Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Detecting Strong Signals in Gene Perturbation Experiments: An Adaptive Approach With Power Guarantee and FDR Control

Leying Guan, Xi Chen & Wing Hung Wong

To cite this article: Leying Guan, Xi Chen & Wing Hung Wong (2019): Detecting Strong Signals in Gene Perturbation Experiments: An Adaptive Approach With Power Guarantee and FDR Control, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1635484

To link to this article: https://doi.org/10.1080/01621459.2019.1635484

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–9: Theory and Methods
https://doi.org/10.1080/01621459.2019.1635484

Detecting Strong Signals in Gene Perturbation Experiments: An Adaptive Approach With Power Guarantee and FDR Control

Leying Guan[a], Xi Chen[a], and Wing Hung Wong[a,b]

[a]Department of Statistics, Stanford University, Stanford, CA; [b]Biomedical Data Sciences, Stanford University, Stanford, CA

ABSTRACT

The perturbation of a transcription factor should affect the expression levels of its direct targets. However, not all genes showing changes in expression are direct targets. To increase the chance of detecting direct targets, we propose a modified two-group model where the null group corresponds to genes which are not direct targets, but can have small nonzero effects. We model the behavior of genes from the null set by a Gaussian distribution with unknown variance τ^2. To estimate τ^2, we focus on a simple estimation approach, the iterated empirical Bayes estimation. We conduct a detailed analysis of the properties of the iterated EB estimate and provide theoretical guarantee of its good performance under mild conditions. We provide simulations comparing the new modeling approach with existing methods, and the new approach shows more stable and better performance under different situations. We also apply it to a real dataset from gene knock-down experiments and obtained better results compared with the original two-group model testing for nonzero effects.

ARTICLE HISTORY
Received July 2018
Accepted June 2019

KEYWORDS
Empirical Bayes; ϵ-contamination; Gene knock-down.

1. Introduction

The transcriptional regulatory networks, formed by transcription factors (TFs) and their targets, are believed to play an important role regulating embryonic stem (ES) cell pluripotency (Niwa et al. 1998, 2000; Chambers and Smith 2004; Loh et al. 2006; Kim et al. 2008; Chen et al. 2008). A multitude of inference methods exist in the literature for the identification of such networks using observational gene expression data (Friedman et al. 2000; Murphy et al. 1999; Kim et al. 2004; Lebre et al. 2010). On the other hand, there is also intense interest in using perturbation experiments in the study of gene regulation. For example, the TF knock-down experiment is expected to very informative identifying potential targets of a TF because it depicts a less complex picture and can provide evidence for causal relationships (Geier et al. 2007; Werhli et al. 2006).

Traditionally, potential targets of the TF are usually identified as the subset of differentially expressed genes between the control and experiment group. However, when an important TF has been knocked down, it is almost always the case that the proportion of significantly changed genes is much larger than expected (Ivanova et al. 2006; Zhou et al. 2007). As a concrete example, consider the dataset analyzed in this study, which is from the knock-down experiment for two TFs which play an important role regulating ES cell pluripotency (see Section 6 for details). In this dataset, the number of differentially expressed genes is very large, while the number of likely direct targets (from external CHIP-seq data assessing TF binding) is significantly smaller.

There are two popular explanations for this phenomenon:

1. The theoretic null distribution of the test statistics(often z-score and other analogous quantities) for zero effect is not accurate.
2. There are a large number of genes showing nonzero but small changes of gene expression level, as effects of the perturbation.

Proposed solutions include modifying the null distribution of z-score empirically(Efron (2007, 2008)) and applying a cutoff to fold-change as a second-layer filter; the latter has been extremely popular in practice (Nichols et al. 1998; Zhou et al. 2007; Vaes et al. 2014). While both of these approaches can narrow down the selected, the former tackles the problem mainly based on the first explanation while the latter adopts the second implicitly, and results can be different in general (Witten and Tibshirani (2007)). For the knock-down experiment, the latter seems preferable because it considers both the change magnitude and the nonzero significant level, which is more related to what scientists care about; however, this approach lacks a natural quantitative justification.

Here, we propose a simple model to combine these two perspectives. By using a Gaussian distribution with unknown variance to describe the underlying behavior of genes in the null group, our model assumes that there can be relatively small nonzero effects even for the null genes. Assuming that the number of genes with large effect size is small, we test for the presence of such large effects relative to the background null variance. Although this model is motivated by the knock-down experiment, it can be applied to more general multiple-

CONTACT Leying Guan *leying.guan@yale.edu* Department of Biostatistics, Yale University, New Haven, CT, 06520.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

testing setting where both the significance and the effect size matter.

Our approach is related to the method of maximal agreement cut by Henderson and Newton (2015). They similarly pointed out that testing approaches which measure evidence against the null hypothesis tend to over-populate the candidate list with those associated with small variance, while approaches that consider only the magnitude will overlook the noise. While sharing the same spirit, our method does not aim to find the top α% subset of genes maximizing the expected overlap with the truth, with some assumed prior for all genes. Instead, we are interested in identifying the subset of genes which could not be described well by the prior describing the majority.

In the setting of knock-down experiment, our model describes a scenario different from the one assumed in approaches testing for differentially expressed genes. In our model, it is assumed that, perhaps due to propagation through the gene regulatory network, when we collect the data, a lot and even all genes may have been influenced once a TF has been knocked down, and we take this possibility into consideration. We show that in this scenario, it is still possible to test for strong effect if (1) the direct target tends to have larger effect size, and (2) there are enough null hypotheses to estimate the variance under the null.

The article is organized as follows: We describe our model in Section 2.1 and the procedure to estimate the null variance in Section 2.2. In Section 3, we study the properties of the estimating procedure of τ^2; in Section 4, we extended the model to the noncentered case and the case of two sample testing with unequal variance. We provide simulations in Section 5 and real data examples in Section 6.

2. Statistical Model and Estimation Procedure

2.1. Statistical Model

We assume there is a control group with m_0 replicates and an experiment group with m_1 replicates after knocking down one TF of interest. The expression levels for N genes are measured for each replicate. Let $x_{i,j}$ be the measurement for gene i in replicate j from the experiment group, and $z_{i,j}$ be the measurement for gene i in replicate j from the control group. Without loss of generality, assume the mean level of $z_{i,j}$ is 0 and the mean level of $x_{i,j}$ is μ_i

$$x_{i,j} \sim N(\mu_i, \sigma_i^2), \quad \forall i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m_1,$$
$$z_{i,j} \sim N(0, \sigma_i^2), \quad \forall i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m_0.$$

To do inference on μ_i, we can look at the two-sample test statistics

$$\bar{x}_i - \bar{z}_i = \frac{\sum_{j=1}^{m_1} x_{i,j}}{m_1} - \frac{\sum_{j=1}^{m_0} z_{i,j}}{m_0} \sim N(\mu_i, \sigma_i^2(\frac{1}{m_1} + \frac{1}{m_0})), \quad \forall i = 1, 2, \ldots, N.$$

Or the paired sample test statistics to remove the batch effect ($m_0 = m_1$)

$$\overline{x_i - z_i} = \frac{\sum_{j=1}^{m_1}(x_{i,j} - z_{i,j})}{m_1} \sim N(\mu_i, \sigma_i^2 \frac{1}{m_1}), \quad \forall i = 1, 2, \ldots, N.$$

As there is no fundamental difference between these two tests in our later analysis, we will omit the notation z_i and use the following common notations for simplicity

$$\bar{x}_i \sim N(\mu_i, \sigma_{\bar{x}_i}^2), \forall i = 1, 2, \ldots, N,$$
$$\hat{\sigma}_{\bar{x}_i}^2 \sim \sigma_{\bar{x}_i}^2 \frac{\chi_{m-k}^2}{m-k}, \forall i = 1, 2, \ldots, N,$$

where $\sigma_{\bar{x}_i}^2 = \frac{\sigma_i^2}{n}$, with n being the effective sample size and $\hat{\sigma}_{\bar{x}_i}^2$ is the usual unbiased variance estimate of $\bar{x}_i$. In the two-sample case, $n = \frac{m_1 m_0}{m}$, $m = m_1 + m_2$, $k = 2$ and in the paired sample case, $n = m = m_1$, $k = 1$.

Following the widely used two group model (Efron 2008), let A_0, A_1 denote the sets of nulls and nonnulls, respectively, and $\gamma = \frac{|A_1|}{N}$ denote the proportion of nonnulls. We assume that the μ_i's in A_0 and A_1 are generated from different distributions

$$\mu_i \sim \begin{cases} N(0, \tau^2) & \forall i \in A_0, \\ g_i(.) & \forall i \in A_1. \end{cases}$$

For each gene $i \in A_1$, g_i is some unknown density function. The parameter τ can be viewed as describing the range of normal behavior.

In contrast to the original two-group model, which corresponds to $\tau = 0$, we allow τ to take positive values. By relaxing this assumption on τ, we are able to detect relatively abnormal behavior compared with the background signal. If we know τ, the p-value for the new null hypothesis for gene i can be derived. Let $u_i = \frac{\bar{x}_i}{\sqrt{\tau^2 + \hat{\sigma}_{\bar{x}_i}^2}}$, $\bar{x}_i \sim N(0, \tau^2 + \sigma_{\bar{x}_i}^2)$, under the null hypothesis, u_i is a Welch statistics (Welch 1947) in the limit case with the degree of freedom for the first "variance estimate" τ^2 being ∞. Usual analysis controlling False discovery rate (FDR) or family-wise error rate (FWER) carries through under our extended model in this case.

We emphasize that the parameter τ itself is informative because it characterizes how influential a stimulus is—in our case, how dramatically the whole system changes after we have knocked down a TF. The value of τ reflects the importance of the TF: it can be set according to either prior knowledge or estimated from the data. The second approach is usually more feasible, as we lack a quantitative characterization of this kind of importance, and it can vary under different environments even for the same TF.

2.2. Estimation Procedure

For $i \in A_0$, $\bar{x}_i \sim N(0, \sigma_{\bar{x}_i}^2 + \tau^2)$ marginalizing out μ_i. If A_0 is known, the empirical Bayes estimate of τ^2 with estimated variance $\hat{\sigma}_{\bar{x}_i}^2$ for $\sigma_{\bar{x}_i}^2$ is given by $\hat{\tau}_{A_0}^2 = \frac{1}{|A_0|}\sum_{i \in A_0}(\bar{x}_i^2 - \hat{\sigma}_{\bar{x}_i}^2)$. Let δ be a predetermined small value, the adjusted form $\hat{\tau}_{A_0}^2 = [\frac{1}{|A_0|}\sum_{i \in A_0}\bar{x}_i^2 - (1+\delta)\hat{\sigma}_{\bar{x}_i}^2]_+$, is often preferred to reduce the error for small τ^2 and to ensure nonnegativity. Similar form of estimate is analyzed by Johnstone (2001b,a) in the context of estimating noncentrality of χ^2-distribution with known variance.

Let $F_i(.)$ denote the distribution of $\frac{\bar{x}_i^2}{\tau^2 + \hat{\sigma}_{\bar{x}_i}^2}$ when $i \in A_0$, which is the distribution of the square of a Welch's statistic as

mentioned before, and $\tilde{F}_i(x) = 1 - F_i(x)$. The Iterated empirical Bayes estimation (ITEB) procedure is given below, which starts from the whole set (as $A_0 \cup A_1$) and then iteratively remove potential outliers on the tail based on the current estimate of τ^2. It stops when no point needs to be further removed.

Iterated empirical Bayes estimation (ITEB) of τ^2
Input: $\{(\bar{x}_i, \hat{\sigma}^2_{\bar{x}_i}), \forall i = 1, 2, \ldots, N\}$, significance level α_1, α_2 and δ. By default, $\alpha_1 = 0.1$, $\alpha_2 = 0.01$ and $\delta = \sqrt{\frac{8}{N}}$.
Output: $\hat{\tau}^2$, the estimated τ^2.
Initialization: $S_0 = \{1, 2, \ldots, N\}$ be the initial estimate of the null set, and $\hat{\tau}^2_{S_0} = \frac{[\sum_{i=1}^N \bar{x}_i^2 - (1+\delta)\sum_{i=1}^N \hat{\sigma}^2_{\bar{x}_i}]_+}{N}$.
For $k = 1, 2, \ldots,$ do

1. Update the p-value for each gene $p_i = \tilde{F}_i(\frac{\bar{x}_i^2}{\hat{\tau}^2_{S_{k-1}} + \hat{\sigma}^2_{\bar{x}_i}})$. The ordered p values from small to large are $p_{(1)}, p_{(2)}, \ldots, p_{(N)}$. Let i^* be the largest index, such that $p_{(i^*)} \le \frac{i^*}{N}\alpha_1$.

2. Let $J_k^1 = \{i \in A : p_i \le p_{(i^*)}\}$, and $J_k^2 = \{i \in A : p_i \le \alpha_2\}$ and remove $J_k := J_k^1 \cap J_k^2$. Update $S_k = S_{k-1} \setminus J_k$ and $\hat{\tau}^2_{S_k} = \frac{[\sum_{i \in S_k} \bar{x}_i^2 - (1+\delta)\sum_{i \in S_k} \hat{\sigma}^2_{\bar{x}_i}]_+}{|S_k|}$.

3. If $S_k = S_{k-1}$, return $\hat{\tau}^2 = \hat{\tau}^2_{S_k}$

We have also described two other estimation methods and compared them with ITEB in the supplement: the truncated MLE method and the central matching (CM) method. The detailed descriptions of the truncated MLE and the CM estimator are given in Appendix C. These two methods have also been applied to estimating the empirical null distribution in the traditional two group test problem (Efron et al. 2001; Efron 2012), and we have adapted them to our problem here. In Appendix D, we compare performance of the three estimators in different scenarios and discuss their strengths and weaknesses. ITEB is most computationally efficient and is found to have better performance overall when the nonnull proportion γ is small. Thus, we will focus on ITEB and we provide detailed analysis of its properties.

3. Properties of ITEB

We study the estimation quality of ITEB as the number of hypotheses $N \to \infty$. For simplicity, we analyze the algorithm under following mild conditions and notations with $\delta = 0$. Let $\lambda(\alpha) := \max_i \tilde{F}_i^{-1}(\alpha)$. The degree of freedom for the variance estimate is m for all i, and $K :=$ the number of iterations needed for the algorithm to stop. Since in this section we only use the mean level $\bar{x}_i$ and its estimated variance $\hat{\sigma}^2_{\bar{x}_i}$, with slight abuse of notation, let $x_i := \bar{x}_i$, $\hat{\sigma}_i^2 := \hat{\sigma}^2_{\bar{x}_i}$, $\sigma_i^2 := \sigma^2_{\bar{x}_i}$. For the two levels α_1 and α_2 in the ITEB algorithm, we let $0 < \alpha_1 < \frac{1}{2e}$ be a fixed value, and let $\alpha_2 \to 0$ at a slow rate to simplify the notations in the proof (we always let $\frac{N\alpha_2}{\log^2 N}$ bounded away from 0).

Assumption 3.1. The degree of freedom for variance estimates $m \ge 5$ is a constant and the nonnull proportion $\gamma < 1 - c$ for some positive constant c. The ratio of variances of different genes is bounded: there exists a positive constant C such that $\frac{\max_i \sigma_i^2}{\min_i \sigma_i^2} \le C$.

Without loss of generality, we rescale $\min_i \sigma_i^2 = 1$, then $C = \max_i \sigma_i^2$. We do not require τ^2 to be positive or a constant. It can be 0 or decay to 0 as $N \to \infty$.

Assumption 3.2. There exist constants L and ϵ, such that

$$\forall i \in A_1, \quad E[x_i^2 - (1+\epsilon)\hat{\sigma}_i^2 \mid x_i^2 - (1+\epsilon)\hat{\sigma}_i^2 \le L(\tau^2 + 1)] \ge (1+\epsilon)\tau^2.$$

Remark 3.1. If $i \in A_0$, we have $E[x_i^2 - \hat{\sigma}_i^2] = \tau^2$. Assumption 3.2 states that, for $i \in A_1$, $x_i^2 - \hat{\sigma}_i^2$ has expectation nonnegligibly bigger than τ^2, and it is not purely driven by observations from its tail.

Assumption 3.3. The nonnull proportion $\gamma \to 0$ as $N \to \infty$.

Theorem 3.1. (Lower bound of the variance estimate) Let $R_K := |J_K \cap A_1|$, where J_K is the rejected set at the last step $k = K$ from ITEB. Let $\Delta_1 = \sqrt{\frac{\log N}{N}}(\tau^2 + C)$, $t_l = \max(\frac{\log^2 N}{N}, \min(\frac{l}{N}, 2\alpha_2))$, $\Delta_{2,l} = 3(\tau^2 + C)t_l \log\frac{1}{t_l}$ and $\tau_l^2 = [\tau^2 - \Delta_1 - \Delta_{2,l}]_+$. Under Assumptions 3.1 and 3.2, we have $P(\cup_{l=0}^{N\gamma}\{R_K = l, \hat{\tau}^2 \ge \tau_l^2\}) \to 1$.

As $\frac{\Delta_1 + \Delta_{2,l}}{\tau^2 + C} \to 0$ for all l, Corollary 3.1 is a direct result of Theorem 3.1.

Corollary 3.1. Under Assumptions 3.1 and 3.2, for any $\delta > 0$, $lim_{N\to\infty} P(\hat{\tau}^2 \ge [\tau^2 - \delta(\tau^2 + C)]_+) = 1$.

Theorem 3.2. (Upper bound of the variance estimate) Under Assumptions 3.1 and 3.3, suppose $\alpha_1 > 0$ is fixed and $\alpha_2 \to 0$ at a rate slow enough: $\gamma\lambda(\alpha_2) \to 0$. Then, for any $\delta > 0$, we have $\lim_{N\to\infty} P(\hat{\tau}^2 \le \tau^2 + \delta(\tau^2 + C)) = 1$.

We next show that these results can usually lead to good performance in the follow-up analysis in practice. Theorem 3.3 states that our estimate of τ^2 can successfully control the FDR if we reject the hypotheses in the set J_K.

Theorem 3.3. (FDR control) Under Assumptions 3.1 and 3.2, if we reject all hypothesis in J_K, we have $\lim_{N\to\infty} FDR \le \alpha_1$.

Remark 3.2. Note that at the given level α_1, α_2 in the ITEB algorithm, J_K^1 will correspond to the set of rejections using the BH (Benjamini–Hochberg procedure) (Benjamini and Hochberg 1995) and J_K will correspond to the set of rejections which are both rejected by the BH procedure and with p-values no greater than α_2. The extra requirement that the p-value is no greater than a reasonable small value α_2 is desirable in many large-scale hypotheses testing settings, including the knock-down experiment.

Theorem 3.4. (Power analysis) Let $\phi_{i,\alpha} = \mathbb{1}_{p_i \le \alpha}$ and let $\phi^*_{i,\alpha}$ be the oracle decision rule knowing τ^2: for any level α, $\phi^*_{i,\alpha} = \mathbb{1}_{x_i^2 > \tilde{F}_i^{-1}(\alpha)(\hat{\sigma}_i^2 + \tau^2)}$. Let $z_i^2 = \frac{x_i^2}{\tau_i^2 + \sigma_i^2}$ where $\tau_i^2 = E[\mu_i^2]$ for

$i \in A_1$. Under Assumptions 3.1 and 3.3, if we further assume that the density of z_i^2 is upper bounded by a constant, and the tail probability for z_i^2 decays sufficiently fast

$$\lim_{w\to\infty} \sup_{\delta>0} \sup_{i\in A_1} \frac{P(z_i^2 \le w(1+\delta))}{P(z_i^2 \le w)(1+\delta)} \le 1.$$

Then we have $\lim_{N\to\infty} \inf_{i\in A_1} \inf_{\alpha\ge 0}(P(\phi_{i,\alpha} = 1) - P(\phi^*_{i,\alpha} = 1)) \ge 0$.

Remark 3.3. Recall that the follow-up p-value p_i for hypothesis i is $\tilde{F}_i(\frac{x_i^2}{\hat{\tau}^2+\hat{\sigma}_i^2})$. Thus, Theorem 3.4 says that the test ϕ_i based on the estimated variance $\hat{\tau}^2$ is asymptotically as powerful as the optimal test based on the (unknown) true variance τ^2.

Proofs of Theorems 3.1, 3.2, 3.3, and 3.4 are given in Appendix A.

4. Extensions

4.1. Noncentered Null Distribution

We have been assuming that μ_i in the null set is generated according to $N(0,\tau^2)$. The ITEB estimation approach is easily extended to the setting where the null distribution might not be centered and $\mu_i \sim N(\epsilon,\tau^2)$ with a small noncentrality ϵ for $i \in A_0$. In ITEB, ϵ can be approximated by

$$\hat{\epsilon} = \frac{\sum_{i=1}^N \bar{x}_i I_{[\bar{x}_i\in(-\delta_0,\delta_0)]}}{\sum_{i=1}^N I_{[\bar{x}_i\in(-\delta_0,\delta_0)]}},$$

where $\delta_0 > 0$ is a reasonable cutoff and we treat $x_i \in (-\delta_0,\delta_0)$ to be from A_0. We can form an ITEB estimation for the noncentered case by replacing $\bar{x}_i$ with $\bar{x}_i - \epsilon$ in the ITEB algorithm.

4.2. Two-Sample Test With Unequal Variance

For hypothesis i, the observations from the experiment and control groups, $x_{i,j}$ and $z_{i,j}$, can have different variances. It is straightforward to generalize ITEB to this situation if we want to perform a two-sample test. We know that ITEB takes in $\{\bar{x}_i - \bar{z}_i\}$ and $\{\hat{\sigma}^2_{\bar{x}_i-\bar{z}_i}\}$. In the unequal variance setting, we can estimate $\sigma^2_{\bar{x}_i-\bar{z}_i}$ by

$$\hat{\sigma}^2_{\bar{x}_i-\bar{z}_i} = \frac{\sum_{j=1}^{m_1}(x_{i,j}-\bar{x}_i)^2}{m_1(m_1-1)} + \frac{\sum_{j=1}^{m_0}(z_{i,j}-\bar{z}_i)^2}{m_0(m_0-1)},$$

whose degree of freedom is approximated by

$$df_\sigma = \frac{(\frac{\sum_{j=1}^{m_1}(x_{i,j}-\bar{x}_i)^2}{m_1} + \frac{\sum_{j=1}^{m_0}(z_{i,j}-\bar{z}_i)^2}{m_0})^2}{\frac{1}{(m_1-1)}(\frac{\sum_{j=1}^{m_1}(x_{i,j}-\bar{x}_i)^2}{m_1})^2 + \frac{1}{(m_0-1)}(\frac{\sum_{j=1}^{m_0}(z_{i,j}-\bar{z}_i)^2}{m_0})^2}$$

We approximate $F_i(.)$, the distribution of the test statistics $\frac{(\bar{x}_i-\bar{z}_i)^2}{\tau^2+\sigma^2_{\bar{x}_i-\bar{z}_i}}$, by $F_{1,df}(.)$, the F distribution with degree of freedoms $(1, df)$, where df is approximated by $df = (\frac{\tau^2}{\hat{\sigma}^2_{\bar{x}_i-\bar{z}_i}} + 1)^2 df_\sigma$ (Satterthwaite 1946).

5. Simulation: Detection of Large Signal

We consider the two-sample setting with equal-variance and generate data under various values of τ and nonnull proportion $\gamma = \frac{|A_1|}{N}$. Specifically, we fix $N = 15{,}000$, $m = 5$ for both the control and experiment group, for any given τ and γ, where $\gamma = 1\%, 5\%$ and $\tau = 0, 0.1, \ldots, 1, 1.5, 2, 2.5, 3$, we generate the true mean and variance as below.

1. Let $\mu_i = 0$ in the control group, and in the experiment group, we generate them as follows:
$$\mu_i \sim \begin{cases} N(0,\tau^2) & \forall i \in A_0, \\ \pm U[1, \max(3, 10\tau)] & \forall i \in A_1, \end{cases}$$
where $U[1, \max(3, 10\tau)]$ is the uniform distribution between 1 and $\max(3, 10\tau)$, and the signs of μ_is will be half positive and half negative.
2. We sample the variances σ_i^2 from its empirical distribution from the real dataset, and we scale them to have mean level 1.

We compare the following approaches:

- ITEB estimate of τ^2, followed Welch's t-test.
- t-test with the null hypothesis testing for nonzero effect.
- EBarray (Kendziorski et al. 2003; Yuan and Kendziorski 2006), which is a two-group empirical Bayes method providing a posterior probability for having nonzero effect. We choose the "LNNMV" method to fit the data as suggested by the authors.
- Fold-change rankings with a threshold for t-test p-values being 10^{-5} (referred to as "fchange"). For hypotheses that do not pass this cutoff, we rank them based on p values from the t-test, after those who have passed the cutoff.
- Rvalue ranking, which finds the α% of genes maximizing the overlap between this gene list and gene list with differential change above the upper α% quantile of a prior normal distribution (Henderson and Newton 2015).

The proposed model, the rvalue ranking and the fchange all consider the magnitude and significance explicitly. If the p-value cutoff is correctly set for the fchange, we expect these three methods to share more in common in terms of their ROC curves (average sensitivity versus FDP), while only the proposed model allows you to set a cutoff based the significance level of whether a hypothesis is different from the background. LNNMV models observations from the control and the experiment directly instead of modeling their differences, also, although it smoothes the variance estimations with an inverse chi-square prior, the prior is not related to the effect sizes across hypotheses.

Figures 1 and 2 provide ROC curves across 20 repetitions as the significance level in the testing step is varied. We see that the new approach performs the best or one of the best across different experiments. When $\tau = 0$, ITEB procedure and t-test behave similarly, and ITEB becomes almost identical as the fchange when τ is large. ITEB and rvalue result in very similar results across different settings (ITEB is slightly better in the more sparse case with $\gamma = 1\%$), however, besides that ITEB provides information about where to set a cutoff, the ITEB approach is much faster considering the run time. All methods except for the t-test are stable across a wide range of τ.

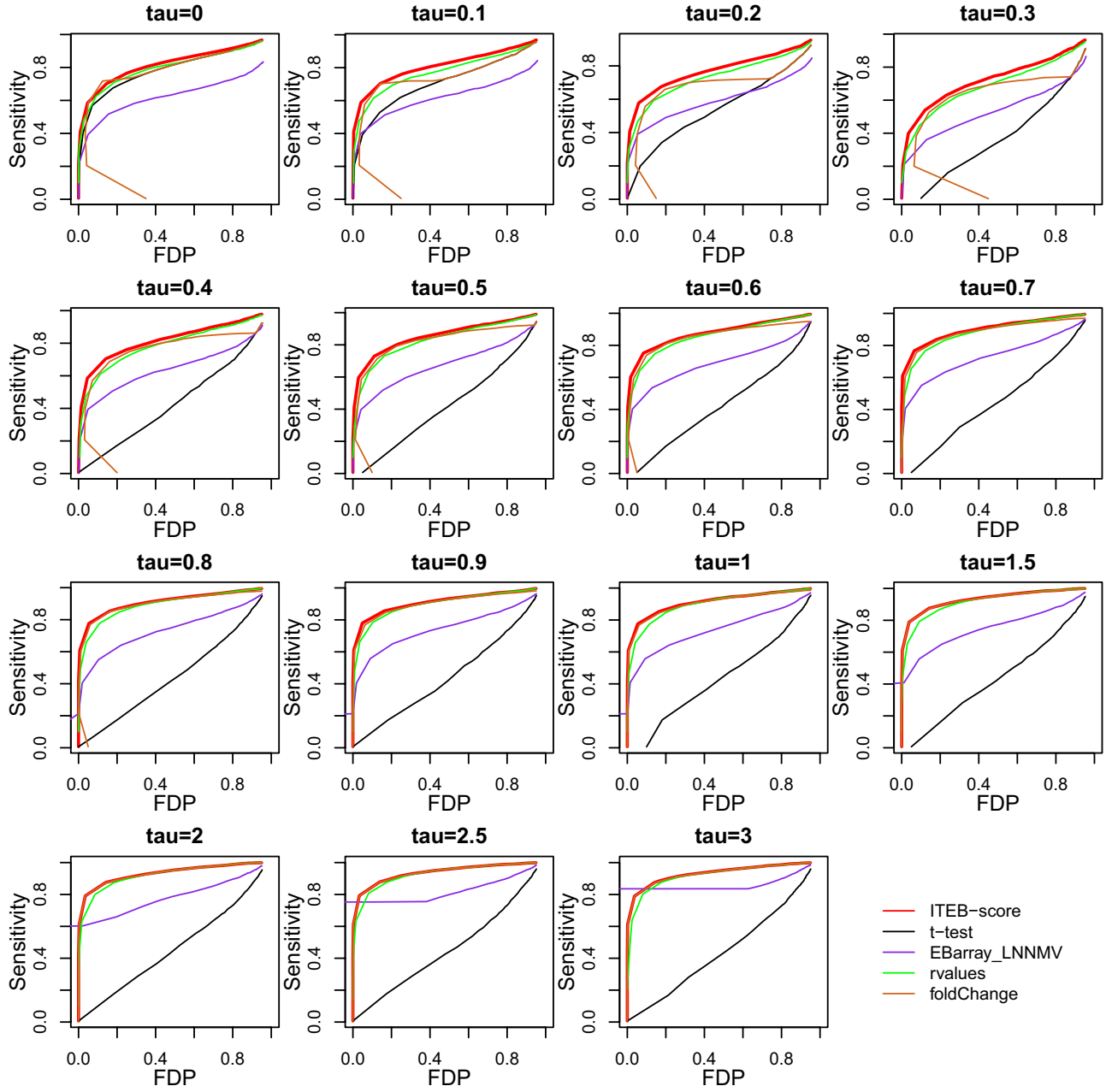

Figure 1. ROC curve for $\gamma = 1\%$.

6. Real Data Examples

In this section, we apply our approach to data from two knock-down experiments described below. The quality of the results is evaluated by the enrichment of ChIP-seq peaks (for the perturbed TF) in active enhancers/promoters for the selected genes. Note that the ChIP-seq data for ES cells are external to the data used to select the genes, and it provides an orthogonal information to access how likely the selected genes are direct targets of the TFs.

We perform gene knock-down experiments on 2 TFs on the mouse ES cell line R1. For each TF, RNA interference (RNAi) delivered using nucleofection was used to knock down its expression. Puromycin selection was introduced 18 h later at 1 μ g/ml, and the medium was changed daily. 30, 48, and 72 h after puromycin selection, the cells were collected for RNA isolation. After the experiments, Microarray hybridizations were performed on the MouseRef-8 v2.0 expression beadchip arrays (Illumina, CA). More details of the experiments can be found in Appendix E. Quantile normalization is performed in the first step to reduce the batch effect, and for the same reason, for each sample in the experiment group, we consider the paired test statistics with each pair being a pair of independent experiment and control samples from the same batch and time point. We have eight-paired observations for both POU5F1 and NANOG, and we take the log difference between the gene expression levels in a knock-down sample and its corresponding control sample to further reduce the batch effect. Table 1 summarizes the data we have at different time points. Figure 3 shows results from nine random realizations of the t-SNE (Maaten and Hinton 2008) plot using the top 1000 genes with

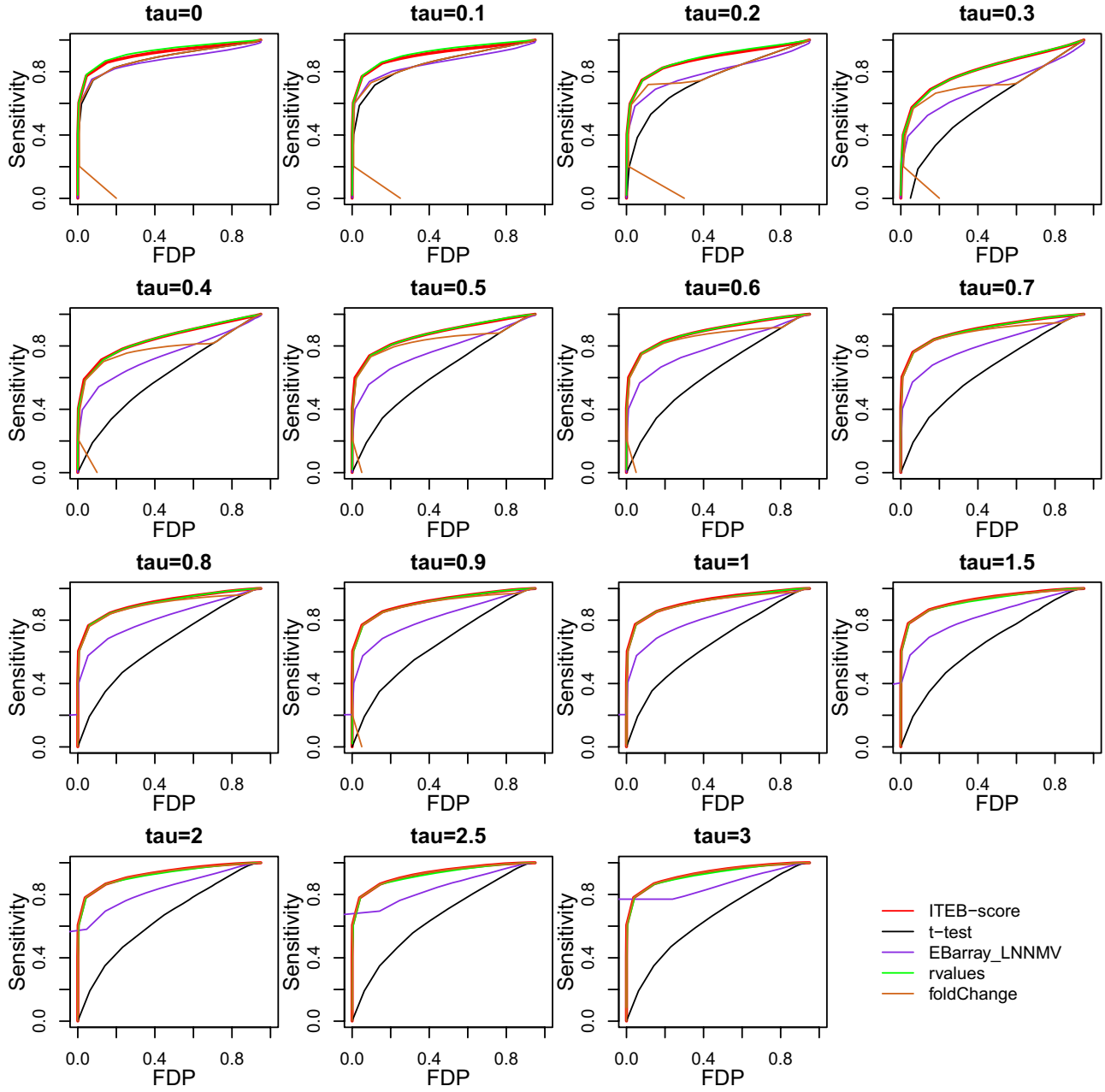

Figure 2. ROC curve for $\gamma = 5\%$.

Table 1. Information of the knock-down datasets.

	30 h	48 h	72 h
POU5F1	4 pairs	4 pairs	–
NANOG	2 pairs	4 pairs	2 pairs

largest variance across experiments. Each data point in the t-SNE plot represents one sample (paired) in the experiment. We use the colors black, red and green to represent data at time points 30, 48, and 72 h, respectively. From the results, we see that differences between time points within the same knock-down experiment is comparable to the differences across batches, and they are very small compared with the differences between two knock-down experiments. To compare the targets of two TFs, we will regard different times points in the same knock-down experiment as replicates of each other.

ChIP-seq data and enhancer-gene association data: To evaluate the quality of the selected gene set, we use two external datasets: the ChIP-seq data are from Chen et al. (2008) and the enhancer-gene association data are from Mumbach et al. (2017). The ChIP-seq data contain results using chromatin immunoprecipitation coupled with ultra-high-throughput DNA sequencing (ChIP-seq) to map the binding locations of 13 sequence-specific TFs, including POU5F1 and NANOG. The enhancer-gene association data are generated by using the HiChIP method where the authors performed H3K27ac HiChIP in mouse ES cells. H3K27ac is a histone modification mark characteristic of active enhancers and promoters in the cell. HiChIP using H3K27ac mark as bait will provide enhancer–gene interaction information. We can evaluate the quality of the selected gene set by examining whether the binding sites of POU5F1 and NANOG are enriched near the

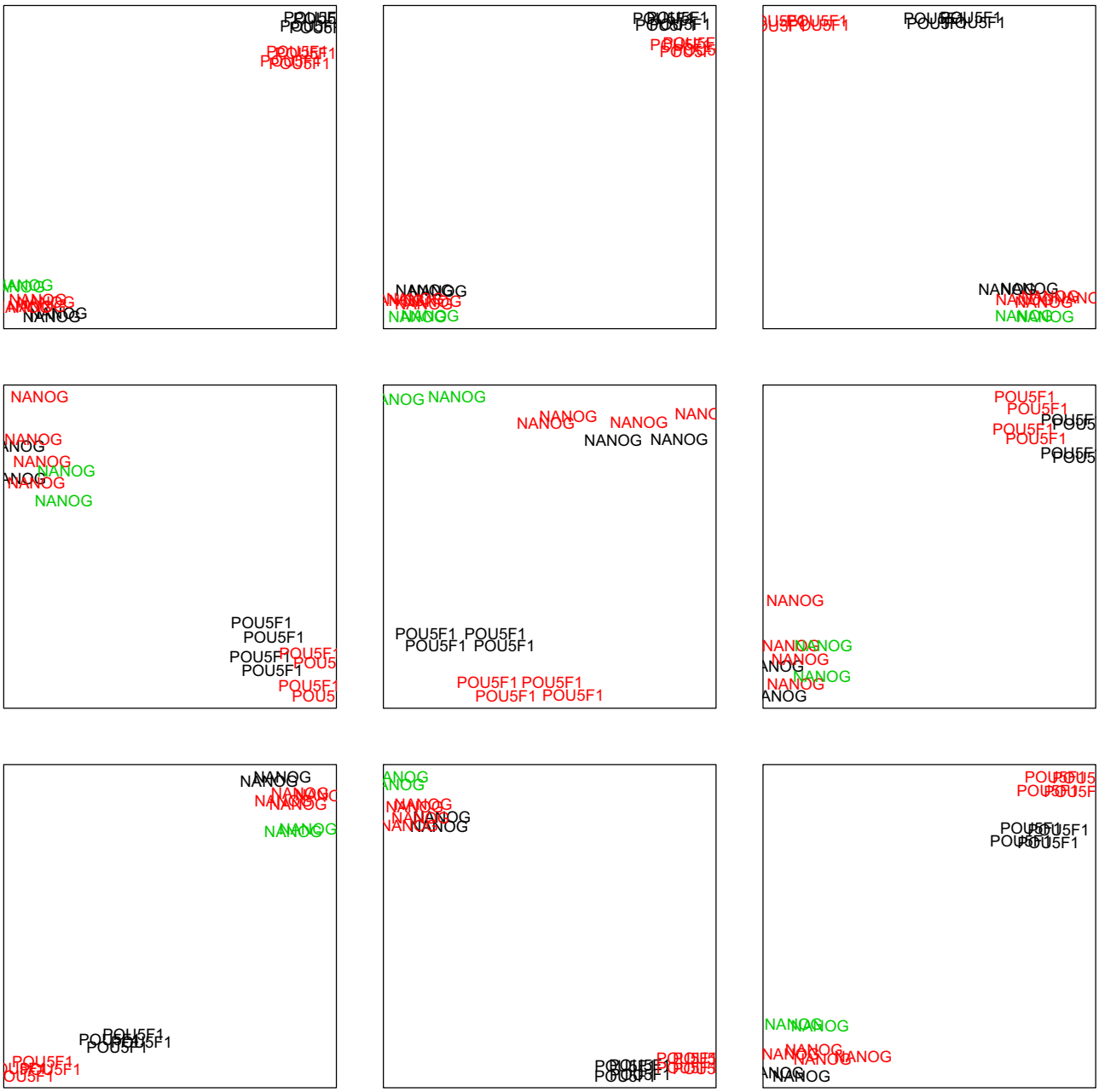

Figure 3. Clustering with t-SNE algorithm.The nine plots here are 9 different random tsne-plot realizations.Black, red, and green colors represent experiments from 30 hr, 48 hr and 72 hr, respectively. Note that there are two replicates from the same batch in a single experiment (same day, same TF, same batch) that are almost identical to each other.

active enhancers/promoters of the selected genes in the ES cell.

Let us call the approach based on p-value using a simple t-test S_0, the approach based on EBarray S_1, and the approach based on p-value using ITEB S_2. We will focus on those genes with significant decrease in their expression levels after POU5F1/NANOG knock-down. For each method, we set the cutoff using the BH procedure with targeted FDR level at 0.01. Accordingly, we set the cut-off level $\alpha_1 = 0.01$ and we select 87 genes after knocking down POU5F1 and 43 genes after knocking down NANOG using S_2. These numbers are 2274 and 1267 using S_0, 2362 and 886 using S_1. Neither S_0 nor S_1 provides informative candidate lists with this criterion. To have a meaningful comparison, we also consider the case where we control FWER at 0.01, which is quite a stringent criterion and under which, S_0 selects 144 genes for POU5F1 and 49 genes for NANOG, S_1 selects 1091 genes for POU5F1 and 254 genes for NANOG.

We say that there is supporting evidence of a gene being a direct target of a TF if this TF has at least one ChIP-seq peak within x kilobase(kb) away from the gene's active enhancers/promoter. As we change x in a large range of value, Figure 4 gives the percentages of genes with this supporting evidence in the selected gene sets using S_0, S_1 with FWER control and S_2 with FDR control. Figure 4 also shows the percentages of all genes with this supporting evidence (referred to as "all" in the figure), and the percentages of bottom 2000 genes with this supporting evidence (referred to as "bottom").

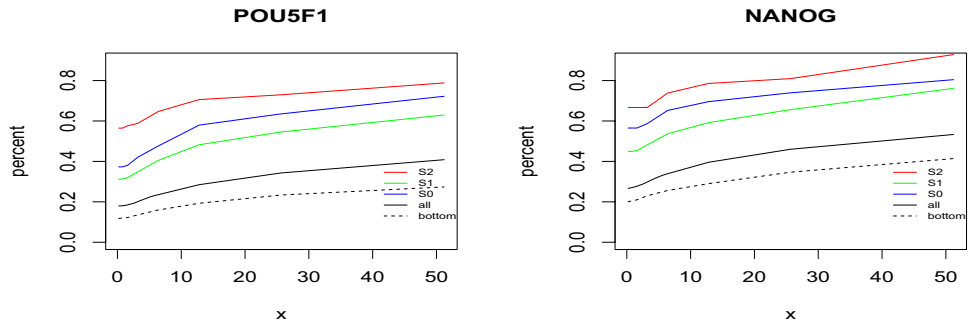

Figure 4. Percent of genes with ChIP-seq nearby versus x for the selected gene sets (S_0/S_1: genes selected using t-test/EBarray with FWER control, S_2: genes selected based on ITEB with FDR control, bottom: the bottom 2000 genes which show the smallest changes, all: all genes). The x-axis is the threshold we use to define whether a gene has a ChIP-seq peak near its enhancer/promoter and the y-axis is the percentage of selected genes with ChIP-seq nearby.

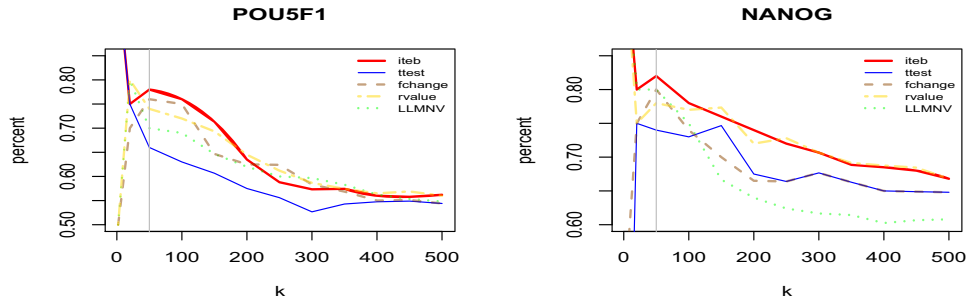

Figure 5. Percent of genes with ChIP-seq nearby versus selected gene size. The x-axis is the threshold of the ranking, and we only consider the top k genes from each ranking list.

Table 2. S_0, S_1, and S_2 results with FDR/FWER level set at 0.01.

	TF	Percent (negative control)	Size (S_0)	Percent (S_0)	Size (S_1)	Percent (S_1)	Size (S_2)	Percent (S_2)
FDR	POU5F1	0.21	2274	0.48	2362	0.48	87	0.73
	NANOG	0.32	1267	0.62	886	0.61	43	0.81
FWER	POU5F1	0.21	144	0.62	1091	0.53	31	0.74
	NANOG	0.32	49	0.74	254	0.64	20	0.8

The third, fifth, seventh, and nineth columns are the percent of genes with Chip-seq+Hi-C support in the negative control set, the gene set selected by S_0 (t-test), the gene sets selected by S_1 (EBarray) and S_2 (ITEB t-test), respectively.

The bottom 2000 genes are those showing less significance after the knock-down based on the ITEB p-value (we consider this set to be the negative control). From Figure 4, we see that the ChIP-seq enrichment is quite significant for the selected genes comparing with both the negative controls and all genes. The selected gene set using S_2 is significantly better than the gene set selected using S_1 and S_0. Table 2 shows the result with $x = 20$.

Figure 5 provides quality evaluation of the top K genes in different ranking lists, respectively. The vertical gray line is where the top 50 genes is. Besides S_0, S_1, and S_2, we also include the results using rvalues and fchange. Comparing S_2 with S_0 and S_1, not only S_2 provides a candidate gene set with much smaller size and higher quality, it provides a gene ranking list with higher quality. In terms of the gene ranking list, the rvalue provides a similar list as that from ITEB ($\sim$ 90% overlappings in the top 200 genes for both TFs).

7. Discussion

In this article, motivated by the problem of identifying TF targets based on data from the knock-down experiment, we have proposed to test for large effect size instead of non-zero effect size in the two-group model where a Gaussian distribution with nonzero variance is used for the effect in the null group. We have considered three approaches (ITEB, truncated MLE and CM) to estimate this nonzero variance adaptively, and recommend ITBE for its computational efficiency, strong performance in simulation and attractive theoretical properties. Although we have focused on the Gaussian setting here, the idea of testing for strong signal and the approaches to estimate the null distribution can be applied to problems involving other data types.

The model itself is related to the "g-modeling" (Carroll and Hall 1988; Efron 2014), the "ϵ-contamination" (Huber 1964; Chen et al. 2016) and the "robust Bayesian analysis" (Berger and Berliner 1986; Gaver and O'Muircheartaigh 1987; Berger et al. 1994). However, it should be noted that our approach has a different goal from the g-modelings. We only estimate the shape of the null distribution of the mean parameter while the g-modeling models the marginal distribution of the mean parameter considering both the nulls and the non-nulls. Our purpose for estimating the null effect distribution is to set a cutoff for the strong signals adaptively, which is not the case for the g-modeling. Although our model can be considered as a special case of the ϵ-contamination model in the parameter space and a special case of the robust empirical Bayesian analysis, the use of these models for large effects have not been studied and, to the best of our knowledge, methods with provable guarantees on power and FDR have not been demonstrated previously.

Acknowledgement

We would like to thank Zhou Fan for his helpful feedbacks. This work is supported by NIH Grants R01HG007834 and R01GM109836, NSF Grants DMS1721550 and DMS1811920

Supplementary Material

Online Supplementary Materials.pdf: This is the Appendix contains (1) proofs to Theorems in the paper (Appendix A, B), (2) details of estimating τ^2 using the truncated MLE and the CM methods (Appendix C), (3) simulations comparing ITEB, truncated MLE and the CM methods (Appendix D), and (4) details of how the knock-down experiment is performed (Appendix E) (.pdf file)

Data and code: Inside the directory "data", we have data files containing gene expression levels from the knock-down experiments (*.csv files) and a file containing experiment conditions for each column of the previous files (*.xlsx). Inside the directory "code", we provide R code used to generate results in Section 6. (.zip file)

References

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [3]

Berger, J., and Berliner, L. M. (1986), "Robust Bayes and Empirical Bayes Analysis with ε-contaminated Priors," *The Annals of Statistics*, 461–486. [8]

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., and Dasgupta, A. (1994), "An Overview of Robust Bayesian Analysis," *Test*, 3, 5–124. [8]

Carroll, R. J., and Hall, P. (1988), "Optimal Rates of Convergence for Deconvolving a Density," *Journal of the American Statistical Association*, 83, 1184–1186. [8]

Chambers, I., and Smith, A. (2004), "Self-renewal of Teratocarcinoma and Embryonic Stem Cells," *Oncogene*, 23, 7150–7160. [1]

Chen, M., Gao, C., and Ren, Z. (2016), "A General Decision Theory for Huber's ε-contamination Model," *Electronic Journal of Statistics*, 10, 3752–3774. [8]

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., and Loh, Y. H. (2008), "Integration of External Signaling Pathways With the Core Transcriptional Network in Embryonic Stem Cells," *Cell*, 133, 1106–1117. [1,6]

Efron, B. (2007), "Correlation and Large-scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [1]

——— (2008), "Microarrays, Empirical Bayes and the Two-groups Model," *Statistical Science*, 1–22. [1,2]

——— (2012), *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Vol. 1), Cambridge: Cambridge University Press. [3]

——— (2014), "Two Modeling Strategies for Empirical Bayes Estimation," *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 29, 285–301. [8]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160. [3]

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, 7, 601–620. [1]

Gaver, D. P., and O'Muircheartaigh, I. G. (1987), "Robust Empirical Bayes Analyses of Event Rates," *Technometrics*, 29, 1–15. [8]

Geier, F., Timmer, J., and Fleck, C. (2007), "Reconstructing Gene-regulatory Networks from Time Series, Knock-out Data, and Prior Knowledge," *BMC Systems Biology*, 1, 11–26. [1]

Henderson, N. C., and Newton, M. A. (2015), "Making the Cut: Improved Ranking and Selection for Large-scale Inference," *Journal of the Royal Statistical Society*, Series B, 78, 781–804. [2,4]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [8]

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. R. (2006), "Dissecting Self-renewal in Stem Cells With RNA Interference," *Nature*, 442, 533. [1]

Johnstone, I. (2001a), "Thresholding for Weighted χ^2," *Statistica Sinica*, 11, 691–704. [2]

——— (2001b), Chi-square Oracle Inequalities," *Lecture Notes–Monograph Series*, 36, 399–418. [2]

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003), "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles," *Statistics in Medicine*, 22, 3899–3914. [4]

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008), "An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells," *Cell*, 132, 1049–1061. [1]

Kim, S., Imoto, S., and Miyano, S. (2004), "Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data," *Biosystems*, 75, 57–65. [1]

Lebre, S., Becq, J., Devaux, F., Stumpf, M. P., and Lelandais, G. (2010), "Statistical Inference of the Time-varying Structure of Gene-Regulation Networks," *BMC Systems Biology*, 4, 130–145. [1]

Loh, Y.-H., Wu, Q., Joon-Lin, C., Vega, V. B., Zhang, W., Chen, X., Bourque, G., Joshy, G., Leong, B., Liu, J., and Wong, K. Y. (2006), "The oct4 and Nanog Transcription Network Regulates Pluripotency in Mouse Embryonic Stem Cells," *Nature Genetics*, 38, 431. [1]

Maaten, L. v. d., and Hinton, G. (2008), "Visualizing Data Using t-sne," *Journal of Machine Learning Research*, 9, 2579–2605. [5]

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., and Wei, Y. (2017), Enhancer Connectome in Primary Human Cells Identifies Target Genes of Disease-associated DNA Elements," *Nature Genetics*, 49, 1602. [6]

Murphy, K., and Mian, S. (1999), "Modelling Gene Expression Data Using Dynamic Bayesian Networks," Technical Report, Computer Science Division, University of California, Berkeley, CA. [1]

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998), "Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor oct4," *Cell*, 95, 379–391. [1]

Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998), "Self-renewal of Pluripotent Embryonic Stem Cells is Mediated Via Activation of STAT3," *Genes & Development*, 12, 2048–2060. [1]

Niwa, H., Miyazaki, J.-i., and Smith, A. G. (2000), "Quantitative Expression of oct-3/4 Defines Differentiation, Dedifferentiation or Self-renewal of ES Cells," *Nature Genetics*, 24, 372. [1]

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114. [4]

Vaes, E., Khan, M., and Mombaerts, P. (2014), "Statistical Analysis of Differential Gene Expression Relative to a Fold Change Threshold on Nanostring Data of Mouse Odorant Receptor Genes," *BMC Bioinformatics*, 15, 39–57. [1]

Welch, B. L. (1947), "The Generalization of Student's' Problem When Several Different Population Variances are Involved," *Biometrika* 34, 28–35. [2]

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006), "Comparative Evaluation of Reverse Engineering Gene Regulatory Networks With Relevance Networks, Graphical Gaussian Models and Bayesian Networks," *Bioinformatics*, 22, 2523–2531. [1]

Witten, D., and Tibshirani, R. (2007), "A Comparison of Fold-change and the t-statistic for Microarray Data Analysis," *Analysis*, 1776, 58–85. [1]

Yuan, M., and Kendziorski, C. (2006). A Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification," *Biometrics*, 62, 1089–1098. [4]

Zhou, Q., Chipperfield, H., Melton, D. A., and Wong, W. H. (2007), "A Gene Regulatory Network in Mouse Embryonic Stem Cells," *Proceedings of the National Academy of Sciences*, 104, 16438–16443. [1]