Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Adaptive local false discovery rate procedures for highly spiky data and their application RNA sequencing data of yeast *SET4* deletion mutants

Mark Louie Ramos [1,4], **DoHwan Park**[*,1], **Johan Lim** [2], **Junyong Park** [2], **Khoa Tran** [3], **Eric Joshua Garcia** [3], and **Erin Green** [3]

[1] Department of Mathematics and Statistics, University of Maryland Baltimore County, Maryland, USA

[2] Department of Statistics, Seoul National University

[3] Department of Biological Sciences, University of Maryland Baltimore County, Maryland, USA

[4] Department of Mathematics and Physics, University of Santo Tomas, Manila, Philippines

Received zzz, revised zzz, accepted zzz

Chromatin dynamics are central to the regulation of gene expression and genome stability. In order to improve understanding of the factors regulating chromatin dynamics, the genes encoding these factors are deleted and the differential gene expression profiles are determined using approaches such as RNA-sequencing. Here, we analyzed a gene expression dataset aimed at uncovering the function of the relatively uncharacterized chromatin regulator, Set4, in the model system *Saccharomyces cerevisiae* (budding yeast). The main theme of this paper focuses on identifying the highly differentially-expressed genes in cells deleted for Set4 (referred to as Set4Δ mutant dataset) compared to the wild type yeast cells. The Set4Δ mutant data produce a spiky distribution on the log fold changes of their expressions, and it is reasonably assumed that genes which are not highly differentially-expressed come from a mixture of two normal distributions. We propose an adaptive local false discovery rate (FDR) procedure, which estimates the null distribution of the log fold changes empirically. We numerically show that, unlike existing approaches, our proposed method controls FDR at the aimed level (0.05) and also has competitive power in finding differentially expressed genes. Finally, we apply our procedure to analyzing the Set4Δ mutant dataset.

Key words: Empirical Approximation; False Discovery Rate; Mixture of Normal; Multiple Testing

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX (please delete if not applicable)

1 Introduction

In gene expression studies, the abundance of product (different RNA species) from individual genes is measured in response to a particular perturbance of cells, such as the incorporation of a mutation into a master regulator of gene expression or treatment with a chemical or cellular stress stimulant. Often, this perturbance results in a very specific pattern of gene expression changes, in which a relatively small subset of the genes tested exhibit high differential expression. It is of interest to identify genes that experience highly differentiated expression as a direct consequence of the genetic or chemical change. The endeavor of identifying these genes from among many thousand candidates presents a multiple testing problem where some common statistic is used as a measure of differentiation. In addition, the sheer number of

*Corresponding author: e-mail: dhpark@umbc.edu, Phone: +1-765-586-2018, Fax: +1-410-455-1066

hypotheses that need to be tested makes false discovery rate (FDR) a reasonable error rate to consider. The BH procedure based on p-values in Benjamini and Hochberg (1995) is a typical approach in multiple testing problems that controls false discovery rate (FDR).

A local FDR method by Efron (2004) uses the empirical Bayes approach in order to estimate relevant parameters in the null distribution of test statistics. In Efron, the null distribution is assumed to be normal with unknown parameters. Both the BH procedure and the local FDR procedure may fail in controlling a given error rate if any of the assumptions on the null distribution is violated.

In the motivating study, our investigation focused on the gene expression consequences of removing the chromatin regulator Set4 from *Saccharomyces cerevisiae* yeast cells (Tran *et al.*,2018). The Set4 protein contains a common signature of chromatin-associated proteins and previous work has shown it plays an important role in regulating gene expression in response to stress (Serratore *et al.*, 2018). Notably, Set4 is a conserved protein (Zhang *et al.*, 2017), with human counterparts that contribute to both cancer progression and neurodevelopmental disorders (Deliu *et al.*, 2018) and this type of study in yeast frequently improves our understanding of how these chromatin regulators control gene expression in the pathological conditions (Tran and Green, 2019). The study involves the examination of about 8,600 coding and non-coding genes in order to identify highly differentiated candidate genes using the logfold change statistic. This statistic measures the extent by which each gene changes from their wild type when Set4 is absent from cells. Highly positive or negative logfold values are indicative of the strongly differentiatially-expressed genes that provide important biological insight.

Some studies that have also dealt with logfold change in the same context of gene differentiation assume the distribution of this statistic to be approximately Normal (Kammers *et al.*, 2015), making standard tests such as the student's t-test applicable (McCarthy and Smyth, 2009). For multiple testing, this implies that Efron's local FDR procedure is likewise applicable. However, at least in the particular situation of the Set4Δ mutant data set, assuming normality of the logfold change does not seem to be reasonable. Cells often have very specific responses to genetic or chemical disruptions, and the majority of the genes in the genome will not show differential expression following the perturbance compared to standard conditions. This means that many of the genes are expected to have logfold change values that are very close to zero. However, there is also often a subset of genes that show small changes, most likely due to indirect consequences of genetically or chemically perturbing the cell. Either group is not of interest to the biologist and so belong to the null distribution. Thus, the resulting null distribution of the test statistic is heavily concentrated around zero, making the distribution have a very high peak, and is a mixture of two distributions. It is assumed that each component retains the general assumption of normality for logfold change data, and so this mixture is considered to be a mixture of two normal distributions. Since Efron's local FDR procedure is based on a single distribution for the null distribution, it is not directly applicable to the motivating dataset where there are two different sources of insignificant genes. Efron's approach may lead the local FDR procedure to fail in controlling a given error rate. Given the biological context leading experimenters to believe in two sources of undifferentiated genes, the null distribution is modelled as a mixture of two normal distributions. Nonetheless, for this estimation of the null distribution, an interval type zero assumption initially used in the work of Efron (2004) can still be applied. In particular, the work of Park *et al.*(2011) expanded this assumption for use in estimating a mixture of two normal distributions. Specifically, they used an adaptive procedure to model the null distribution as a mixture of two normal distributions based on the assumption that for some interval around the center of the data, all data in the interval came purely from the null distribution. The choice of this interval is critical since a selected interval affects the estimation of the null distribution. The method in Park *et al.*(2011) is based on selecting an interval in which the data are used for estimating the null distribution. The choice of the optimal interval is based on widening interval and the calculation of some criterion corresponding to the data in the interval. This approach may be sensitive and unstable for the case of high-peaked data since a very small increase of interval causes a dramatic change of data in the interval leading to unstable estimation of the null distribution. From our simulations shown later, we see that the method in Park *et al.*(2011) performs very conservatively.

When we use the idea of widening interval from the center of the whole data, any desirable criterion for selecting an optimal interval should be sensitive at detecting a change from purely null to mixture of null and alternative distributions.

Using the idea of widening interval, Gauran *et al.*(2018) proposed a criterion based on change-point detection to identify the point where behavior of the criterion indicates contamination of data from the alternative distribution. One main idea of this paper is that we modify the criterion of change-point detection considering the very spiky peaked phenomenon to have more stable estimation procedures. Furthermore, we also improve the method in Park *et al.*(2011) to avoid serious conservative decisions.

The remainder of the article is organized as follows. In Section 2, a brief background on the existing methods and their limitations are provided. Section 3 describes a framework for the newly proposed methods including three algorithms. Section 4 highlights the simulation studies for a variety of cases while Section 5 shows the real application of the methods to the Set4Δ mutant data set.

2 Development of local FDR Method

2.1 Original Method

In this section, we briefly review the FDR procedure and provide our model and its estimation procedure. When multiple hypotheses are tested simultaneously, control over type I error becomes an important issue (Benjamini and Hochberg, 2011). Different type I error controlling procedures have been developed to address this issue (Storey and Tibshirani, 2003). For example, family-wise error rate (FWER) seeks to control the probability of encountering at least one false discovery at a given nominal rate (Aickin and Gensler, 1996). Procedures that control this error rate, such as the Bonferroni Procedure or Holm's procedure are typically conservative (Shaffer, 1995).

On the other hand, false discovery rate (FDR) is the expected proportion of falsely rejected null hypotheses. Multiple testing procedures controlling FDR are likely to reject more hypotheses than those controlling FWER. Benjamini and Hochberg (1995) developed the procedure controlling FDR for independent p-values while the procedure in Benjamini and Yekutieli (2011) is valid under arbitrary dependence. Efron (2004) constructed the local FDR procedure which are based on the posterior probability of the observed data from the view point of empirical Bayes. When there are $z_1, \ldots, z_n$, it is of interest to test the following hypotheses simultaneously: for $1 \leq i \leq n$,

$$H_{0i} : z_i \sim f_0 \; vs. \; H_{1i} : z_i \sim f_1. \tag{1}$$

Each z_i is modeled as

$$f(z) \quad = \quad p_0 f_0(z) + (1 - p_0) f_1(z) \tag{2}$$

where $f_0(z) = \frac{1}{\sigma_0}\phi((z - \mu_0)/\sigma_0)$ for $\phi(z) = \frac{1}{\sqrt{2pi}} e^{-\frac{z^2}{2}}$ and $p_0 = P(H_{0i})$ for all $1 \leq i \leq n$. Based on Efron (Efron, 2004), the local false discovery rate for a given statistic z is defined as

$$fdr(z) = \frac{p_0 f_0(z)}{f(z)} \tag{3}$$

which requires the estimates of (f_0, f, p_0). $f(z)$ can be estimated using Poisson-regression or splines. $f_0(z)$ is estimated using the "zero-assumption," which states that elements found around the central peak of the distribution come exclusively from the null distribution. With this assumption and the assumption that the null distribution is normal, quadratic splines can be used in order to accomplish the estimation. Finally, the estimates of $f(z)$ and $f_0(z)$ can be used to estimate p_0. With these estimates, the local FDR at any value of the statistic can be computed and may be interpreted as the average number of false discoveries that would occur if hypotheses associated with this statistic are rejected.

2.2 Misspecification of the Null Distribution

Misspecification of the null distribution poses a serious problem in any simultaneous inference approach (Efron, 2004). Despite of this, limited research was found on examining the general validity of the local FDR assumption that the null distribution of test statistics is normal. Schwartzman (2008) addressed the misspecification problem for the null distribution by extending Efron's procedure to fit the null to exponential families in general. However, this extension is also insufficient if the null does not belong to the exponential family. Robin *et al.* (2007) and Jeong *et al.* (2018) used semiparametric methods to model the mixture distribution for purposes of application in the local FDR procedure but retained the null distribution as normal. Other studies were also found that modified Efron's local FDR procedure according to the demands of particular research contexts such as functional MRI data (Lee *et al.*, 2016) or soil content data (Chauveau *et al.*,2014). Efron's local FDR procedure uses a more flexible null distribution; a family of normal distributions with empirically estimated parameters instead of a theoretical null distribution such as a standard normal distribution. However, such an extension of the null distribution may not be enough in many practical situations where the null distribution is more complicated based on the experimental context. In particular, we consider an additional extension based on scientific reason that the null distribution is a mixture of two normal distributions which is suitable for the Set4Δ mutant data. Park *et al.*(2011) used a mixture of two normal distributions for the null distribution to explain a heavy tailed null distribution with a couple of estimation procedures for relevant parameters. They highlighted that a mixture of two normal distribution has some advantage in fitting a heavy-tailed null distribution than a normal distribution. On the other hand, the data from Set4Δ mutant data are mostly centered around zero due to the high-peaked phenomenon which may cause some difficulties in the estimation of all relevant parameters in local FDR procedures. It will be demonstrated that the existing method is unstable in fitting the spiky data with mixture of very small variance and wider one and produce fairly conservative decisions under such data. We will address some difficulties and then propose three methodologies to overcome the difficulties or improve some existing procedure in the following sections.

3 Proposed model and estimation Procedure

In Efron's local FDR procedure, a single normal distribution is used for f_0 in (2) (Efron, 2004), however f_0 may have a more complicated form such as a mixture of two distributions for some scientific reason as in the case of Set4Δ mutant data described in the introduction. Instead of a single normal distribution for f_0, we consider a mixture of two normal distributions as follows:

$$f_0(z) = \eta\phi_1(z) + (1-\eta)\phi_2(z) \tag{4}$$

where ϕ_1 and ϕ_2 are each the densities of normal distributions with corresponding parameters (μ_1, σ_1^2) and (μ_2, σ_2^2). The type of high-peaked data is explained by a normal distribution with a small variance, say $\sigma_1 << \sigma_2$, so that ϕ_1 corresponding to σ_1 represents a "spiky" distribution with almost a zero-inflation phenomenon observed at the center of data. Regarding f_1 in (2), it is not necessary to assume anything on f_1 if we estimate f directly, however we put a condition called the zero assumption as follows:

$$f(z) = p_0 f_0(z) \text{ for } z \in [-c, c] \tag{5}$$

which is also used in other studies (Efron, 2004; Park *et al*., 2011; Gauran *et al.*, 2018). From the zero assumption in (5), we can avoid the issue of identifiability in the mixture model of f_0 and f_1 since the support of f_1 is $(-\infty, -c) \cup (c, \infty)$. More importantly, it is not of interest to distinguish between the different types of undifferentiated genes from ϕ_1 and ϕ_2 found in $[-c, c]$. Instead, we need only to estimate f_0 for calculating $fdr(z)$ in (3). Regarding c, c in (5) is unknown in practice, however Efron (2004) used an ad-hoc choice of c which may affect the estimation of parameters in the null distribution. Thus, in the calculation of local FDR, a good estimate of unknown c is needed to have desirable properties of multiple testing procedure since both underestimating or overestimating c is problematic. Underestimating c can

cause instability of parameter estimates while overestimating c can lead to a considerable loss of power. In the following sections, we identify and address some difficulties in estimating c for the spiky distributed data set in local FDR procedure and provide (i) estimation of (f_0, f, p_0) for a given c and (ii) estimation of c itself.

3.1 Estimation of (f_0, f, p_0) for a fixed c

When c in (5) is known, estimation procedure is based on truncated data, $z_i \in \mathcal{I} = \{z_i : |z_i| \leq c\}$. The parameters in f_0 are estimated using the EM algorithm based on the conditional distribution of $f_0(z)$ given $|z| \leq c$. More specifically, we use the conditional density of z given $|z| \leq c$ and estimate all parameters $(\mu_1, \sigma_1, \mu_2, \sigma_2, \eta)$ in $f_{0,c}(z) = \frac{f_0(z)}{\int_{-c}^{c} f_0(z)dz}$ as in Dempster *et al.* (1977) where $f_0(z)$ is defined in (2). Regarding the estimation of f, any nonparametric density estimation procedure such as poisson regression may be used. Specifically, a method using splines that is included in the local FDR procedure of Efron (2004) was used. This is implemented by the locfdr function of the `locfdr R-package, https://cran.r-project.org/web/packages/locfdr/locfdr.pdf`. Finally for estimating p_0, we utilize the estimators of f and f_0 and use an estimator of p_0 by integrating both sides in (5) leading to the following restricted estimator of p_0,

$$\hat{p_0} = \min\left(1, \frac{\hat{F}(c) - \hat{F}(-c)}{\hat{F}_0(c) - \hat{F}_0(-c)}\right), \tag{6}$$

where $\hat{F}$ and $\hat{F}_0$ are the cumulative distribution functions from estimated $\hat{f}$ and $\hat{f}_0$, respectively.

3.2 Choice of c

The remaining step is to estimate the zero-assumption interval $[-c, c]$ or simply c. In this regard, Efron (2004) used the prefixed interval which makes a strong assumption that may not be applicable in all situations. In the high-peaked data set, there are several problems in selection of c. Park *et al.* (2011) proposed a couple of methods to choose c based on comparison of the estimate of p_0 and the relative frequencies in $[-c, c]$ for different c values. One of those methods is based on the idea of goodness-of-fit (GoF) test selecting the best interval with the largest p-value from GoF tests for different intervals. In high-peaked data, a large portion of the data are concentrated around the spiky part which dominates other observations including the differentiated expressions. Our simulation studies later will demonstrate that this procedure is fairly conservative so there is a room for improvement.

On the other hand, Gauran *et al.* (2018) proposed a loglikelihood-based criteria using the idea of change point detection in multiple testing problem for discrete data. For different c_is with $c_{i-1} < c_i$ for $i = 1, 2, ..., I$ and $c_0 = 0$, $[-c_i, c_i]$ can either contain points only from f_0 or from the mixture of f_0 and f_1.

For the data $z \in [-c, c]$, the correct likelihood is obtained from $f_0(z)$ which is larger than $f(z)$ due to the zero assumption, $f(z) = p_0 f_0(z) < f_0(z)$ and $0 < p_0 < 1$. Conversely, the correct likelihood of the data from the outside of $[-c, c]$ is $f(z)$ which tends to be larger than $f_0(z)$. From this idea, the sign of $\log \frac{f_0(z)}{f(z)}$ can be an indication of whether z belongs to either inside or outside of $[-c, c]$. The cumulative sum of log ratio corresponding to the data in $[-c, c]$ tends to increase and then starts to decrease when the data from outside of $[-c, c]$ start to be included.

Based on this criterion and the predetermined grid points of c_i for $i = 1, \ldots, I$, Gauran *et al.* (2018) used the EM algorithm to have estimates $(\hat{p}_{0,c_i}, \hat{f}_{0,c_i}, \hat{f})$ for each c_i and then considered the following criterion

$$c^* = \text{argmax}_{c_i : i=1,\ldots,I} \mathcal{L}(c_i) \tag{7}$$

for $\mathcal{L}(c) = \sum_{z_j:|z_j|\le c} \log \frac{\hat{f}_{0,c_i}(z_j)}{\hat{f}(z_j)}$.

Note that f_0 and p_0 are estimated for different intervals $[-c_i, c_i]$, so in fact, $\hat{f}_0$ and $\hat{p}_0$ depend on a given interval, say $\hat{f}_0 = \hat{f}_{0,c_i}$ and $\hat{p}_0 = \hat{p}_{0,c_i}$ corresponding to $[-c_i, c_i]$ while $\hat{f}$ is estimated with the whole data one time. We suppress c_i in $\hat{f}_{0,c_i}$ and $\hat{p}_{0,c_i}$ for notational simplicity. For the criterion in (7) to be working well, it is necessary that $\hat{f}_0 = \hat{f}_{0,c_i}$ for different $c_i < c$ should be similar so that $\mathcal{L}(c_i)$ is monotone increasing in $c_i < c$. Figure 1 illustrates unstable estimates of p_0 and f_0 for different intervals from simulated data.

[[Figure 1]]

The broken curve in Figure 1 represents the estimate of f while the solid curves represent estimates of $\hat{f}_0$ corresponding to different intervals $[-c_i, c_i]$. There are clear differences among $\hat{f}_0$s especially around the peak, however, $\hat{f}_{0,c_i}$ tends to be close to each other as z moves further away from the center. In high-peaked data, there are condensed data around the center which may lead to large differences in the criterion based on $\mathcal{L}(c_i)$ in (7) although there are slight differences among $\hat{f}_0$s for different c_is. Thus, we consider some adjustment which can avoid such unstable behaviors of $\mathcal{L}(c_i)$ in (7) for the case of high-peaked data.

In practice, differentiated genes from f_1 are clearly distinguished from those from ϕ_1 with $\sigma_1(< \sigma_2)$ and are mixed with data from ϕ_2. From this point, the value c in (5) is located outside of ϕ_1 with $\sigma_1(< \sigma_2)$ which are the part of the data where the high peak is located. To reflect this idea, gene expression z_i with $|z_i| > 2\sigma_1$ or $|z_i| > 3\sigma_1$ is assumed to be generated from either ϕ_2 or f_1. In terms of avoiding instability of $\mathcal{L}(c_i)$ due to unstable estimate of $\hat{f}_0$ for different c_is, it seems to be reasonable to exclude the condensed high-peaked section of the data in $\mathcal{L}(c_i)$ which is the main idea of overcoming the difficulty in selecting c. Based on this idea, we propose two algorithms called **Fixed a** and **Flex a** with a smoothed curve of $\mathcal{L}(c_i)$. Additionally, we also propose another method called **Hybrid** where we use an initial estimate of $[-c, c]$ using some known method such as goodness-of-fit(GoF) in Park et al. (Park *et. al.*,2011) and attempt to improve upon it by using the criterion in (7). The first proposed algorithm is based on only removing the data around spiky part. Given that the component is normal, it is considered that $[-a, a]$ contains about 95 percent of elements coming from this component, where $a = 2\hat{\sigma}_1$ ($\hat{\sigma}_1 < \hat{\sigma}_2$). Thus, these elements do not relevantly contribute to the criterion. Based on this, a modified criterion was constructed as

$$c^* = \text{argmax}_{c_i:i=1,\ldots,I} \left(\mathcal{L}(c_i) - \mathcal{L}(a)\right) \tag{8}$$

where a is constant for all c_i. Based on the previous sections, we propose the following algorithm for **Fixed a** as follows:

Algorithm 1 (Fixed a):

Step 1: Obtain an initial estimate for $\sigma_1^2(< \sigma_2^2)$, the variance of the spikier component of the null distribution using EM algorithm on some sufficiently large c such that $[-c, c]$ contains 80 percent of the data.

Step 2: Let $a = 2\hat{\sigma}_1$ where $\hat{\sigma}_1$ is from Step 1.

Step 3: Let $c_1 = a$ and take $c_1 < c_2 < \cdots < c_I$ so that $[-c_i, c_i]$ contains a fixed number of data more than $[-c_{i-1}, c_{i-1}]$.

Step 4: Do EM estimation of parameters in $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2, \eta)$ for the data subset in $[-c_i, c_i]$ and compute $\mathcal{L}(c_i) - \mathcal{L}(a)$.

Step 5: Choose $c^* = \text{argmax}_{c_i:i=1,\ldots,I}(\mathcal{L}(c_i) - \mathcal{L}(a))$ and the corresponding estimator of $\boldsymbol{\theta}$.

In **Fixed a** , the main idea is that $\mathcal{L}(c_i)$ should have a monotone increasing pattern in $c_i \in [0, 2\hat{\sigma}_1]$ from $\frac{f_0(z)}{f(z)} > 1$ for $|z| < 2\sigma_1$ however, we actually estimate all the parameters in f_0 for different c_is. $\mathcal{L}(c_i)$ for

different c_i values are computed based on different $\hat{f}_0$s of which estimation errors may cause instability of $\mathcal{L}(c_i)$. From the instability of $\hat{f}_0$, the monotonicity of $\mathcal{L}(c_i) = \sum_{z_j:|z_j|<c_i} \frac{\hat{f}_0(z_j)}{\hat{f}(z_j)}$ is not guaranteed for $c_i < c$. **Fixed a** is designed for eliminating the instability by removing the part of high-peaked portion of the data around the center of the whole data. The choice of a in Step 2 in **Fixed a** is ad-hoc, so there may be a room for improvement. We consider a more flexible procedure which uses more values of a rather than a fixed value of a. Furthermore, there is another problematic phenomenon in $\mathcal{L}(c_i)$ such that multiple small bumps in the plot of $\mathcal{L}(c_i)$ leading to several local maximums. Ideally, we should have one maximum of $\mathcal{L}(c)$, however this does not happen due to estimation errors in $\hat{f}_0$ and $\hat{f}$. To remedy this, we consider a smoothed curve of $\mathcal{L}(c)$ by ignoring small variations of $\mathcal{L}(c)$ and detect a single global maximum from the smoothed curve. For these two purposes such as (i) more flexible a to remove spiky part and (ii) detect a single global maximum by removing small variations, a is treated as a tuning parameter and smoothed curve of $\mathcal{L}(c)$ is fitted for each a. Considering these two aspects, we propose another Algorithm, **Flex a** as follows:

Algorithm 2 (Flex a):

For a given K, we take $a_1 < \ldots < a_K$.

Step 1: For a given a_k, compute $\mathcal{L}(c_i) - \mathcal{L}(a_k)$ for all $c_i > a_k$, $i = 1, ...I$ using **Fixed a**.

Step 2: Fit $(c_i, \mathcal{L}(c_i) - \mathcal{L}(a_k))$ for $c_i > a_k$ using quartic (4th order) polynomial regression.

Step 3: Identify the number of local maximums of the fitted polynomial regression in Step 2.

Step 4: If both of the following conditions are true, stop and accept the result in Step 1. Otherwise, repeat Step 1 using a_{k+1}.

- Condition 1: There is exactly one maximum from the fitted regression curve located in the range of the c_i
- Condition 2: At least one value of the first derivative of the quartic polynomial at the first few smallest cutoff values is positive.

The **Flex a** approach seeks to choose an **a** that ignores enough of the data around the peak of the distribution to make $\mathcal{L}(c)$ stable. By fitting a quartic polynomial, the fitted curve becomes a proxy for the behavior of the criterion across different values of c. The quartic polynomial is selected since it is flexible enough to accommodate both desirable and undesirable behaviors of the curve. If Condition 1 is false, then it means that the criterion goes up again within the range of c_i after reaching a first maximum point, which is contrary to the intended behavior of $\mathcal{L}(c)$ with respect to c. If Condition 2 is false, then it means that $\mathcal{L}(c)$ does not start in an increasing state, which is again contrary to its intended behavior. Thus, if either condition is false, then it means that the plot is not behaving as it should if enough of the bias from $\mathcal{L}(c)$ has been removed by a. Thus, in this case, the algorithm proceeds to repeating the process with an increased a. In essence, **Flex a** is based on removing some part of the condensed high-peaked part adaptively.

A demonstration of how the algorithm adjusts a until achieving the intended behavior of $\mathcal{L}(c)$ is shown in the Figure 2.

The Figure 2 shows the plot for a specific simulated data set (seed=2) at different values of a selected, where a_k is selected as some factor of $\hat{\sigma}_1$ for $\hat{\sigma}_1 < \hat{\sigma}_2$ such as $\{1, 1.25, 1.5, 1.75, 2, 2.5\}$. As shown from Figure 2, at $a = 2\hat{\sigma}_1$ ($\hat{\sigma}_1 < \hat{\sigma}_2$), the plot shows two visible local maximums within the range of data which are not desirable. By increasing a_k, the plot with $a = 2\hat{\sigma}_1$ shows the fitted line shows one maximum which we actually expect in the sense that we have one maximum. In the example of Figure 2, when we observe the phenomenon from $2\hat{\sigma}_1$ from the given grid points of a_k, **Flex a** chooses $a = 2\hat{\sigma}_1$ in this example.
[[Figure 2]]

As mentioned, **Fixed a** and **Flex a** should estimate f_0 for different c_is which may cause some unexpected pattern of $\mathcal{L}(c)$ from different estimation errors. To avoid this, we propose another algorithm, called **Algorithm 3 (Hybrid)**, which does not request multiple estimation of f_0 for different c_i values.

Algorithm 3 (Hybrid)

Step 1: Use a conservative interval $[-c_I, c_I]$ and estimate $(\mu_1, \sigma_1, \mu_2, \sigma_2, \eta, p_0)$ using the data in $[-c_I, c_I]$

Step 2: Create equally-spaced grid points $0 < c_1 < \cdots < c_I$ for cutoff values in $[-c_I, c_I]$ chosen in Step 1.

Step 3: Compute the criterion $\mathcal{L}(c_i)$ for $1 \leq i \leq I$ using Algorithm 1 with $a = 0$.

Step 4: Select the cutoff value with maximum criterion value as the best cutoff, i.e., $c^* = \text{argmax}_{1 \leq i \leq I} \mathcal{L}(c_i)$

The rationale behind **Hybrid** is that since $(\mu_1, \sigma_1, \mu_2, \sigma_2)$ in f_0 and (η, p_0) are estimated only once based on $[-c_I, c_I]$ chosen in Step 1 using the GoF approach (Park *et. al.*,2011), so the variation of $\mathcal{L}(c)$ from differently estimated parameters for different c_is can be eliminated while **Fixed a** and **Flex a** include such variations. Thus, in **Hybrid**, a can be safely set to be zero and the expected behavior of criterion values versus cutoff values will be observed. The initial choice of $[-c_I, c_I]$ in Step 1 should be conservative in the sense that the interval can include the true interval $[-c, c]$ for the Zero assumption in (5) so that $[-c, c]$ can be included in $[-c_I, c_I]$. However, we do not want to take an arbitrarily large interval, so we use GoF which tends to select a wider interval than the true interval in the Zero assumption. **Hybrid** attempts to improve upon the conservative result of the GoF approach by re-detecting the change point using **Fixed a**.

4 Simulation studies

In this section, we present simulation studies to compare the performance of the three proposed algorithms (**Fixed a**, **Flex a**, and **Hybrid**). We also provide results using the GoF approach in Park *et al.* (2011) and Efron's local FDR procedure (Efron, 2004) for comparison. Each of these methods is evaluated across different sets of null and alternative distribution configurations. Across all settings, the means for the null distributions are set to be zero to reflect the biological assumption on the non-interesting genes. If the ratio of σ_1 and σ_2 is large, it is expected to have a very high peak around the center of the data. In each setting, we generate $8,000$ data and $p_0 = 0.9$ of them are generated from f_0 while the rest are generated from f_1, i.e. $7,200$ and 800 are generated from f_0 and f_1, respectively. η in f_0 is set to be 0.6.

As alternative distributions, we consider several different settings: (i) a truncated t-distribution with 20 degrees of freedom and truncation excluding the region $[-0.5, 0.5]$, (ii) a pair of shifted truncated t-distributions shifted by ± 1.5 and likewise truncated to exclude the region $[-0.5, 0.5]$, and (iii) a mixture of gamma(shape=2, rate=1) and -gamma(shape=2, rate=1). We evaluate the performance of each method using false discovery rate (FDR) and true positive rate (TPR). Since we simulate each setting $1,000$ times, we define the empirical FDR and TPR as follows:

$$FDR = \frac{1}{1,000} \sum_{i=1}^{1,000} \frac{V_i}{R_i} I(R_i > 0),\ TPR = \frac{1}{1,000} \sum_{i=1}^{1,000} \frac{S_i}{800} \tag{9}$$

where 800 in TPR is the number of the true alternative from f_1, and V_i, R_i and S_i are the numbers of falsely rejected hypotheses, the number of rejected hypotheses and the number of true positives from ith simulated data set, respectively.

The succeeding simulation results are divided into two subsections. In the first subsection, the focus is on examining the performance of the different methods under the alternative distribution described in (i). As such, both σ_1 and σ_2 were varied. In the second subsection, the focus is on examining the performance and parameter estimation of the methods under different alternative distributions described in (ii) and (iii) specifically in the context of data similar to the Set4Δ mutant dataset. Thus for this subsection, σ_1 is kept

Table 1
FDR and TPR estimates for each method across various null distributions $0.6N(0,\sigma_1^2) + 0.4N(0,\sigma_2^2)$. Alternative distribution is t distribution truncated at ± 0.5. The number in (.) represents standard deviation.

Null distribution	σ_1^2 / σ_2^2	0.02 / 0.1		0.02 / 0.08		0.02 / 0.05		0.08 / 0.1		0.06 / 0.1		0.01 / 0.1		0.005 / 0.1	
$\widehat{FDR}$	Fixed a	0.0187	(0.0168)	0.0042	(0.0021)	0.0004	(0.0003)	0.0000	(0.0000)	0.0000	(0.0000)	0.0839	(0.0392)	0.2541	(0.1125)
	Flex a	**0.0427**	(0.0243)	**0.0285**	(0.0166)	0.0086	(0.0050)	0.0017	(0.0023)	**0.0059**	(0.0039)	**0.0453**	(0.0605)	0.0130	(0.0543)
	Hybrid	0.0036	(0.0107)	0.0039	(0.0077)	0.0076	(0.0072)	0.0030	(0.0044)	0.0057	(0.0067)	0.0052	(0.0194)	**0.0051**	(0.0330)
	GoF	0.0007	(0.0021)	0.0009	(0.0020)	0.0000	(0.0004)	0.0006	(0.0019)	0.0008	(0.0025)	0.0008	(0.0019)	0.0013	(0.0021)
	Efron's locfdr	0.2784	(0.0242)	0.1943	(0.0243)	**0.0673**	(0.0154)	**0.0043**	(0.0035)	0.0095	(0.0053)	0.5032	(0.0218)	0.6673	(0.0107)
$\widehat{TPR}$	Fixed a	0.6028	(0.0723)	0.6033	(0.0917)	0.6717	(0.0751)	0.2039	(0.0582)	0.2452	(0.0341)	0.7298	(0.1362)	0.8329	(0.1629)
	Flex a	**0.7165**	(0.0949)	**0.7693**	(0.0591)	0.8403	(0.0304)	0.4945	(0.0315)	**0.5819**	(0.0259)	**0.6522**	(0.1634)	0.4547	(0.1575)
	Hybrid	0.4588	(0.1056)	0.5749	(0.1082)	0.7602	(0.1665)	0.4752	(0.1153)	0.5182	(0.1283)	0.4829	(0.1046)	**0.5242**	(0.0679)
	GoF	0.3998	(0.0787)	0.4830	(0.1082)	0.1312	(0.1845)	0.2576	(0.2146)	0.2838	(0.2117)	0.4288	(0.0633)	0.4832	(0.0491)
	Efron's locfdr	0.9732	(0.0143)	0.9693	(0.0149)	**0.9719**	(0.0149)	**0.5192**	(0.0258)	0.5754	(0.0265)	1.0000	(0.0000)	1.0000	(0.0000)

constant at 0.02 while σ_2 is varied. Figure 3 presents the setup of null and alternative distributions for the different settings.

4.1 Primary Simulation Outcomes

The first set of simulations vary both the larger and the smaller variance of the null distributions in order to get a general idea of the comparative performance of the different methods. These settings use (i) truncated t-distribution as the alternative distribution and they are visualized in Figure 3(A). Table 1 shows the FDR and TPR defined in (9) from $1,000$ replications under each setting. We use a nominal level of FDR=0.05 throughout the simulations. As expected, Efron's local FDR procedure based on a single normal distribution is not able to control FDR in most of settings where σ_1 and σ_2 are significantly different. This is because Efron's local FDR recognizes ϕ_1 ($\sigma_1 < \sigma_2$) as a null distribution and tends to reject many data from ϕ_2 falsely which leads to the failure of controlling a given level of FDR. The Efron's FDR approach is expected to control a given FDR for the case of similar values of σ_1 and σ_2 under which a single normal distribution behaves similar to the mixture of ϕ_1 and ϕ_2. The **Fixed a** approach is not able to control a given level of FDR for some of the settings since a pre-fixed a such as $a = 2\hat{\sigma}_1$ is not expected to cover universal cases. On the other hand, **Flex a** achieves the closest empirical FDR to the nominal level 0.05 among all procedures. As expected, the Goodness-of-fit (GoF) approach produces the most conservative FDRs. The **Hybrid** approach is a modification of the GoF approach that seeks to correct its over-conservative nature by using the changepoint criterion to select a less conservative c. As shown in Table 1, it is able to do this in each setting, producing a higher FDR each time that is still below the nominal α level.

In terms of empirical TPR, Table 1 shows that Efron's local FDR achieves the highest TPRs across all settings. However, it fails in controlling the given FDR in almost every case, only being able to do so when the variances of the two null distribution components are close enough together such that the null distribution better resembles a single normal. Thus, it is not meaningful to consider the Efron's local FDR for comparison of TPR. Since the **Fixed a** method is also unable to control FDR for some cases, it is also excluded in the TPR comparison. Among the remaining methods, **Flex a** yields the highest TPRs in general and is almost always better than the **Hybrid** in terms of TPR across all settings. The **Hybrid** produced a better TPR than the **Flex a** only in the setting where the $\sigma_1^2 = 0.005$ and $\sigma_2^2 = 0.10$. This observation is further explored in another set of simulations. Also as expected, the GoF approach has the lowest TPRs and the **Hybrid** approach tends to improve the TPR of GoF approach. This improvement is observed to be limited when the variance components of the null are farther apart but becomes larger as those components are set to be closer to each other. [[Figure 3]]

4.2 Comparison between alternative distributions and parameter estimation

In this section, we present more simulation studies to demonstrate the behavior of the different approaches under different alternative distributions. The smaller variance is set at $\sigma_1^2 = 0.02$ while σ_2^2 is varied. In the previous section, we used the f_1 which is the truncated t distribution with the support $(-\infty, c) \cup (c, +\infty)$. In this section, we consider a shifted t distribution with truncation. More specifically, when f_1 is the conditional density of $z = \pm 1.5(t)$ given $|z| > 0.5$. The Zero assumption is still satisfied with this f_1. This setup is represented by Figure 3(C)

In addition to this, we are also interested in how the Zero assumption in (5) affects results when the assumption is actually not satisfied by the setup. For this, f_1 is modeled based on gamma distribution and random sign (± 1) leading to $z_i = (2x_i - 1)y_i$ where $X \sim Bernoulli(0.5)$, $Y \sim gamma(2, 1)$ and x_i and y_i are independent. We call this as a $\pm gamma(2, 1)$. There is no truncation here for f_1, so c in the Zero assumption is simply 0 which means there does not exist an interval $[-c, c]$ defined in (5). This setup is represented in Figure 3(B).

Results for the shifted truncated t-distribution alternative are shown in Table 2. Consistent with Table 1, the empirical FDRs are shown to be controlled for the **Flex a**, **Hybrid**, and GoF approaches across null distribution settings as σ_2^2 is decreased from 0.12 to 0.05 while the Efron's local FDR fails in controlling a given FDR. We also have the results of TPRs similar to Table 1 in the sense that **Flex a** obtains the highest TPRs for three cases and the Efron's local FDR approach has the best TPRs with somewhat inflated empirical FDR when σ_1 and σ_2 are similar (e.g., $\sigma_1^2 = 0.02$ and $\sigma_2^2 = 0.05$). The **Hybrid** improves the GoF approach, but it is still very limited improvement. In terms of parameter estimation, Table 2 shows that σ_1^2 is estimated well across the three methods (**Flex a**, **Hybrid**, GoF). On the other hand, σ_2^2 was found to be overestimated by GoF and underestimated by the **Flex a** approach. The **Hybrid** comes closest to estimating the null distribution parameters, however, in terms of estimating c, **Flex a** comes much closer than the **Hybrid** which overestimates of c. Considering the superiority of the **Flex a** approach over the **Hybrid** approach, the estimation of c is more important than estimation of (σ_1^2, σ_2^2). Since the **Hybrid** approach inherits the estimate of c from the GoF approach, the **Hybrid** approach should be much more conservative compared to the **Flex a** approach unless there is a much better initial estimate of c.

Table 3 shows parallel results for the $\pm$ gamma(2,1) as f_1. Although there does not exist an interval $[-c, c]$ in (5), the results are still consistent with those in Table 2, with the main difference being a loss in TPR across all methods. Compared to the truncated f_1 used in Table 1 and 2, $\pm$ gamma(2,1) is more overlapped with f_0 which leads to the loss of powers of testing procedures. From Table 3, even when there exist slightly contaminated data from f_1 around the center of the whole data. the proposed approaches with the Zero assumption in (5) is robust to such contamination from f_1.

4.3 Robustness under model missspecification

We have discussed the cases that the null distribution consists of two normals and one of them represents the spiky distribution. In practice, there may be some situation that such a spiky distribution also consists of more than one normal distribution. In this section, we consider the cases that the null distribution is a mixture of more than two normal distributions. As shown in the simulations in the previous section, Efron's local fdr based on a single normal for the null distribution performs well when the null distribution is a mixture of two normals with similar variances. From this, when the null distribution is a mixture of more than two normals and some of them have similar variances, the mixture of two normal distributions is expected to be robust to the number of components in the mixture model. To deomstrate this, we conduct some simulations where the null distribution is a mixture of four normals instead of two such as $f_0 = 0.3(N(0, \sigma_1^2) + N(0, \sigma_2^2)) + 0.2(N(0, \sigma_3^2) + N(0, \sigma_4^2))$. We considere two cases: the first one is $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.005, 0.05, 0.10, 0.15)$ which is the case that two components represent spiky distributions and the other two components represent non-spiky distributions. The second case is $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.11, 0.13, 0.1, 0.3)$ which is the case that all of them have similar variances. The alternative distribution used is the same as that in Table 2. Table 4 shows that the results across different

Table 2
Comparison of results using different methods across various null distributions $0.6N(0, 0.02) + 0.4N(0, \sigma_2^2)$ with alternative distribution set as $t \pm 1.5$ truncated at ± 0.5. The number in (.) represents standard deviation

Null	σ_1^2	0.02		0.02		0.02		0.02	
Distribution	σ_2^2	0.12		0.1		0.08		0.05	
$\widehat{FDR}$	Flex a	**0.0584**	(0.0592)	**0.0525**	(0.0376)	**0.0288**	(0.0146)	0.0046	(0.0030)
	Hybrid	0.0050	(0.0135)	0.0052	(0.0105)	0.0040	(0.0048)	0.0032	(0.0031)
	GoF	0.0015	(0.0025)	0.0014	(0.0024)	0.0016	(0.0022)	0.0009	(0.0014)
	Efron's locfdr	0.3532	(0.0246)	0.2788	(0.0265)	0.1903	(0.0262)	**0.0615**	(0.0154)
$\widehat{TPR}$	Flex a	**0.8188**	(0.0942)	**0.8672**	(0.0689)	**0.8974**	(0.0318)	0.9159	(0.0140)
	Hybrid	0.6945	(0.0624)	0.7534	(0.0571)	0.8188	(0.0358)	0.8889	(0.0539)
	GoF	0.6669	(0.0479)	0.7194	(0.0442)	0.7830	(0.0433)	0.7488	(0.2127)
	Efron's locfdr	0.9870	(0.0076)	0.9842	(0.0082)	0.9818	(0.0087)	**0.9820**	(0.0087)
$\widehat{\sigma_1^2}$	Flex a	0.0251	(0.0111)	0.0229	(0.0051)	0.0235	(0.0023)	0.0236	(0.0045)
	Hybrid	0.0211	(0.0056)	0.0216	(0.0075)	0.0203	(0.0062)	0.0212	(0.0031)
	GoF	0.0207	(0.0014)	0.0204	(0.0013)	0.0196	(0.0014)	0.0210	(0.0034)
$\widehat{\sigma_2^2}$	Flex a	0.0832	(0.0365)	0.0631	(0.0232)	0.0497	(0.0094)	0.0399	(0.0023)
	Hybrid	0.1369	(0.0298)	0.1068	(0.0254)	0.0760	(0.0137)	0.0515	(0.0255)
	GoF	0.1514	(0.0244)	0.1239	(0.0213)	0.0930	(0.0200)	0.1666	(0.1872)
$\widehat{\eta}$	Flex a	0.5427	(0.0390)	0.5375	(0.0204)	0.5332	(0.0117)	0.5195	(0.0314)
	Hybrid	0.6075	(0.0420)	0.5878	(0.0525)	0.5520	(0.0350)	0.5551	(0.0741)
$\widehat{c}$	Flex a	0.5768	(0.1643)	0.4898	(0.1257)	0.4470	(0.0658)	0.5291	(0.1158)
	Hybrid	0.7676	(0.1277)	0.7244	(0.1163)	0.6360	(0.1061)	0.5787	(0.1839)

Table 3
Comparison of results using different methods across various null distributions $0.6N(0, 0.02) + 0.4N(0, \sigma_2^2)$ with alternative distribution set as $\pm$gamma(2,1). The number in (.) represents standard deviation

Null	σ_1^2	0.02		0.02		0.02		0.02	
Distribution	σ_2^2	0.12		0.1		0.08		0.05	
$\widehat{FDR}$	Flex a	**0.0661**	(0.0610)	**0.0501**	(0.0406)	**0.0303**	(0.0151)	0.0038	(0.0029)
	Hybrid	0.0082	(0.0201)	0.0060	(0.0117)	0.0035	(0.0045)	0.0032	(0.0029)
	GoF	0.0018	(0.0030)	0.0016	(0.0025)	0.0016	(0.0020)	0.0009	(0.0014)
	Efron's locfdr	0.3998	(0.0278)	0.3163	(0.0310)	0.2141	(0.0309)	**0.0641**	(0.0180)
$\widehat{TPR}$	Flex a	**0.7509**	(0.0817)	**0.7766**	(0.0671)	**0.8123**	(0.0314)	0.8243	(0.0179)
	Hybrid	0.6534	(0.0560)	0.6941	(0.0459)	0.7390	(0.0292)	0.8156	(0.0308)
	GoF	0.6250	(0.0384)	0.6646	(0.0356)	0.7143	(0.0332)	0.7522	(0.1349)
	Efron's locfdr	0.8709	(0.0155)	0.8674	(0.0156)	0.8638	(0.0160)	**0.8624**	(0.0162)
$\widehat{\sigma_1^2}$	Flex a	0.0276	(0.0126)	0.0236	(0.0067)	0.0240	(0.0029)	0.0237	(0.0045)
	Hybrid	0.0218	(0.0071)	0.0219	(0.0080)	0.0201	(0.0049)	0.0215	(0.0019)
	GoF	0.0207	(0.0016)	0.0204	(0.0015)	0.0197	(0.0014)	0.0200	(0.0027)
$\widehat{\sigma_2^2}$	Flex a	0.0788	(0.0342)	0.0663	(0.0245)	0.0500	(0.0098)	0.0418	(0.0034)
	Hybrid	0.1238	(0.0295)	0.0996	(0.0221)	0.0761	(0.0106)	0.0457	(0.0126)
	GoF	0.1399	(0.0221)	0.1158	(0.0187)	0.0889	(0.0156)	0.1113	(0.2297)
$\widehat{\eta}$	Flex a	0.5334	(0.0378)	0.5346	(0.0223)	0.5312	(0.0126)	0.5222	(0.0311)
	Hybrid	0.5865	(0.0442)	0.5694	(0.0471)	0.5459	(0.0254)	0.5345	(0.0362)
$\widehat{c}$	Flex a	0.5772	(0.1606)	0.5140	(0.1452)	0.4410	(0.0815)	0.5238	(0.1177)
	Hybrid	0.7215	(0.1412)	0.6883	(0.1162)	0.6221	(0.0995)	0.5138	(0.1065)

Table 4
Comparison of results where the null is a mixture of four Normal distributions $0.3(N(0,\sigma_1^2)+N(0,\sigma_2^2))+0.2(N(0,\sigma_3^2)+N(0,\sigma_4^2))$ with alternative distribution set as $t \pm 1.5$ truncated at ± 0.5. The number in (.) represents standard deviation

Null Distribution	σ_1^2, σ_2^2 / σ_3^2, σ_4^2	0.005 , 0.05 / 0.10 , 0.15		0.11 , 0.13 / 0.1, 0.3	
$\widehat{FDR}$	Flex a	**0.0194**	(0.0314)	**0.0223**	(0.0153)
	Hybrid	0.0069	(0.0109)	0.0196	(0.0195)
	GoF	0.0056	(0.0040)	0.0127	(0.0146)
	Efron's locfdr	0.5796	(0.0200)	0.4503	(0.0224)
$\widehat{TPR}$	Flex a	**0.7306**	(0.0781)	**0.5307**	(0.0710)
	Hybrid	0.7101	(0.0809)	0.5046	(0.0935)
	GoF	0.7160	(0.0256)	0.4276	(0.1624)
	Efron's locfdr	0.8735	(0.0102)	0.9248	(0.0067)

methods. As in previous simulations, **Flex a**, **Hybrid**, and GoF are all able to retain control over FDR under this missspecification of the null distribution, and that **Flex a** remains superior to the other two in terms of TPR. Efron's local FDR procedure continues to suffer from misspecification and is unable to control FDR in each case.

5 Real Data : Analysis of Set4Δ Mutant Dataset

The study is interested in the identification of genes within the genome of the budding yeast *Saccharomyces cerevisiae* that are differentially expressed when key gene expression machinery is mutated. Specifically, this work focused on a yeast protein, known as Set4, which was recently identified as a chromatin regulator and there are many open questions about its functions (Tran *et al.*,2018). Previous work demonstrated that the gene expression regulator Set4 is important for protecting cells during oxidative stress induced by hydrogen peroxide treatment and activates stress response genes to help cells survive during stress. However, the full complement of genes controlled by Set4 is not yet determined. In order to identify the genes dependent on Set4 during stress, the gene expression profile of the whole genome in cells where Set4 is absent (knock-out condition, Set4Δ) and cells where it is present (wildtype condition) are determined. In addition, cells are treated with hydrogen peroxide to induce oxidative stress. Approximately 8000 RNA species are evaluated using this method, each represented by a weighted count value under knock-out and wildtype conditions. From among the 8000 RNA species (about 6000 coding genes and 2000 non-coding genes), the objective is the identification of "interesting" genes based on values of the statistic $log_2 \frac{\bar{KO}}{\bar{WT}}$, where $\bar{WT}$ is a mean weighted count measure of mutations under wild type condition and $\bar{KO}$ is a mean weighted count measure of mutations under knock-out condition. The counts are weighted based on the length of each gene and the total number of reads and so is a continuous value rather than a discrete one. A gene is considered "interesting" if the weighted count measure is very different between the KO and WT settings, and so correspondingly, if $\mid log_2 \frac{\bar{KO}}{\bar{WT}} \mid \gg 0$. For each gene, the experiment is replicated thrice and the resulting weighted counts are averaged. Afterward, the $log_2 \frac{\bar{KO}}{\bar{WT}}$ is computed, thus yielding a single statistic for each gene. A basic histogram of the logfold change statistics is shown in Figure 4.
[[Figure 4]]

Figure 4 shows that the distribution of the test statistics has a very high peak but still have prominent tails. This is consistent with cellular biology as cells often have very specific responses to genetic or chemical disruptions. Thus, the majority of the genes in the genome will not show differential expression

Table 5 Set4Δ Mutant Data Results

	# of interesting genes	Cutoff c	μ_1	σ_1^2	μ_2	σ_2^2	η
Flex a	505	0.400	0.000	0.004	-0.002	0.042	0.615
Hybrid	328	0.664	-0.001	0.005	-0.003	0.082	0.668
GoF	264	0.851	-0.020	0.116	0.000	0.006	0.299
Efron's locfdr	1348	NA	-0.001	0.006	NA	NA	NA

following the perturbance compared to standard conditions. However, there is also often a subset of genes that show small changes, most likely due to indirect consequences of genetically or chemically perturbing the cell. Neither of these groups of cells are of interest and so both belong to the null distribution. Only the genes that are highly differentially expressed are considered to be directly affected by the genetic or chemical change, and are of the most interest to the biologist investigating the specific outcome of a given cellular perturbance. Thus, it is reasonable to assume that the null distribution is a mixture of two components. For the Set4Δ mutant dataset, there are 8630 simultaneous tests conducted. For estimating f via splines, the most extreme 30 values of the logfold change statistic needed to be removed for estimation to work. An FDR level of $\alpha = 0.05$ was used for all procedures. **Flex a**, **Hybrid**, and the GoF approach were used. Table 4 shows a summary of the results. Results of applying Efron's local FDR are also presented. As expected, using just the local FDR approach leads to a very liberal set of interesting cases (rejections). However, the assumption of a single normal null distribution is contrary to biological grounding, and having over 1300 rejections is not realistic. As expected, the goodness-of-fit approach is the most conservative. The Hyrbid approach is less conservative than goodness-of-fit but not by much since it uses goodness-of-fit as a baseline. The **Flex a** approach gives the most liberal results while still controlling FDR. Figure 5 shows how each test, **Flex a**, **Hybrid**, and GoF, identifies interesting genes. It is shown that the number of significant genes selected by different methods have similar patterns to those from simulation studies. In particular, the sets of the selected genes are nested in the sense that all of the genes identified by GoF and **Hybrid** are also identified by **Flex a** etc. Considering the simulation studies, GoF and **Hybrid** approches are considered to be fairly conservative and Efron's local fdr fails in controlling a given level of FDR. On the other hand, the result from **Flex a** in Set4Δ mutant data is believed to be the most reasonable among all the methods in terms of controlling a FDR and obtaining significant genes.

[[Figure 5]]

6 Conclusion

Different change-point detection based procedures that were constructed were able to control FDR in the setting where the null distribution is a highly peaked mixture of two normal distributions, where one component has a very small variance. Simulation studies showed how in this situation, the standard local FDR procedure fails at controlling type 1 error. Furthermore, accommodating for the issue by modeling the null distribution as a mixture of two normal distributions using a goodness-of-fit approach was shown to produce overly conservative outcomes. Novel methods developed to address these problems applied changepoint detection to determine the largest interval in the dataset that still comes purely from the null distribution. Estimating the ormal null mixture from this interval was found to retain control over FDR while producing more competitive power than the goodness-of-fit approach.

Of the two viable changepoint detection procedures constructed, the **Flex a** approach, which used a tuning parameter in order to remove an unstable portion of the dataset, was found to be superior in terms of TPR in almost every simulation setting used compared to the **Hybrid** approach which used the result from goodness-of-fit as a baseline and then applied changepoint detection. When the two components of the null distribution were close enough to make the distribution appear similar to a single normal, the

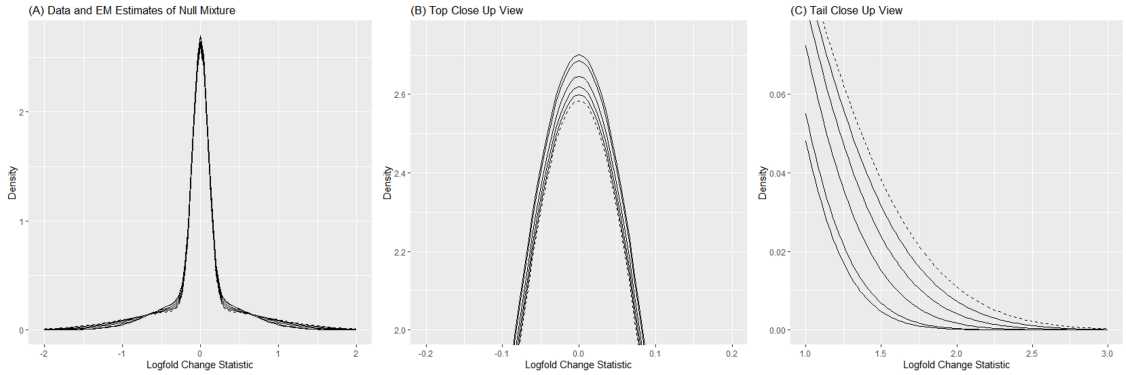

Figure 1 Bias of Estimator in Equation 5. The broken curve represents the estimate of f while the solid curves represent some EM estimates of f_0 at different fixed $[-c, c]$'s. Close-up view is shown on the figure to the right.

standard local FDR method was still able to control FDR and produced the best TPR. This shows that the new approaches are best used in contexts where the null distribution can be soundly assumed as deviating considerably from a single normal distribution.

Finally, application of the different methods on the Set4Δ mutant dataset showed that the methods which accounted for the null distribution being a mixture of two normal distributions rather than a single normal produced more reasonable results. Among these, the **Flex a** method was able to identify the most number of interesting candidates for the investigator to conduct further experimentation on.

Acknowledgements We thank the Associate Editor and the anonymous reviewer for their many constructive comments and suggestions that have improved the paper. This research was supported by the National Research Foundation of Korea funded by the Korea government (MSIT) (no. NRF-2019H1D3A2A02102167, NRF-2020R1A2C1A01100526), the New Faculty Startup Fund from Seoul National University, and the National Institute of Health (no. NIH-R01GM124342) to EMG.

Conflict of Interest

The authors have declared no conflict of interest. (or please state any conflicts of interest)

References

Aickin, M. and Gensler, H.(1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health*, **86**, 726–728.

Benjamini, Y. and Hochberg, Y. (2011).Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*,**57**,289–300

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Chauveau, D., Saby, N., Orton, T., Lemercier, B., Walter, C. and Arrouays, D.(2014). Large-scale simultaneous hypothesis testing in monitoring carbon content from French soil database — A semi-parametric mixture approach.*Geoderma*,**219**,117–124.

Dempster, A.and Laird, N.and Rubin, D.(1977).Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society, Series B*, **39**,1–38.

Deliu, E., Arecco, N., Morandell, J., Dotter, C., Contreras, X. and Girardot, C., Kasper, E., Kozlova, A., Kishi, K., Chiaradia, I., Noh, K., and Novarino, G.(2018).Haploinsufficiency of the intellectual disability gene SETD5 disturbs developmental gene expression and cognition.*Nat Neurosci.*, **21**,1717–1727.

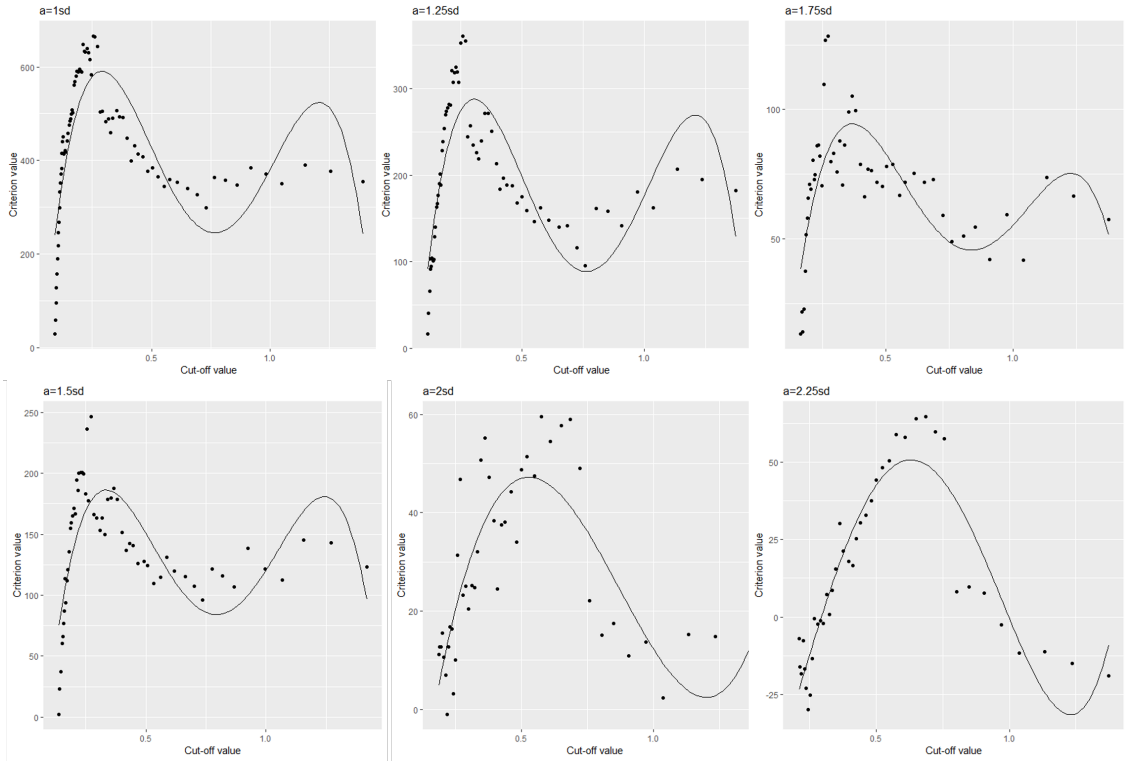

Figure 2 Plot behavior at different a. Increasing incrementally from $\widehat{\sigma_1}$ by $0.25\widehat{\sigma_1}$, the plot is shown to be well-behaved by $a = 2\widehat{\sigma_1}$. Algorithm 2 selects this a and proceeds to using the best c determined from it.

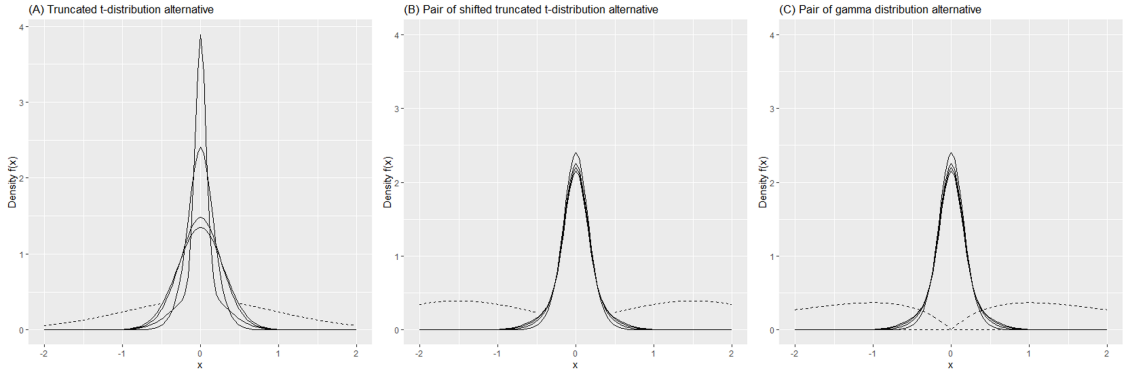

Figure 3 Setup for simulation using different values of σ_1^2 and σ_2^2. and different alternative distributions. Solid lines represent null distributions. The range of null distributions used according to peakedness is represented. Dashed lines are alternative distributions. Only 10% of the simulated data come from the alternative.

Efron, B.(2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**,96–104.

Efron, B., Turnbull, B., Narasimhan, B. and Strimmer, K.(2015). package "locfdr". https://cran.r-project.org/web/packages/locfdr/locfdr.pdf, (Accessed: 2019-5-2)

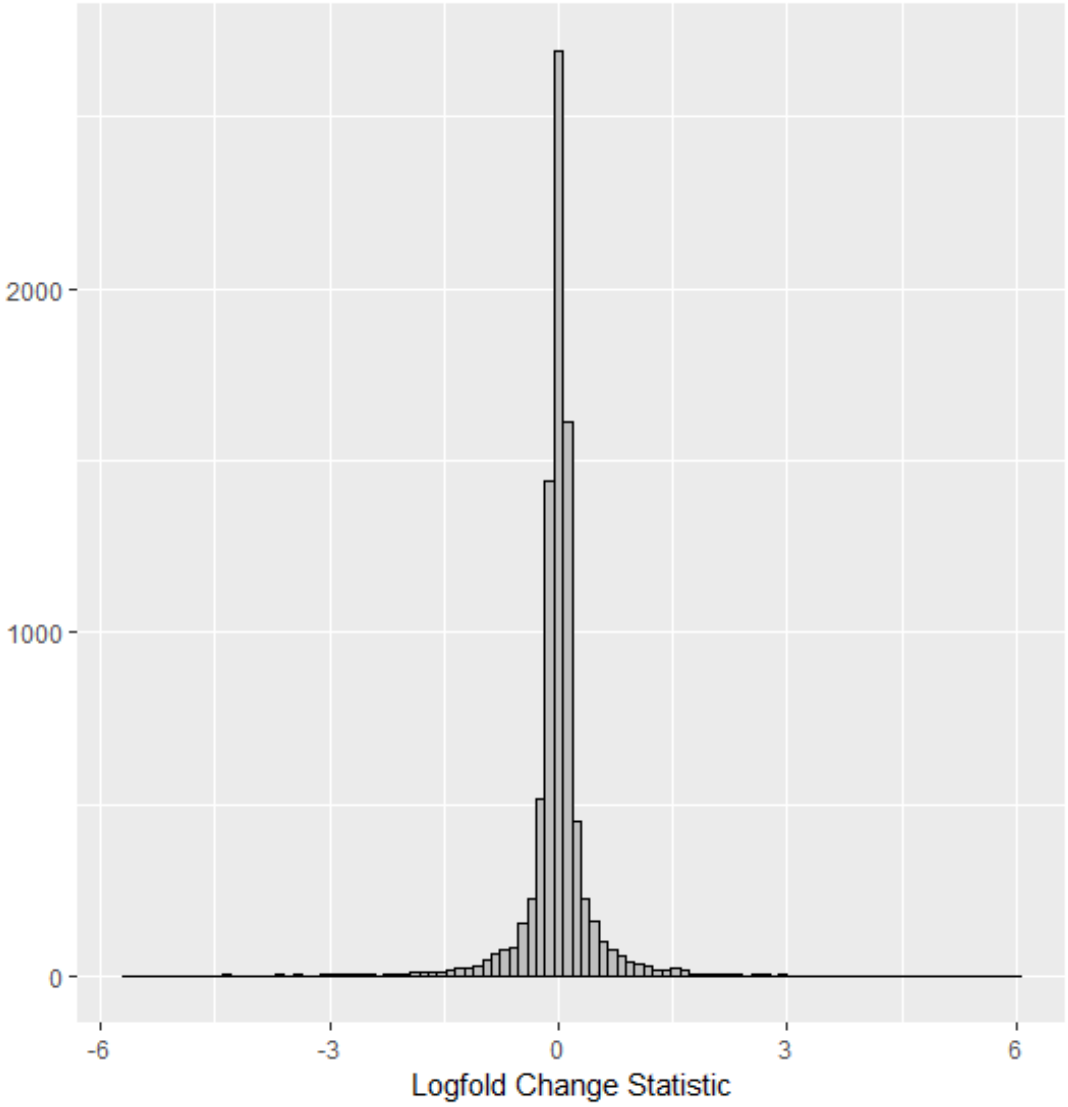

Figure 4 Histogram of Logfold Change Statistic values from the motivating study.

Gauran, I. and Park, J. and Lim, J. and Park, D.and Zylstra, J. and Peterson, T. and Kann, M. and Spouge, J.(2018). Empirical Null Estimation Using Zero-Inflated Discrete Mixture Distributions and its Application to Protein Domain Data.*Biometrics*, **74**,458–471.

Jeong, S.-O., Choi, D. and Jang, W.(2020). A semiparametric mixture method for local false discovery rate estimation, *The Annals of Applied Statistics*, **14**,1242–1257 .

Kammers, R., Cole, R., and Tiengwe, C., and Ruczinski, I.(2015). Detecting significant changes in protein abundance, *EUPA Open Proteomics*, **7**,11–19

Kwong, K., Holland, B. and Cheung, S.(2002).A modified Benjamini–Hochberg multiple comparisons procedure for controlling the false discovery rate. *Journal of Statistical Planning and Inference*, **104**,351–362.

Lee, N., Kim, A., Park, C., and Kim, S.(2016). An improvement on local FDR analysis applied to functional MRI data.*Journal of Neuroscience Methods*, **267**, 115–125.

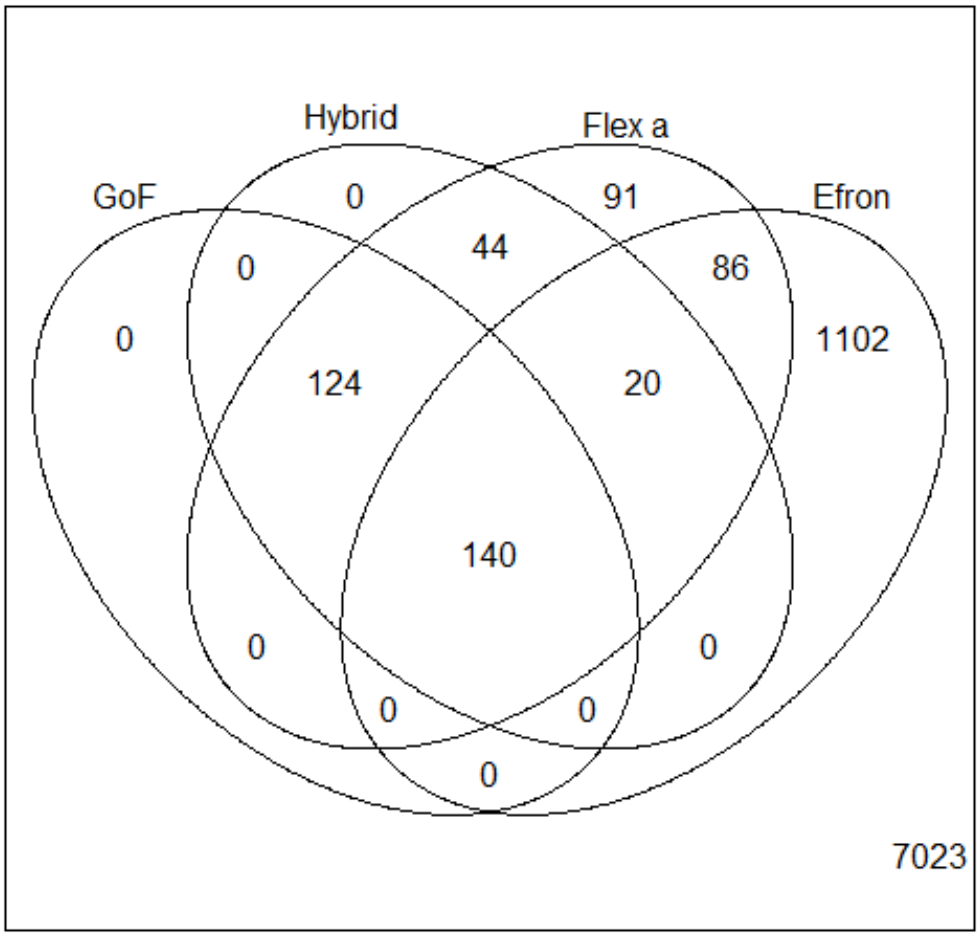

Figure 5 Interesting genes identified by each test. Efron's local FDR identifies the most but control over FDR is doubtful. Flex a identifies every gene identified by any of the remaining methods and is expected to control FDR based on simulations.

McCarthy, D., and Smyth, G.(2009). Testing significance relative to a fold-change threshold is a TREAT.*Bioinformatics*, **25**, 765–771.

Park, D. and Park, J. and Zhong, X. and Sadelain, M.(2011).Estimation of empirical null using a mixture of normal and its use in local false discovery rate.*Computational Statistics Data Analysis* **55**,2421–2432.

Robin, S., Bar-Hen, A., Daudin, J. and Pierre, L.(2007).A semi-parametric approach for mixture models: Application to local false discovery rate estimation,*Computational Statistics Data Analysis*,**51**,5483–5493.

Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families *The Annals of Applied Statistics*, **2**,1332—1359.

Serratore, ND., Baker, KM., Macadlo, LA., Gress, A., Powers, B. and Atallah, N., Westerhouse, K., Hall, M., Weake, V. and Briggs, S.(2018). A Novel Sterol-Signaling Pathway Governs Azole Antifungal Drug Resistance and

Hypoxic Gene Repression in Saccharomyces cerevisiae. *Genetics*, **208**, 1037–1055.

Shaffer, J. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**,561–584.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genome-wide studies, *PNAS*,**100**,9440–9445

Tran, K., Jethmalani, Y., Jaiswal, D. and Green, E. (2018). Set4 is a chromatin-associated protein, promotes survival during oxidative stress, and regulates stress response genes in yeast.*JBC*

Tran, K. and Green, E. (2019).SET domains and stress: uncovering new functions for yeast Set4.*Curr Genet.*, **65**, 643–648.

Tran, K., Jethmalani, Y., Jaiswal, D. and Green, E.(2018).Set4 is a chromatin-associated protein, promotes survival during oxidative stress, and regulates stress response genes in yeast.http://www.jbc.org/content/early/2018/08/06/jbc.RA118.003078.full.pdf,(Accessed: 2019-03-11)

Zhang, X. and Novera, W. and Zhang, Y. and Deng, L.(2017).MLL5 (KMT2E): structure, function, and clinical relevance. *Nat Neurosci.*, **74**,2333–2344.

Zhao, Q. and Small, D. and Su, W. (2018). Multiple Testing When Many p-values are Uniformly Conservative, with Application to Testing Qualitative Interaction in Educational Interventions. *JASA*, **114**,1291–1304.

