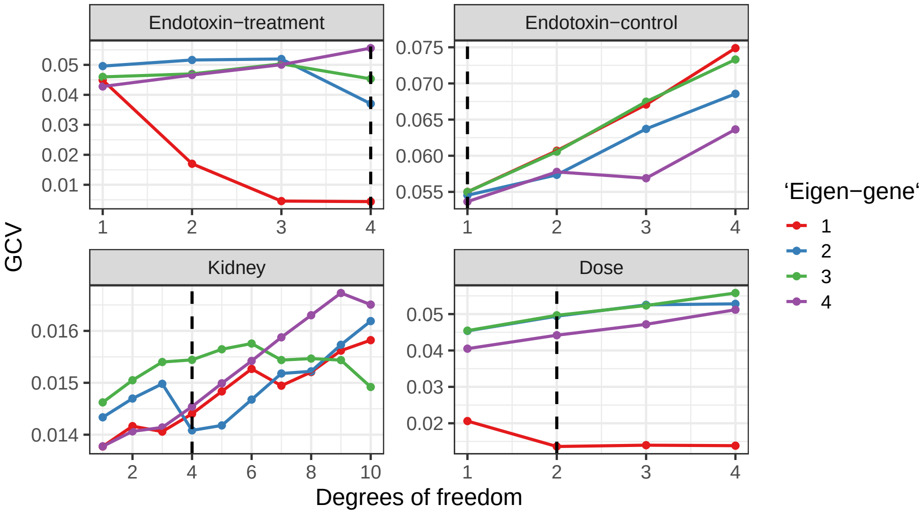

Figure 2: Cross validation error versus degrees of freedom for the first four eigen-genes. The dotted line indicates the chosen degrees of the freedom in the study.

E2 concentration, where the E2 concentrations considered were 0, 10, 30, 60, and 100 pM (25 total samples). After 48 hours exposed to E2, RNA samples were hybridized onto U133a GeneChips with 22,283 probes. The expression values were log-normalized 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 degrees of freedom d for the natural cubic 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 $d = 1, 2, ..., 10$ for the kidney study (an intercept term was included in the model). 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 (Figure 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 Appendix 6.1). Compared to the F-statistic, the moderated F-statistic shrinks

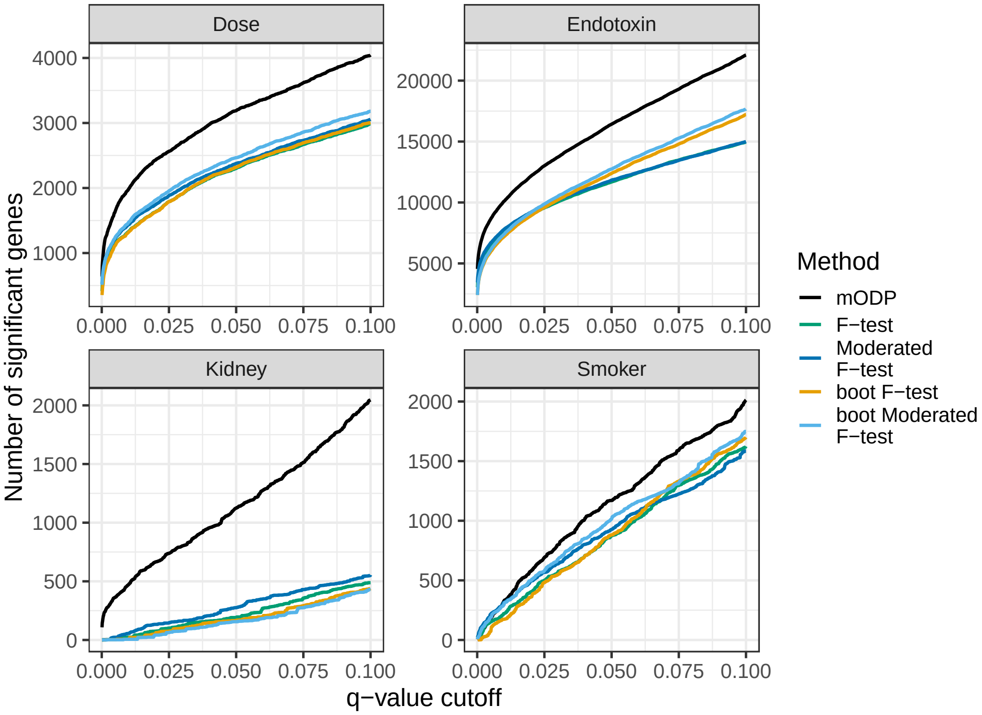

Figure 3: Observed number of discoveries at various q-value cutoffs from the modular optimal discovery procedure (black), F-test (green), bootstrap F-test (orange), moderated F-test (blue) and bootstrap moderated F-test (light blue). These methods were applied to four different studies: endotoxin (longitudinal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

the sample variance towards a pooled variance. This shrinkage allows for more stable inferences with low sample size studies [6]. 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 Appendix 6.2). 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 false discovery rates (Figure 3). At each false discovery rate threshold, the mODP finds substantially more differentially expressed genes compared to the other methods. For example, when applying a false discovery rate of 0.1, mODP detects 1481, 273, 6637, and 887 more differentially expressed genes in the kidney, smoker, endotoxin and dose studies, respectively. Additionally, the mODP finds nearly all of the differentially expressed genes detected by the other methods (Figure 4). Finally, we find that the mODP has the lowest estimated proportion of true nulls across all studies (Table 1). This suggests that the mODP estimates a higher expected number of

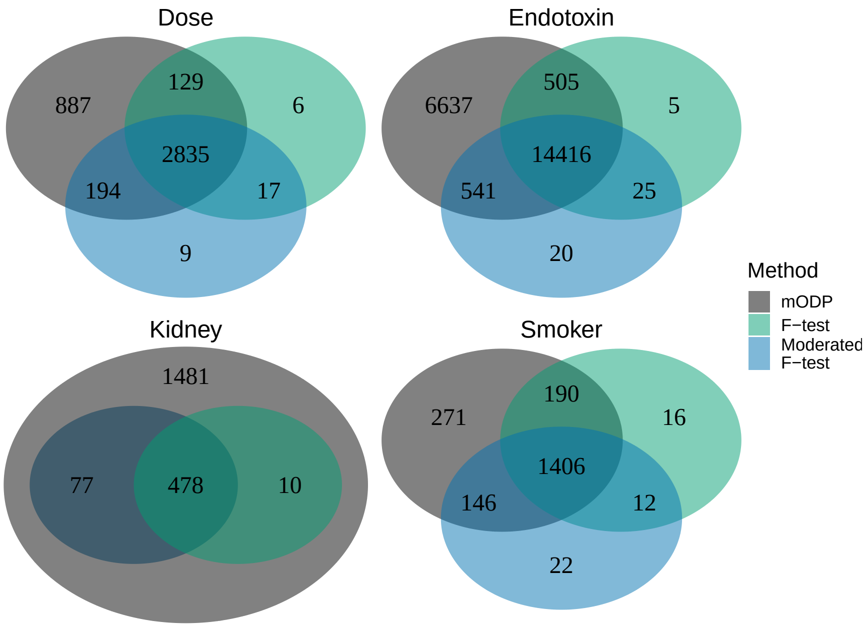

Figure 4: Venn diagram of the total discoveries at a false discovery rate of 0.1 for the studies in Figure 3. Only the modular optimal discovery procedure (black), F-test (green) and moderated F-test (blue) are shown.

alternative genes.

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 [15, 16]. 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 Appendix 6.4 for additional details). We find that the mODP has the largest proportion of true positives across all gene sets compared to other methods (Figure 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 (i.e., alternative and null) based on a false discovery rate 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

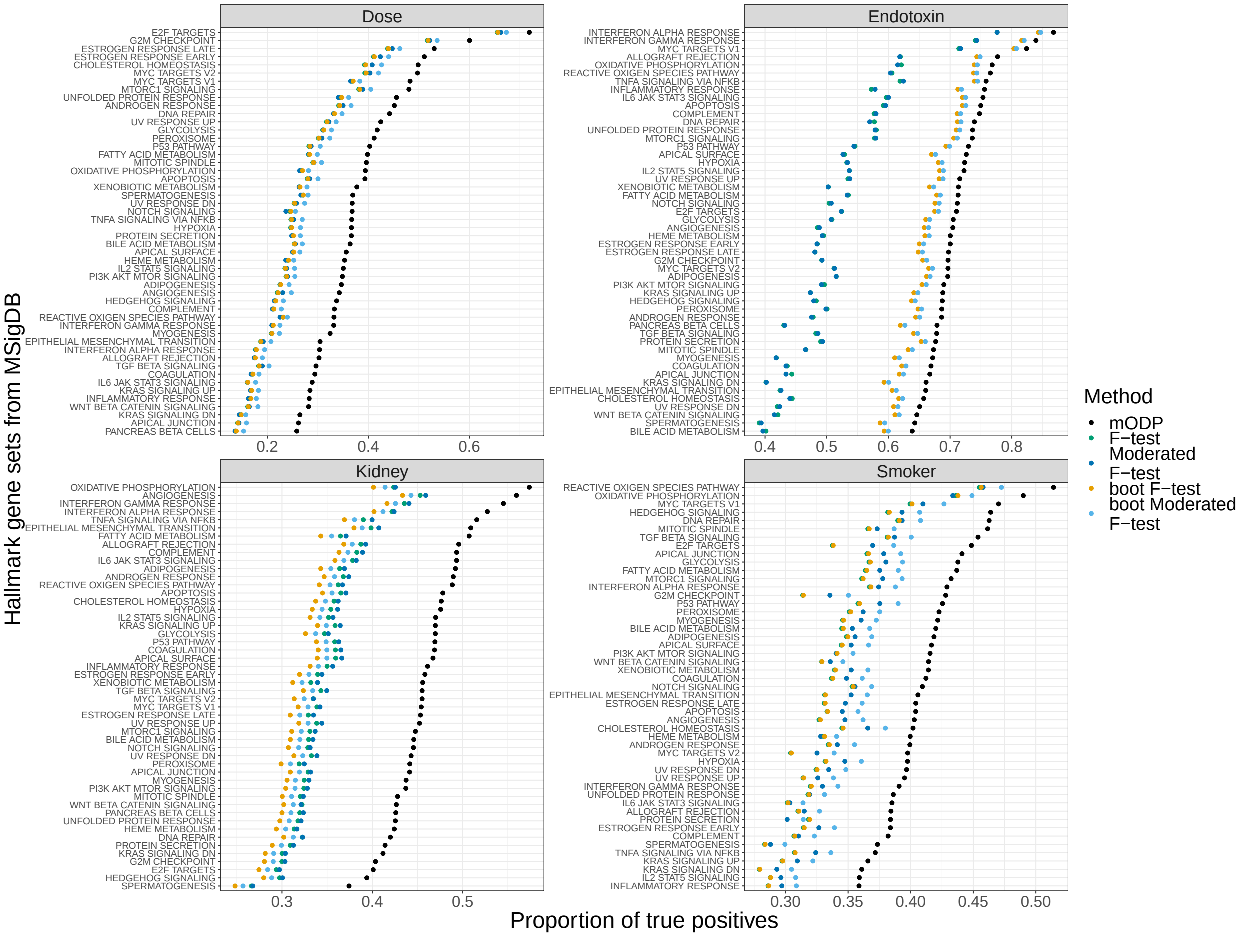

Figure 5: Proportion of true positives for the Hallmark gene sets from MSigDB. Enrichment results for each method are shown for the dose, endotoxin, kidney, and smoker studies.

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 were chosen to keep the observed proportion of true nulls fixed. For more details see Appendix 6.3.

To compare the testing procedures in the simulated datasets, we calculated the estimated false discovery rate (FDR) and the total number of discoveries. We find that the mODP controls the FDR in all simulated studies (Figure 6b). Furthermore, substantially more differentially expressed genes were detected relative to the other testing procedures (Figure 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 very small sample sizes. When the number of unique gene expression patterns is increased, the power of the mODP decreases. This is because the number of unique gene expression curves does not change the power of the F-test and moderated F-test.

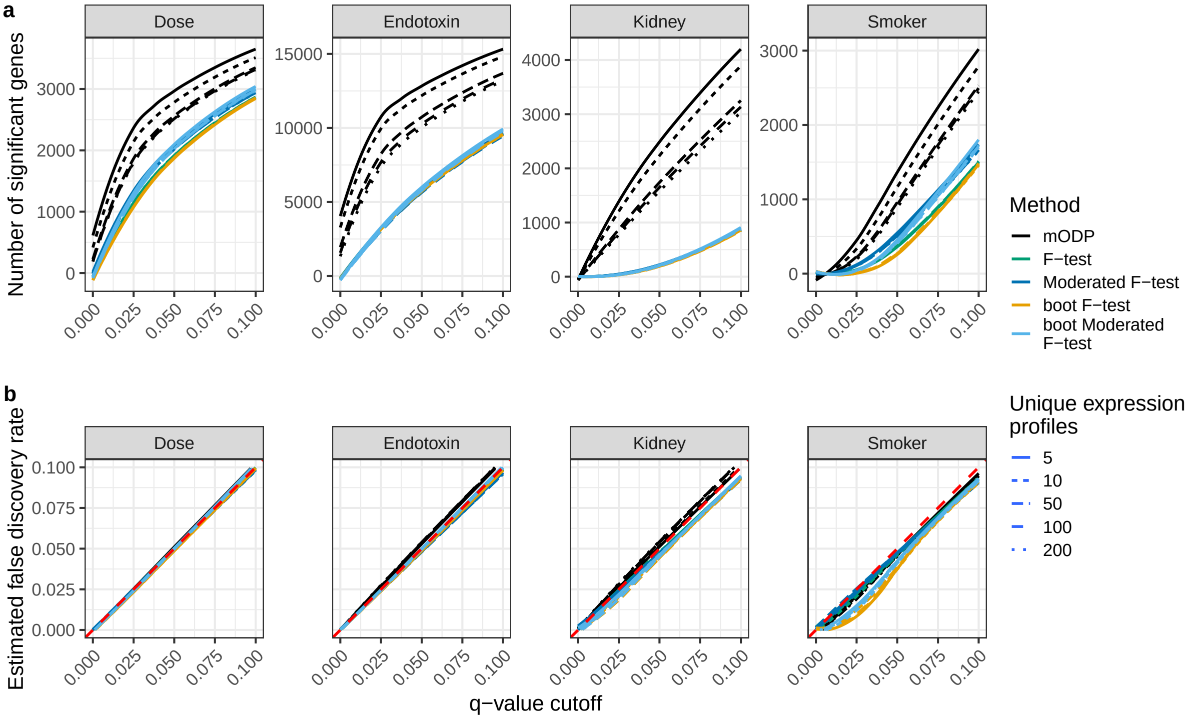

Figure 6: Simulation results for the dose, endotoxin, kidney, and smoker studies using 5, 10, 50, 100, and 200 unique gene expression profiles (linetype). The modular optimal discovery procedure (black), F-test (green), moderated F-test (blue), bootstrap F-test (orange) and bootstrap moderated F-test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

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 [3, 2, 17], 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 genomic study, genes share similar patterns of gene expression based on their functional roles. 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 functionally related genes, we varied the number of unique gene expression profiles with simulated data. We find that the ODP loses power as the number of unique gene expression profiles increases: there are fewer functionally related genes and so there is less information that can be leveraged in the test statistic. As an extreme case, suppose all genes in a study are functionally unrelated. In this scenario, the ODP has been shown to perform similar to the F-test (see [3]). 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 functional relationships among genes provides the best results. Second, there needs to be an adequate number of samples 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 samples per biological condition to generate a valid empirical null. In the studies considered here, there are at least four biological replicates per condition. Lastly, the appropriate degrees of freedom needs to be carefully chosen to avoid overfitting the spline. To this end, we implemented a procedure from ref. [10] 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 functional 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 [18]. Another improvement is to extend the ODP to generalized linear models where the response variable follows an exponential family distribution. This would enable its extension to genome-wide association studies where complex traits are commonly non-Normal. These avenues will be explored in future work.

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 Appendix

6.1 The F-statistic and moderated F-statistic

Suppose gene expression data y_{ij} and explanatory variable x_{lj} are observed for $i = 1, 2, ..., M$ genes, $j = 1, 2, ..., n$ samples, and $l = 1, 2, ..., d$ explanatory variables; the design matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d)$ is assumed to be full rank. In this case, the null model $y_{ij} = \sum_{l=1}^{d_0} \beta_{il} x_{jl} + \epsilon_{ij}$ is tested versus the alternative model $y_{ij} = \sum_{l=1}^{d} \beta_{il} x_{jl} + \epsilon_{ij}$ where $1 \leq d_0 < d$ and ϵ_{ij} are uncorrelated random errors that follow a Normal distribution with mean zero and variance σ_i^2. We are interested in comparing both models to infer whether $\beta_{il} \neq 0$ for at least one of the $l = d_0 + 1, ..., d$ explanatory variables.

The F-test is a classical testing procedure that can be used to compare nested regression models. The general procedure works as follows. The alternative model is fit using ordinary least squares to the observed data to estimate the parameters $\hat{\beta}_{il}$ and the residual vector $\mathbf{e}_i = \mathbf{y}_i - \sum_{l=1}^{d} \hat{\beta}_{il} \mathbf{x}_l$. Similarly, the null model is fit to estimate the parameters $\hat{\beta}_{il}^{\text{null}}$ and the residual vector $\mathbf{e}_i^{\text{null}} = \mathbf{y}_i - \sum_{l=1}^{d_0} \hat{\beta}_{il}^{\text{null}} \mathbf{x}_l$. The test statistic is defined as

$$F(\mathbf{y}_i) = \frac{\left(\left\| \mathbf{e}_i^{\text{null}} \right\|^2 - \left\| \mathbf{e}_i \right\|^2 \right) / (d - d_0)}{\left\| \mathbf{e}_i \right\|^2 / (n - d)}, \tag{7}$$

where the theoretical distribution under the null hypothesis follows Fisher's F-distribution with $d - d_0$ degrees of freedom in the numerator and $n - d$ degrees of freedom denominator (denoted $F_{d-d_0,n-d}$). Intuitively, if there is no difference between both models then $F(\mathbf{y}_i)$ should be concentrated around 1. Otherwise, large deviations from 1 provide evidence against the null model. The assumption that the F-statistic follows an $F_{d-d_0,n-d}$-distribution under the null hypothesis is only true asymptotically: in practice, there needs to be a large number of samples for reliable inferences.

For small sample sizes, the moderated F-statistic can be used to compare two models. The main issue with small sample sizes is that the sample variance can often be inflated and unreliable to use in the traditional F-test. The moderated F-test is an empirical Bayes procedure that borrows information across genes to provide stable estimates of the sample variance [6]. A rough outline of the hierarchical model is as follows. The inverse variance across genes are assumed to vary as a scaled chi-squared distribution, i.e.,

$$\frac{1}{\sigma_i^2} \sim \frac{1}{\rho_0 \sigma_0^2} \chi^2_{\rho_0}, \tag{8}$$

where ρ_0 is the degrees of freedom and σ_0^2 is a scaling factor. Furthermore, the non-zero effect sizes are assumed to follow a Normal distribution with mean 0 and variance proportional to σ_i^2. The posterior mean of σ_i^2 given the sample variance can be determined from the above hierarchical model, see [6] for more details. This mean value is used as an improved estimate of the sample variances, where the sample variances are shrunken towards the prior estimator σ_0^2 for more stable estimates. More specifically, the moderated F-statistic is defined as

$$F(\mathbf{y}_i) = \frac{\left(\left\|\mathbf{e}_i^{\text{null}}\right\|^2 - \left\|\mathbf{e}_i\right\|^2\right)/(d-d_0)}{\left\|\mathbf{e}_i^*\right\|^2/((n-d)+\rho_0)}, \tag{9}$$

where $\|\mathbf{e}_i^*\|^2 = \|\mathbf{e}_i\|^2 + \rho_0\sigma_0^2$ and the parameters (ρ_0, σ_0^2) are estimated from the data [6]. The theoretical distribution under the null hypothesis for the moderated F-statistic follows an F-distribution with $d - d_0$ degrees of freedom in the numerator and $(n-d)+\rho_0$ degrees of freedom in the denominator, $F_{d-d_0,(n-d)+\rho_0}$. When there are a sufficient number of samples, $\|\mathbf{e}_i^*\|^2 \approx \|\mathbf{e}_i\|^2$ and $(n-d)+\rho_0 \approx n-d$. Thus the statistical power of the moderated F-test and F-test will be similar for large sample sizes.

6.2 Generating an empirical null

A standard bootstrap procedure was implemented to generate an empirical null distribution for the testing procedures. For $i = 1, 2, ..., M$ genes,

1. Assume the model fit for the null model is $\mathbf{y}_i = f_0(\mathbf{x}_i) + \epsilon_i$ and for the alternative model is $\mathbf{y}_i = f_1(\mathbf{x}_i) + \epsilon_i$ where ϵ_i are the random errors.

2. Fit both models to the observed data using ordinary least squares: Estimate $\hat{f}_1(\mathbf{x}_i)$ and $\mathbf{e}_i$ for the alternative model and $\hat{f}_0(\mathbf{x}_i)$ and $\mathbf{e}_i^{\text{null}}$ for the null model. Calculate the test statistic of interest (i.e., the ODP statistic, F-statistic, or moderated F-statistic), denoted by $T_i(\hat{f}_0(\mathbf{x}_i), \hat{f}_1(\mathbf{x}_i))$.

3. Adjust the residuals by calculating studentized residuals: $\mathbf{e}_i^s = \mathbf{e}_i(1 - \text{Tr}(\mathbf{P}))$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix under the alternative model.

4. For $b = 1, 2, ..., B$ bootstrap samples, sample n observations from the studentized residuals (with replacement) to obtain $\mathbf{e}_i^{s*(b)}$. Add these residuals to the null model fit $\mathbf{y}_i^{*(b)} = \hat{f}_0(\mathbf{x}_i) + \mathbf{e}_i^{*(b)}$.

5. Fit both models to $\mathbf{y}_i^{*(b)}$ and obtain $\hat{f}_0^{*(b)}(\mathbf{x}_i)$ and $\hat{f}_1^{*(b)}(\mathbf{x}_i)$ estimates under the null and alternative models, respectively. Calculate $T_i^{*(b)}(\hat{f}_0^{*(b)}(\mathbf{x}_i), \hat{f}_1^{*(b)}(\mathbf{x}_i))$ for $b = 1, 2, ..., B$ bootstrap samples. Note that the hyperparameters of the moderated F-statistic (d_0, σ_0) are fixed and so $\|\mathbf{e}_i^*\|^2 = \left\|\mathbf{e}_i^{*(b)}\right\|^2 + d_0\sigma_0^2$.

6. Calculate the empirical p-values as

$$p_i = \frac{\sum_{b=1}^{B} \mathbf{1}\left(T_i^{*(b)}(\hat{f}_0^{*(b)}(\mathbf{x}_i), \hat{f}_1^{*(b)}(\mathbf{x}_i)) \geq T_i(\hat{f}_0(\mathbf{x}_i), \hat{f}_1(\mathbf{x}_i))\right)}{B}.$$

For the ODP, the studentized residuals need to be rescaled by the observed sample variance. This enforces that the sample variance remains the same for all bootstrap iterations. Thus step (4) is $\mathbf{e}_i^{s'*(b)} = \mathbf{e}_i^{s*(b)} \frac{\hat{\sigma}_i}{\hat{\sigma}_i^{s*(b)}}$ where $\hat{\sigma}_i = \frac{\|\mathbf{e}_i\|}{\sqrt{n-d}}$ is the sample standard deviation of the residuals from the original alternative model (d explanatory variables) and $\hat{\sigma}_i^{s*(b)} = \frac{\left\|\mathbf{e}_i^{*(b)}\right\|}{\sqrt{n-1}}$ is the standard deviation from the resampled residuals.

It is important to note that additional steps in the above algorithm may need to be taken when handling longitudinal data. See ref. [10] for more details.

6.3 Simulation details

The primary objective in the simulation study is to generate replicate datasets of the observed studies. We use the biological signal from each study as a baseline: both models are fit to estimate the gene expression curves under the alternative and null models. The genes assigned to the alternative model had q-values $< .1$ while genes assigned to the null model had q-values $> .1$. The number of unique curves from the alternative model was varied by randomly selecting from the population of genes assigned to the alternative model. For each study and number of unique gene expression curves $g = 5, 10, 50, 100, 200$, the procedure is outlined below:

1. Use the estimated proportion of true nulls $\hat{\pi}_0$ to randomly assign the m genes to either the alternative ($m(1 - \hat{\pi}_0)$) or null ($m\hat{\pi}_0$) models. Genes assigned to the alternative model followed a unique gene expression curve g, $f_1^g(\mathbf{x}_i)$. Alternatively, the null genes were randomly sampled from the population of null model fits $f_0^*(\mathbf{x}_i)$.

2. Using the observed signal-to-noise ratio (SNR) distribution from the alternative model, calculate an appropriate SNR_M such that the estimated number of differential expressed genes at a false discovery rate of 0.1 is close to the observed study. This was done by trial and error: $\mathbf{y}_{ig} = f_1^g(\mathbf{x}_i) + \sigma_\text{i}^*$, where σ_i^* is randomly sampled from the population of standard deviations $\frac{\sigma_\text{g}}{\sqrt{\text{SNR}_\text{M}}}$ for all g.

	ODP	F-test	Mod. F-test	boot. F-test	boot. mod. F-test
Dose	0.672	0.803	0.813	0.804	0.793
Endotoxin	0.349	0.605	0.615	0.388	0.397
Kidney	0.585	0.687	0.685	0.684	0.676
Smoker	0.600	0.681	0.681	0.680	0.669

Table 1: Estimated proportion of true nulls.

3. Randomly sample from the population of standard deviations in the previous step to add noise to the alternative model $\mathbf{y}_{ig} = f_1^g(\mathbf{x}_i) + \sigma_{\mathrm{i}}^*$ and the null model $\mathbf{y}_i = f_0^*(\mathbf{x}_i) + \sigma_{\mathrm{i}}^*$ for all genes; call this simulated dataset $\mathbf{Y}^*$.

4. Apply the testing procedures to $\mathbf{Y}^*$ and calculate p-values.

5. Repeat steps (3-4) 500 times and calculate the average number of discoveries and the average false discovery rate for all testing procedures.

The estimated proportion of true nulls are shown in Table 1 and the estimated $\mathrm{SNR_M}$ for the dose, endotoxin, kidney and smoker studies are the 0.86, 0.45, 0.35, and 0.8 quantiles of the SNR distribution, respectively.

6.4 True positive enrichment analysis

Consider $i = 1, 2, ..., m$ test statistics z_i calculated on a gene-by-gene basis from a biological study. Given these test statistics, we propose a new summary statistic for gene sets based on the enrichment of true positives. The procedure works as follows. For each gene i, we can calculate the local false discovery rate based on the chosen test statistic:

$$\mathrm{lfdr}(z_i) = \mathrm{Pr}\{\mathrm{null}|z_i\} = \pi_0 \frac{f_0(z_i)}{f(z_i)},$$

where π_0 is the prior probability that a hypothesis test is null, $f_0(z_i)$ is the null density, and $f(z_i)$ is a mixture of the null and alternative densities [7]. Next, we average the local false discovery rate in gene set S,

$$\mathrm{TPE}(S) = \frac{\sum\limits_{i \in S} (1 - \mathrm{lfdr}(z_i))}{|S|}.$$

As an example, if we calculated TPE$(S) = 0.9$ then it corresponds to a gene set with an average of 0.9 true positives. Thus this gene set has a high proportion of true positives.

The advantages of working in this framework are (i) it is computationally fast to calculate TPE(S) for all gene sets, (ii) the interpretation of important gene sets is more intuitive, and (iii) covariate-adjusted local false discovery rates can easily be incorporated to improve statistical power.

References

[1] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

[2] John D. Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, jun 2007. doi: 10.1111/j.1467-9868.2007.005592.x. URL https://dx.doi.org/10.1111/j.1467-9868.2007.005592.x.

[3] John D. Storey, J. Y. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, 8(2): 414–432, aug 2006. doi: 10.1093/biostatistics/kxl019. URL https://dx.doi.org/10.1093/biostatistics/kxl019.

[4] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, 04 2001. doi: 10.1073/pnas.091062498. URL https://www.ncbi.nlm.nih.gov/pubmed/11309499.

[5] Xiangqin Cui, J. T. Gene Hwang, Jing Qiu, Natalie J. Blades, and Gary A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 01 2005. doi: 10.1093/biostatistics/kxh018. URL http://dx.doi.org/10.1093/biostatistics/kxh018.

[6] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, 2004. doi: 10.2202/1544-6115.1027. URL https://www.degruyter.com/view/j/sagmb.2004.3.issue-1/sagmb.2004.3.1.1027/sagmb.2004.3.1.1027.xml.

[7] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001. doi: 10.1198/016214501753382129. URL https://doi.org/10.1198/016214501753382129.

[8] Ingrid Lönnstedt and Terry Speed. Replicated microarray data. *Statistica Sinica*, 12(1):31–46, 2002. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24307034.

[9] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(2):R29, Feb 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-2-r29. URL https://doi.org/10.1186/gb-2014-15-2-r29.

[10] John D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36): 12837–12842, sep 2005. doi: 10.1073/pnas.0504609102. URL https://dx.doi.org/10.1073/pnas.0504609102.

[11] Graham E. J Rodwell, Rebecca Sonu, Jacob M Zahn, James Lund, Julie Wilhelmy, Lingli Wang, Wenzhong Xiao, Michael Mindrinos, Emily Crane, Eran Segal, Bryan D Myers, James D Brooks, Ronald W Davis, John Higgins, Art B Owen, and Stuart K Kim. A transcriptional profile of aging in the human kidney. *PLOS Biology*, 2(12):e427–, 11 2004. doi: 10.1371/journal.pbio.0020427. URL https://doi.org/10.1371/journal.pbio.0020427.

[12] Steve E. Calvano, Wenzhong Xiao, Daniel R. Richards, Ramon M. Felciano, Henry V. Baker, Raymond J. Cho, Richard O. Chen, Bernard H. Brownstein, J. Perren Cobb, S. Kevin Tschoeke, Carol Miller-Graziano, Lyle L. Moldawer, Michael N. Mindrinos, Ronald W. Davis, Ronald G. Tompkins, Stephen F. Lowry, Inflammation, and Host Response to Injury Large Scale Collaborative Research Program. A network-based analysis of systemic inflammation in humans. *Nature*, 437: 1032 EP –, 08 2005. doi: 10.1038/nature03985. URL https://doi.org/10.1038/nature03985.

[13] Dorothy M Ryan, Thomas L Vincent, Jacqueline Salit, Matthew S Walters, Francisco Agosto-Perez, Renat Shaykhiev, Yael Strulovici-Barel, Robert J Downey, Lauren J Buro-Auriemma, Michelle R Staudt, Neil R Hackett, Jason G Mezey, and Ronald G Crystal. Smoking dysregulates the human airway basal cell transcriptome at copd risk locus 19q13.2. *PloS one*, 9(2):e88051; e88051–e88051, 02 2014. doi: 10.1371/journal.pone.0088051. URL https://www.ncbi.nlm.nih.gov/pubmed/24498427.

[14] Kathryn R Coser, Jessica Chesnes, Jingyung Hur, Sandip Ray, Kurt J Isselbacher, and Toshi Shioda. Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in mcf7/bus breast cancer cells by dna microarray. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):13994–13999, 11 2003. doi: 10.1073/pnas.2235866100. URL https://www.ncbi.nlm.nih.gov/pubmed/14610279.

[15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell systems*,

1(6):417–425, 12 2015. doi: 10.1016/j.cels.2015.12.004. URL https://www.ncbi.nlm.nih.gov/pubmed/26771021.

[16] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics (Oxford, England)*, 27 (12):1739–1740, 06 2011. doi: 10.1093/bioinformatics/btr260. URL https://www.ncbi.nlm.nih.gov/pubmed/21546393.

[17] S. Woo, J. T. Leek, and John D. Storey. A computationally efficient modular optimal discovery procedure. *Bioinformatics*, 27(4):509–515, dec 2010. doi: 10.1093/bioinformatics/btq701. URL https://dx.doi.org/10.1093/bioinformatics/btq701.

[18] Xiongzhi Chen, David G. Robinson, and John D. Storey. The functional false discovery rate with applications to genomics. *bioRxiv*, 2017. doi: 10.1101/241133. URL https://www.biorxiv.org/content/early/2017/12/30/241133.