EMPIRICAL BAYES ANALYSIS OF RNA SEQUENCING EXPERIMENTS WITH AUXILIARY INFORMATION

By Kun Liang

University of Waterloo

Finding differentially expressed genes is a common task in high-throughput transcriptome studies. While traditional statistical methods rank the genes by their test statistics alone, we analyze an RNA sequencing dataset using the auxiliary information of gene length and the test statistics from a related microarray study. Given the auxiliary information, we propose a novel nonparametric empirical Bayes procedure to estimate the posterior probability of differential expression for each gene. We demonstrate the advantage of our procedure in extensive simulation studies and a psoriasis RNA sequencing study.

1. Introduction. Consider a recent RNA sequencing (RNA-seq) study of psoriasis vulgaris disease in Jabbari et al. (2012), where the expression levels of 18151 genes were measured on three pairs of lesional and nonlesional skin samples collected from three patients. Psoriasis vulgaris, or psoriasis, in short, is a chronic autoimmune disease characterized by skin inflammation and affects 2–4% of the population in western countries (Parisi et al., 2013). The main objective of the study is to find the differentially expressed (DE) genes between lesional and nonlesional skin samples to further our understanding of the disease mechanism. RNA-seq is the new generation of high-throughput technology that can measure tens of thousands of gene expression levels simultaneously. Comparing to the microarray technology that has been in use in the past two decades, RNA-seq provide more precise measurement of gene expression levels, especially for lowly or very highly expressed genes (Kukurba and Montgomery, 2015). High-throughput technology such as microarray and RNA-seq are typically expensive, and the sample sizes are usually small. For example, as of August 2017, among more than 11 thousands RNA-seq experiment datasets deposited in one of the largest gene expression databases, Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo), the median sample size is only 8.

We want to detect DE genes but also need to limit the number of false positives. The false discovery rate (FDR), which is defined as the expected proportion of false positives, has become the common error rate to control
in the literature. We can test the differential expression of genes using the limma-voom method (Law et al., 2014) which computes a list of $t$-statistics, $t_1, \ldots, t_{18151}$, with 8.8 effective degrees of freedom. Then we could easily compute $p$-values from the $t$-statistics and use the linear step-up procedure of Benjamini and Hochberg (1995), the BH procedure, to control the FDR at a certain target level. Alternatively, we can transform the $t$-statistics to $z$-values and use the empirical Bayes method that estimates the local false discovery rate (Lfdr) according to the two-group model (Efron and Tibshirani, 2002). Details of the $z$-value transformation can be found at the beginning of Section 2.

Much of the multiple testing literature, including the above BH procedure and empirical Bayes approach and their variants, treat all hypotheses as exchangeable, and all null hypotheses are considered equally likely to be true. However, we often know additional information about the tests that are potentially useful. In our psoriasis example, each hypothesis is associated with a human gene, whose coding region has a certain length. In Swindell et al. (2014), a strong gene length effect is reported in both microarray and RNA-seq studies of psoriasis skin. Figure 1 shows the RNA-seq $z$-values versus the ranks of gene coding region length, and we also plot the smoothing spline fit of the $z$-values as a solid blue line. Figure 1 suggests that the $z$-
values tend to be more negative when the coding regions are long, especially for the top 2000 or so longest genes.

The gene coding region length represents a type of general covariate information that can be used in many genetic studies. Other examples of general covariate information include the minor allele frequency of single nucleotide polymorphism (SNP) in genome-wide association studies (GWAS) and the SNP-gene distance in expression quantitative trait loci (eQTL) mapping studies (Ignatiadis et al., 2016), among others.

For any specific genetic study, there could be previous studies conducted under similar experimental conditions (Li and Barber, 2017) or on related diseases (Andreassen et al., 2013). Before the psoriasis RNA-seq study of Jabbari et al. (2012), Gudjonsson et al. (2010) have used microarray to study the gene expression differences between 58 pairs of lesional and non-lesional skin samples from psoriasis patients. The two studies are based on two different technology platforms: microarray quantifies the gene expression level through measuring light intensities of designed probe sets while RNA-seq counts the number of sequenced genome fragments mapped to genes. Roughly, we can think the microarray platform as a targeted analog approach while the RNA-seq platform offers an unbiased digital approach. The two studies also differ in their patient enrollment criteria, which we will discuss in detail in Section 4. Despite many differences between the two studies, the previous microarray study could provide useful information for our RNA-seq study.

Among the 18151 genes measured in the psoriasis RNA-seq study (Jabbari et al., 2012), only 16490 genes can be found in the microarray study of Gudjonsson et al. (2010) due to the platform difference. Figure 2 Panel (a) plots the RNA-seq \( z \)-values against the corresponding microarray \( t \)-statistics. Intuitively, we can interpret Figure 2 as follows: the dense vertical ellipse data point cloud centered at \((0, 0)\) is likely due to non-DE genes in both studies, on the other hand, DE genes with varying effect sizes are likely to yield positively correlated \( z \)-values and \( t \)-statistics, and the corresponding data points appear predominantly in the first and the third quadrants. In Figure 2 Panel (b), we plot the histograms of the RNA-seq \( z \)-values whose corresponding microarray \( t \)-statistics are less than \(-10\) and greater than \(10\) in grey and light brown colors, respectively. The two conditional \( z \)-value distributions are quite different, and there are many more positive \( t \)-statistic and \( z \)-value pairs than negative pairs. One of the main contributions of our proposed method in Section 2 is its ability to model such changing conditional distributions. From both Panels (a) and (b), we draw the following conclusion. Conditionally, if a gene has a large positive \( t \)-statistic from the microarray
study, then it is likely the corresponding RNA-seq z-value will also be large and positive; a similar phenomenon is true for negative pairs of t-statistics and z-values. In other words, given the microarray t-statistic, we would have more knowledge about the prior probability of differential expression and the likely direction and strength of the differential expression.

![Graph](image)

**Fig 2.** Panel (a), scatter plot of RNA-seq z-values versus microarray t-statistics. Panel (b), histograms of z-values whose corresponding t-statistics are less than −10 (grey) and greater than 10 (light brown).

A natural question is how we can utilize the auxiliary covariate information to improve the power of detecting DE genes. The total of 18151 genes in the RNA-seq study can be divided into two groups: one group contains
the 16490 genes with matching microarray data and another group of 1661 genes without matches. For grouped hypotheses, Efron (2008) and Cai and Sun (2009) suggest performing separate empirical Bayes analysis in each group and thresholding the posterior probabilities of being true null across groups. Then the challenge is how to estimate the posterior probabilities in groups with potentially multiple continuous covariates, which is the focus of this paper.

Given multiple covariates, the methods developed by Qu et al. (2012) and Scott et al. (2015) can be used to estimate the posterior probability of no differential expression, but both require strong modeling assumptions. Both methods assume a constant alternative density, which is not a reasonable assumption in our psoriasis application as Figure 2 illustrates. The method of Qu et al. (2012) applies only to t-statistics, and they model the true null probability as an additive function of covariates. More importantly, Qu et al. (2012) restrict the noncentrality parameters in the alternative t distribution to follow a normal distribution with a zero mean. This assumption implies that the alternative distribution is symmetric around zero, but Figure 2 suggests that the DE genes are more likely to be up-regulated in lesional skin and effect sizes in up-regulated genes are larger than down-regulated genes. Scott et al. (2015) work with z-values and thus is not restricted to applications with t-statistics. Furthermore, they model the covariate effect on the true null probability through a regression framework and can model nonlinear effects through spline basis expansion. However, it is unclear how their tuning parameters, such as the number of knots of splines, should be chosen. Similarly, it may be possible for Scott et al. (2015) to model interactions among covariates, but it is unclear how to do so in practice. In their simulation study with interaction effect between covariates, they chose to fit an additive model to show their result is robust to model misspecification. More critically, their implementation relies on a normal mean model where the alternative density is a location mixture of the null Gaussian density. The normal mean model allows flexible alternative distributions, which for example can be asymmetric. On the other hand, the normal mean model may not be suitable if the original test statistics do not follow the normal distribution, such as the t-statistics in our example.

Alternatively, for a target FDR level, frequentist methods such as the BH procedure will return a list of rejections. If the same experiment is repeated many times, frequentist methods can provide guarantees that the long-term average of false discovery proportions will be no greater than the target level. For example, Ignatiadis et al. (2016) control the FDR asymptotically with a single covariate by assigning different weights to p-values. As a comparison,
frequentist methods provide binary decision of rejection or acceptance given a target FDR level, while Bayes methods provide more detailed information of the posterior probability of each null hypothesis being true. Furthermore, the hypotheses in our application naturally form two groups, and it is unclear what is the best strategy to combine frequentist results in our grouped setting.

In Section 2, we derive the optimal procedure given the covariates and propose a novel empirical Bayes method to mimic the optimal procedure. Our proposed method is evaluated through extensive simulation studies in Section 3. We will return to the psoriasis example in Section 4 before we conclude with a discussion in Section 5. All technical proofs, additional simulation and application results are presented in Supplementary Material.

2. The Proposed Approach. Suppose $H_1, \ldots, H_m$ are $m$ null hypotheses of interest, among which $m_0$ are true nulls and $m_1$ are alternatives with $m = m_0 + m_1$. For the $i$th gene in our psoriasis example, the corresponding null hypothesis is

$$H_i : \text{gene } i \text{ is not differentially expressed.}$$

The outcome of a multiple testing procedure can be summarized in Table 1.

The false discovery rate (FDR) is defined as $E\{N_{1|0}/ \max(R,1)\}$, the expected proportion of false discoveries among all rejections. A closely related concept is the marginal FDR (mFDR), which is defined as $E(N_{1|0})/E(R)$. It can be shown that mFDR = FDR + $O(m^{-1/2})$ under independence (Genovese and Wasserman, 2002), i.e., the mFDR is asymptotically equivalent to the FDR. Another related concept is the marginal false non-discovery rate (mFNR) defined as $E(N_{0|1})/E(S)$. Though many optimality definitions are possible, we adopt the criterion in Sun and Cai (2007) where the optimal multiple testing procedure is defined as the one that controls mFDR at a certain level while minimizing mFNR.

2.1. Optimal Procedure for Multiple Testing with Covariates. Consider the commonly used two-group model (Efron et al., 2001): suppose $H_1, \ldots, H_m$ are $m$ null hypotheses of interest, $Y_1, \ldots, Y_m$ are the corresponding test
statistics, and \( \theta_1, \ldots, \theta_m \) are the corresponding true null indicators, i.e., \( \theta_i = 1 \) indicates \( H_i \) is true, and \( \theta_i = 0 \) otherwise. Then

\[
\begin{align*}
\theta_i &\sim \text{Bernoulli}(\pi_0), \\
Y_i | \theta_i &\sim \theta_i F_0 + (1 - \theta_i) F_1,
\end{align*}
\]

where \( \pi_0 \) is a constant true null probability, \( \theta_i \)'s are independent, and \( Y_i \)'s are independent given \( \theta_i \)'s. The marginal cdf of \( Y \) is \( F(y) = \pi_0 F_0(y) + (1 - \pi_0) F_1(y) \) with a probability density function (pdf) of \( f(y) = \pi_0 f_0(y) + (1 - \pi_0) f_1(y) \). Sun and Cai (2007) show that the \( p \)-value based multiple testing procedures can be inefficient when the original test statistics have directions, for example, in our psoriasis example the \( t \)-statistics can be either positive or negative. For efficiency and convenience, we will work with \( z \)-values, which can be easily transformed from the original test statistics. More specifically, in our psoriasis example, we can compute \( z \)-values as

\[
y_i = \Phi^{-1}(G_0(t_i)), \quad i = 1, \ldots, 18151,
\]

where \( \Phi \) is the cumulative distribution function (cdf) of the standard normal distribution and \( G_0 \) is the cdf of \( t_{8.8} \), which is the \( t \) distribution with 8.8 degrees of freedom. This transformation is one-to-one, and we will not lose the directional information. For \( z \)-values, \( F_0 \) is typically assumed to be \( \Phi \), the cdf of the standard normal, or can be estimated empirically from data (Efron, 2004).

For any rejection region \( \Lambda \), Efron and Tibshirani (2002) define the Bayesian FDR as

\[
F_{\text{dr}}(\Lambda) \equiv P(\theta = 1 | y \in \Lambda) = \pi_0 F_0(\Lambda)/F(\Lambda),
\]

where \( F_0(\Lambda) = P(y \in \Lambda | \theta = 1) \) and \( F(\Lambda) = P(y \in \Lambda) \). Under the two-group model, it can be shown that \( \text{mFDR}(\Lambda) = F_{\text{dr}}(\Lambda) \). Define the local FDR (Lfdr) as

\[
\text{Lfdr}(y) \equiv P(\theta = 1 | y) = \pi_0 f_0(y)/f(y).
\]

Efron and Tibshirani (2002) further show that \( F_{\text{dr}}(\Lambda) = E\{\text{Lfdr}(y) | y \in \Lambda\} \), which suggests an empirical FDR estimate of a set of rejections \( \mathcal{R} \) to be \( \sum_{i : y_i \in \mathcal{R}} \text{Lfdr}(y_i) / |\mathcal{R}| \).

Suppose for each hypothesis \( H_i \), there is a \( p \)-variate vector of auxiliary variables \( X_i = (X_{i1}, \ldots, X_{ip}) \) in addition to the primary test statistic \( Y_i \),
where \( p \geq 1 \) is a positive integer. We emphasize that the auxiliary variables we study are the test level information, such as the gene length and microarray \( t \)-statistic for each gene in our psoriasis example. The test level auxiliary information should be distinguished from the auxiliary or covariate information on the subject or sample level, such as age and sex of patients that are commonly used in the literature, for example, see Hummel et al. (2008) and others. Under the two-group model, which we denote as Model I, \( \text{Lfdr}(y) \) is shown to be an optimal statistic to control mFDR while minimizing mFNR (Sun and Cai, 2007). Because we rarely fix the covariate values before we decide which hypotheses to test, we will treat the covariate vector \( X \) as random. Our two-group joint model for both \( X \) and \( Y \), denoted as Model I*, is as follows:

\[
\begin{aligned}
\theta_i & \sim \text{Bernoulli}(\pi_0), \\
X_i, Y_i | \theta_i & \sim \theta_i F_0 + (1 - \theta_i) F_1,
\end{aligned}
\]

where \( \theta_i \)'s are independent and \( (X_i, Y_i) \)'s are independent given \( \theta_i \)'s. The joint pdf is \( f(x, y) = \pi_0 f_0(x, y) + (1 - \pi_0) f_1(x, y) \). The two-group joint model is parallel to the basic two-group model, and we regain the exchangeability after we jointly model the covariates and test statistic. With a slight abuse of notation, here we reuse \( F_0, F_1, f, f_0, f_1 \) for joint cdfs and pdfs instead of their marginal versions as in the two-group model, for the sake of brevity. Then, the joint local FDR can be expressed as

\[
\text{Lfdr}(x, y) = \frac{\pi_0 f_0(x, y)}{f(x, y)}.
\]

**Theorem 1.** Under Model I*, let an optimal rejection region \( \Lambda_{OPT} = \{(x, y) : \text{Lfdr}(x, y) \leq c\} \) for some constant \( 0 < c < 1 \), then for any rejection region \( \Lambda \) with \( \text{mFDR}(\Lambda) = \text{mFDR}(\Lambda_{OPT}) \), we have

\[
\begin{align*}
(a) & \quad F(\Lambda_{OPT}) \geq F(\Lambda); \\
(b) & \quad F_1(\Lambda_{OPT}) \geq F_1(\Lambda); \\
(c) & \quad \text{mFNR}(\Lambda_{OPT}) \leq \text{mFNR}(\Lambda).
\end{align*}
\]

That is, for any optimal rejection region constructed by thresholding \( \text{Lfdr}(x, y) \), there does not exist a better rejection region, in terms of the average number of rejections, or power, or mFNR, than the optimal rejection region given the same level of mFDR.

**Remark 1.** Du and Zhang (2014) derive the optimal rejection region based on bivariate \( p \)-values in their Proposition 1. Parts (a) and (b) of
Theorem 1 can be viewed as an extension of their result to general statistics and covariates. Similar to their proposition, we can relax the condition of $\text{mFDR}(\Lambda) = \text{mFDR}(\Lambda_{\text{OPT}})$ for our parts (a) and (b) to $\text{mFDR}(\Lambda) \leq \text{mFDR}(\Lambda_{\text{OPT}})$ by finding a rejection region $\Lambda'$ such that $\Lambda \subseteq \Lambda'$ and $\text{mFDR}(\Lambda') = \text{mFDR}(\Lambda_{\text{OPT}})$.

Theorem 1 implies that $L_{\text{fdr}}(x, y)$ is the optimal statistic to rank hypotheses for rejection. However, the $L_{\text{fdr}}(x, y)$ formula in (2) is difficult to evaluate. For example, the joint null density $f_0(x, y)$ is not easy to characterize. Furthermore, the marginal null density of the covariate vector $X$, $f_0(x)$, is unknown and cannot even be assumed to belong to a certain distribution family.

In many multiple testing applications, the null distribution of the primary statistic is unaffected by the covariates. In our psoriasis example, the microarray data and the gene length information exist before the data collection and will not affect the null distribution. In this article, we assume the null independence condition that covariates are independent of the primary statistic under the null hypothesis. Under such condition, $f_0(y)$ is typically known or can be estimated. Therefore, we explore the conditional approach that characterizes the distribution of $Y$ given $X$.

If the true null probability depends on the covariates, we can formulate the following Model II:

$$\left\{ \begin{array}{l} \theta_i | X_i \sim \text{Bernoulli}(\pi_0(X_i)), \\ Y_i | \theta_i \sim \theta_i F_0 + (1 - \theta_i) F_1, \end{array} \right.$$ 

where $\pi_0(\cdot)$ is the true null probability function of the covariates, $\theta_i$’s are independent given $X_i$’s, and $Y_i$’s are independent given $\theta_i$’s. The conditional pdf of $Y$ given $X$ can be written as $f(y|x) = \pi_0(x)f_0(y) + (1 - \pi_0(x))f_1(y)$. Model II is considered in Qu et al. (2012) and Scott et al. (2015).

Furthermore, if the alternative distribution also depends on the covariates, we have Model III as follows:

$$\left\{ \begin{array}{l} \theta_i | X_i \sim \text{Bernoulli}(\pi_0(X_i)), \\ Y_i | \theta_i, X_i \sim \theta_i F_0 + (1 - \theta_i) F_1 X_i, \end{array} \right.$$ 

where $F_1 X_i$ is the covariate regulated alternative cdf, $\theta_i$’s are independent given $X_i$’s, and $Y_i$’s are independent given $\theta_i$’s and $X_i$’s. The conditional pdf of $Y$ given $X = x$ is $f(y|x) = \pi_0(x)f_0(y) + (1 - \pi_0(x))f_1(y|x)$, where $f(\cdot|\cdot)$ denotes the conditional density. Model III is the most general model among models I–III, and the models I and II are special cases of Model III.
Under Model III, define the conditional local false discovery rate (CLfdr) as
\[
\text{CLfdr}(y|x) \equiv \Pr(\theta = 1|x, y) = \frac{\pi_0(x)f_0(y)}{\pi_0(x)f_0(y) + (1 - \pi_0(x))f_1(y|x)}.
\]

(3)

Although CLfdr is defined conditionally on the covariate vector \(X\), it can be shown that CLfdr is equivalent to the joint local FDR in (2).

**Theorem 2.** Under Model I* and the null independence condition,
\[
\text{Lfdr}(x, y) = \text{CLfdr}(y|x).
\]

Theorem 2 unifies the joint modeling approach and the conditional modeling approach under the null independence condition. It also indicates that, rather than Model II, Model III is the appropriate conditional model to use.

For notation simplicity, we use \(\gamma\) to denote the generic CLfdr\((y|x)\), with a subscript for the statistic index, and a superscript for the submodel index, if necessary. Thus, \(\gamma_i = \text{CLfdr}(Y_i|X_i)\). As the models I and II are special cases of Model III, the CLfdr formula in (3) can be simplified in submodels. For example, under Model II,
\[
\gamma_i^{II} \equiv \text{CLfdr}^{II}(Y_i|X_i) = \frac{\pi_0(X_i)f_0(Y_i)}{\pi_0(X_i)f_0(Y_i) + (1 - \pi_0(X_i))f_1(Y_i)}.
\]

Theorems 1 and 2 suggest the following optimal procedure: Arrange \(\gamma\) values in ascending order such that \(\gamma(1) \leq \ldots \leq \gamma(m)\). For a fixed target FDR level \(\alpha\), let \(k = \max\{i : \frac{1}{i} \sum_{j=1}^{i} \gamma(j) \leq \alpha\}\), and we reject the first \(k\) hypotheses with the smallest \(\gamma\) values.

The optimal procedure is an oracle procedure because it depends on the optimal (oracle) statistics \(\gamma_i\)’s. To illustrate the difficulty to estimate \(\gamma_i\)’s, we now make some connections with the problem of grouped hypotheses testing. If hypotheses reside in predefined groups, then the group label can be considered as a categorical covariate \(X\), and the optimal ranking statistic can also be expressed in the form of CLfdr\((y|x) = \Pr(\theta = 1|x, y)\) according to Cai and Sun (2009). To reliably estimate the CLfdr’s, there are presumably a limited number of groups each of which contains a large number of observations. In our case, covariates are continuous, and it is unlikely for hypotheses to have identical covariate values. If we group hypotheses by unique covariate values, each hypothesis will form a separate group with its individual true null probability and alternative density. Therefore, the estimation of CLfdr is much more challenging in the case of continuous covariates.
2.2. Nonparametric Estimation of CLfdr. We propose to estimate both $\pi_0(\cdot)$ and $f_1(\cdot|\cdot)$ nonparametrically. As directly estimating $\pi_0(\cdot)$ or $f_1(\cdot|\cdot)$ at each data point is not feasible, a reasonable approach is to borrow strength from neighbors by assuming that $\pi_0(\cdot)$ and $f_1(\cdot|\cdot)$ are gradually changing with respect to the covariate vector $X$.

Because it is difficult to simultaneously estimate $\pi_0(\cdot)$ and $f_1(\cdot|\cdot)$, we will approach the problem in a step-wise fashion. The basic idea is to explore the hierarchy of the models I–III: Model I is a special case of Model II, and both models assume a common $f_1(\cdot)$, which could be estimated relatively easily from the simpler Model I; then fixing $f_1(\cdot)$, we can estimate $\pi_0(\cdot)$ under Model II, which in turn is a special case of Model III; finally, fixing $\pi_0(\cdot)$, we can estimate $f_1(\cdot|\cdot)$ under Model III. We now describe these three steps in details.

**Step 1: Estimation of $f_1(\cdot)$ under Model I**

We start with a global true null probability estimate $\hat{\pi}_0$. The choice of $\pi_0$-estimator necessarily depends on the data generation process that leads to the original primary statistics. For example, if the primary statistics are z-values generated from the normal means model (Efron, 2004; Jin, 2008), the consistent estimator of Jin and Cai (2007) would be a good candidate. For more discussions of $\pi_0$-estimators, see Remark 2. The overall density $f$ can be estimated by regular kernel density estimator, then

$$
\hat{\gamma}_i^I = \frac{\hat{\pi}_0 f_0(Y_i)}{\hat{f}(Y_i)}.
$$

That is, the posterior probability of $Y_i$ being alternative can be estimated as $1 - \hat{\gamma}_i^I$, $i = 1, \ldots, m$. Then we simply estimate $f_1(\cdot)$ through a weighted kernel method with weights proportional to the values of $(1 - \hat{\gamma}_i^I)$’s. More specifically, let $K_h(\cdot) = h^{-1}K(\cdot/h)$ where $K$ is a smooth and symmetric univariate kernel density function. The functional form of the kernel is not crucial for either the density estimation here or the conditional density estimation later, and we will use the Gaussian kernel hereafter. Furthermore,

$$
\hat{f}_1(y) \equiv \hat{f}_{h_1}(y) = \sum_{i=1}^{m} w_i K_{h_1}(Y_i - y),
$$

where $w_i = (1 - \hat{\gamma}_i^I)/\sum_{j=1}^{m}(1 - \hat{\gamma}_j^I)$. The bandwidth $h_1$ is chosen using Silverman’s ‘rule of thumb’ (Silverman, 1986).

**Remark 2.** Most $\pi_0$-estimators are p-value based. Under independence and mild dependence conditions, the conservative $\pi_0$-estimator of Liang and
Nettleton (2012) can be used. Under moderate to strong dependence conditions, Blanchard and Roquain (2009) recommend the Storey-$\alpha$ estimator according to their simulation results. Other $\pi_0$-estimators with good theoretical or empirical properties can also be used, such as Langaas et al. (2005), Meinshausen et al. (2006), and Patra and Sen (2016), among many others.

Step 2: Estimation of $\pi_0(\cdot)$ under Model II

To estimate $\pi_0(\cdot)$, we use an EM-like algorithm with initial values $\hat{\pi}_0^{(0)}(X_i) = \hat{\pi}_0, i = 1, \ldots, m$. In the E-step, we compute

$$\hat{\gamma}_{II}^{(l)}(i) = \frac{\hat{\pi}_0^{(l)}(X_i)f_0(Y_i)}{\hat{\pi}_0^{(l)}(X_i)f_0(Y_i) + (1 - \hat{\pi}_0^{(l)}(X_i))\hat{f}_1(Y_i)},$$

where $l$ indicates the $l$th iteration.

In the M-step, we model $\pi_0(\cdot)$ nonparametrically through low rank thin plate regression splines (Wood, 2003, 2017), where the rank is chosen by generalized cross-validation (Craven and Wahba, 1978). Ideally, we want the average of $\hat{\pi}_0^{(l+1)}(X_i)$ values to match our initial $\hat{\pi}_0$ to stabilize the global $\pi_0$ between iterations and preserve good statistical properties of the initial $\hat{\pi}_0$. In practice, it may not be the case, and we take a simple approach to shift the values of $\hat{\pi}_0^{(l+1)}(X_i)$'s such that their average would equal to the initial $\hat{\pi}_0$. We call this the backfitting step because it is similar to the backfitting algorithm for additive models. More specifically, let $\zeta = \frac{1}{m} \sum_{i=1}^{m} \hat{\pi}_0^{(l+1)}(X_i) - \hat{\pi}_0$, and $\hat{\pi}_0^{(l+1)}(X_i) = \hat{\pi}_0^{(l+1)}(X_i) - \zeta$. If after backfitting, some values of $\hat{\pi}_0^{(l+1)}(X_i)$ are outside of $[0, 1]$, then the adjustment can be done on the logit scale.

Then we iterate between the E and M steps until the change in likelihood between two consecutive iterations is less than a certain threshold, and our CLfdr estimate for the $i$th observation under Model II is

$$\hat{\gamma}_{II}^i = \frac{\hat{\pi}_0(X_i)f_0(Y_i)}{\hat{\pi}_0(X_i)f_0(Y_i) + (1 - \hat{\pi}_0(X_i))\hat{f}_1(Y_i)}.$$

Remark 3. The EM-like algorithm is in a similar spirit to the algorithm in Young and Hunter (2010) under the mixture of regression models. We do not assume linear regression components in our model, and we also have the extra backfitting step. This is because the alternative density $f_1$ is estimated based on the initial $\hat{\pi}_0$ in Step 1, and the backfitting step is merely enforcing self-consistency. If $f_1$ is completely known, the backfitting step becomes unnecessary.
Remark 4. Other nonparametric regression methods, such as local polynomial regression, can be used in the M-step as well. Though it is unlikely, in case some \( \hat{\pi}_0^{(l+1)}(X_i) \) values fall outside of \([0, 1]\), we set them to their nearest bound of 0 or 1. Alternatively, we can use nonparametric regression on the logit scale, similar to Qu et al. (2012). However, we prefer our method because, first of all, it is more difficult to select tuning parameters and perform the backfitting step on the logit scale. Secondly, as Section 5.10 of Wasserman (2006) indicates, fitting nonparametric logistic model or local linear model makes a minor difference in practice.

Step 3: Estimation of \( f_1(\cdot|\cdot) \) under Model III

To estimate \( f_1(\cdot|\cdot) \), we combine the double kernel idea from conditional density literature (Rosenblatt, 1969) and posterior probability weighting idea from Step 1:

\[
\hat{f}_{1h}(y|x) = \frac{\hat{f}_1(x, y)}{\hat{f}_1(x)} = \frac{\sum_{i=1}^{m} w_i K_{h_3}(X_i - x) K_{hy}(Y_i - y)}{\sum_{i=1}^{m} w_i K_{h_3}(X_i - x)}.
\]

where \( K_{h_3}(x) = \prod_{j=1}^{p} K_{h_{3j}}(x_j) \) and \( w_i = (1 - \hat{\gamma}_I^I_i)/(\sum_{j=1}^{m}(1 - \hat{\gamma}_j^I)) \).

The bandwidth vector \( h = (h_{31}, \ldots, h_{3p}, h_y) \) are chosen by cross-validation to minimize the following error measure. Similar to Fan and Yim (2004) and Hall et al. (2004), we define the integrated squared errors as

\[
\text{ISE} = \int [\hat{f}_{1h}(y|x) - f_1(y|x)]^2 g(x) dx dy
\]

\[
= \int \hat{f}_{1h}^2(y|x) g(x) dx dy - 2 \int \hat{f}_{1h}(y|x) f_1(y|x) g(x) dx dy + \int f_1^2(y|x) g(x) dx dy
\]

\[
= I_1 + I_2 + I_3,
\]

where \( g(\cdot) \) is the marginal density for \( x \).

Note that \( I_3 \) does not depend on \( h \) and will be ignored afterwards. The first term \( I_1 \) can be estimated by

\[
\hat{I}_1 = \frac{1}{m} \sum_{i=1}^{m} \int \hat{f}_{1h_{-i}}(y|X_i) dy,
\]

where \( \hat{f}_{1h_{-i}}(y|x) \) is the leave-one-out estimate of conditional density based on \( \{(X_j, Y_j), j \neq i\} \). It can be shown that

\[
I_2 = -2 \int \hat{f}_{1h}(y|x) \frac{f_1(y|x)}{f(y|x)} f(y|x) g(x) dx dy
\]

\[
= -2 \int \hat{f}_{1h}(y|x) \frac{1 - \gamma}{1 - \pi_0(x)} f(x, y) dx dy,
\]
where the last step can be derived from (3). Then it suggests the following estimator
\[
\hat{I}_2 = -2 \frac{1}{m} \sum_{i=1}^{m} \frac{1 - \hat{\gamma}_{i}^I}{1 - \hat{\pi}_0(X_i)} \hat{f}_{1h_{-i}}(Y_i | X_i).
\]

Finally, we update the estimate of \(\pi_0(\cdot)\) with \(\hat{f}_1(\cdot | \cdot)\), similar to what is done in Step 2. Technically, we could iterate the estimation of \(\pi_0(\cdot)\) and \(f_1(\cdot | \cdot)\) till some convergence criterion is satisfied. However, we found that more iterations do not necessarily lead to significant improvement of model fitting, and we update the estimate of \(\pi_0(\cdot)\) one time only after the estimation of \(f_1(\cdot | \cdot)\) for computational efficiency. The final CLfdr estimate is
\[
\hat{\gamma}_i = \frac{\hat{\pi}_0(X_i) f_0(Y_i)}{\hat{\pi}_0(X_i) f_0(Y_i) + (1 - \hat{\pi}_0(X_i)) \hat{f}_1(Y_i | X_i)}.
\]

3. Simulation Study. We evaluate the performance of different procedures with single covariate and multiple covariates. Throughout this section, the results are based on 400 replications, and by default, we set the target FDR level at 0.1. All procedures assume that a \(z\)-value follow \(N(0, 1)\) when the corresponding null hypothesis is true.

3.1. Candidate Procedures. We consider the following procedures:

\(\text{ORC}\): The oracle procedure where CLfdr’s are computed using the true parameters.

\(\text{CLfdr}\): Our procedure that mimics the \(\text{ORC}\) by using the optimal procedure with \(\hat{\gamma}_i\)’s estimated as in (6) under Model III. By default, we estimate the initial \(\pi_0\) by \(\hat{\pi}^{RB}_0\) using the right-boundary (RB) procedure (Liang and Nettleton, 2012). If data are generated according to the normal mean model, we instead estimate \(\pi_0\) by \(\hat{\pi}^{JC}_0\) using the method proposed by Jin and Cai (2007).

\(\text{Lfdr}\): The optimal procedure proposed by Sun and Cai (2007) when covariates are ignored. More specifically, we use the optimal procedure with the \(\hat{\gamma}_i\) values from (4) with \(\hat{\pi}_0 = \hat{\pi}_0^{JC}\) as suggested by Sun and Cai (2007).

\(\text{FDRreg}\): The FDR regression procedure proposed by Scott et al. (2015), which is implemented in the FDRreg R package v0.2 from GitHub (github.com). We used the recommended default method, which estimates the alternative density by predictive recursion (Newton, 2002; Martin and Tokdar, 2012). As in the simulation of Scott et al. (2015), we expanded each covariate in B-spline basis with five equally spaced knots.

\(\text{BH-RB}\): The adaptive BH procedure with \(\pi_0\) estimated by \(\hat{\pi}^{RB}_0\), which is shown to be conservative under independence (Liang and Nettleton, 2012).
3.2. Multiple Covariates. For illustration purpose, we focus on the setting where the covariates are bivariate ($p = 2$). More specifically, we simulated covariates $X_{i1}$ and $X_{i2}$ independently from $\text{Unif}[−1, 1]$. We set $m = 20000$, which is similar to the number of tests in our psoriasis application.

3.2.1. Set 1: normal mean model. We set the true null probability function $\pi_0(x) = 0.6 + 0.3(1−\omega)(x_1^2 + x_2^2 − 2/3) − 0.375\omega x_1 x_2$, where the average true null probability is 0.6 and similar to the estimate in our application in Section 4. The term $x_1^2 + x_2^2 − 2/3$ represents the additive effect of $x_1$ and $x_2$, and the term $x_1 x_2$ represents the interaction effect between $x_1$ and $x_2$. We let $\omega = 0, 0.25, 0.5, 0.75, \text{and } 1$ to represent different proportion of interaction effect. The multiplying constants before the additive and interaction effects are set such that they have similar contributions to the variation of true null probabilities when $\omega = 0.5$. We simulated the $z$-values under the normal mean model. Under true nulls, $Y_i \sim N(0, 1)$. Under the alternatives, $Y_i \sim N(\mu, 1)$, and $\mu = −1.5$ with probability $\frac{1 + x_1 x_2}{2}$ and $\mu = 3$ if otherwise. The signal effect size is stronger on the positive side ($\mu = 3$) than on the negative side ($\mu = −1.5$) to mimic the situation in our psoriasis example.

Similar as the rest of settings, the graphical illustrations of the $\pi_0(\cdot)$ and $f_1(\cdot|\cdot)$ can be found in the Supplementary Material Section 2. Figure 3(a) plots the average realized FDR as a function of $\omega$. The FDR levels of $\text{FDRreg}$ can be slightly inflated while the FDR levels of other methods are close to or below the target level. Figure 3(b) shows the power relative to the oracle procedure. $\text{CLfdr}$ is more powerful than $\text{Lfdr}$ and $\text{BH-RB}$ because $\text{CLfdr}$ can utilize the changing $\pi_0(\cdot)$ and $f_1(\cdot|\cdot)$ information. When $\omega = 0$, $\text{FDRreg}$ is slightly more powerful than $\text{CLfdr}$, which can be attributed to $\text{FDRreg}$’s larger and slightly liberal realized FDR level. As $\omega$ increases, the relative power of $\text{FDRreg}$ is decreasing at a faster speed than $\text{CLfdr}$ because the estimation accuracy of $\pi_0(\cdot)$ by $\text{FDRreg}$ is deteriorating as more interaction effect is present. Figure 3(c) plots the mean square error (MSE) of $\pi_0$ estimates on log 10 scale. As $\omega$ increases, the MSE of $\text{FDRreg}$ steadily deteriorates to the levels of constant $\pi_0$-estimators of $\text{Lfdr}$ and $\text{BH-RB}$. That is, although the power loss of $\text{FDRreg}$ may not seem significant due to a large proportion of strong signals ($\mu = 3$), the estimated $\pi_0(\cdot)$ could be far from the truth.

3.2.2. Set 2: two-sample $t$-statistic. In sets 2–4, we simulated data to mimic microarray or RNA-seq studies where the number of subjects is typically small while the number of genes is large. More specifically, for each of the $m = 20000$ genes, we compare the expression levels of $n$ Treatment group subjects to $n$ Control group subjects, which are generated independently
Fig 3. Simulation results for simulation set 1 (left column) and set 2 (right column). First row, realized FDR; second row, power relative to ORC; third row, mean squared error of $\pi_0(\cdot)$ on log 10 scale.
from $N(\mu, 1)$ and $N(0, 1)$, respectively. We set $n = 5$ such that the resulting regular two-sample $t$-statistics have degrees of freedom of $2n - 2 = 8$, which is close to the degrees of freedom in our psoriasis example. We set the group difference $\mu = 0$ if the null hypothesis is true.

In set 2, the true null probability function is the same as in set 1. When the null hypothesis is false, we set $\mu = -1$ with probability $\frac{2 + x_1 + x_2}{4}$ and $\mu = 2$ with probability $\frac{2 - x_1 - x_2}{4}$. The $z$-values can be computed from the $t$-statistics similar to our psoriasis example.

From Figure 3(d), FDR$_{\text{reg}}$ can have 40–50% inflation in realized FDR levels while the FDR levels of other methods are close to or below the target level. Both FDR$_{\text{reg}}$ and Lfdr rely on the normal mean model, which assumes both true null and alternative $z$-values follow the normal distribution. In most applications, the original test statistics may not be normal and need to be transformed to $z$-values. In our psoriasis example, the true null $t$-statistics follow central $t$ distribution, and the corresponding $z$-values follow $N(0, 1)$. However, under the alternative hypothesis, the $t$-statistics follow non-central $t$ distribution and remain non-normal after the transformation. The $\hat{\pi}_0^{JC}$ estimator used in Lfdr requires the variances of alternative statistics to be no less than the null variance, while FDR$_{\text{reg}}$ further restricts the alternative variance to be exactly the same as the null variance. Such strong model assumption under the alternative can lead to liberal $\pi_0$ estimates and the loss of FDR control, and a similar loss of FDR control for FDR$_{\text{reg}}$ is also reported in Ignatiadis et al. (2016). Lfdr is relatively robust to the violation of the normal mean model, and the realized FDR levels of Lfdr is below the target level in this setting.

Among all methods that maintain FDR control, the order of their power is the same as in set 1. From Figure 3(f), FDR$_{\text{reg}}$ has the worst MSE of $\pi_0$ estimates mainly due to its large negative bias in overall $\pi_0$, which is around -0.18 across different values of $\omega$.

Technically, the method proposed by Qu et al. (2012) can be applied in this simulation setting with $t$-statistics. However, because the effect sizes under alternative are not generated from the zero-mean normal distribution as assumed in Qu et al. (2012), their method’s realized FDR levels can more than double the target FDR level, and their results are not shown here.

3.2.3. Set 3: $t$-statistic with varying effect sizes. In set 3 we study the impact of effect size. The set up is the same as set 2 with $n$ subjects in each of Treatment and Control group. The true null probability $\pi_0(x) = 0.6 + 0.08(x_1 + x_2) + 0.2x_1x_2$, a mixture of additive and interaction effects. When the null hypothesis is false, we set $\mu = -\tau$ with probability $\frac{3 + 2x_1 + x_2}{6}$ and $\mu = \tau$
τ with probability \( \frac{3-2x_1-x_2}{6} \), where the signal effect size \( \tau = 1, 1.25, 1.5, 1.75 \), and 2 to represent moderate to strong signals.

Figure 4(a) plots the average realized FDR as a function of \( \tau \). The realized FDR levels of \( \text{FDR}_{\text{reg}} \) can be inflated more than 30% over the target FDR level, and \( \text{Lfdr} \) can have at most 10% inflation in realized FDR levels. Similar to set 2, the inflated FDR levels of \( \text{FDR}_{\text{reg}} \) and \( \text{Lfdr} \) are due to the violation of the normal mean model, and \( \text{Lfdr} \) is less sensitive to such violation than \( \text{FDR}_{\text{reg}} \). Only \( \text{CLfdr} \) and BH-RB maintain FDR control, and \( \text{CLfdr} \) is more powerful than BH-RB by utilizing the information in changing \( \pi_0(\cdot) \) and \( f_1(\cdot|\cdot) \).

3.2.4. Set 4: dependent t-statistic with varying effect sizes. Here we investigate the performance of various methods under dependence. The data are generated similar to set 3 except we form random blocks of genes of size 20 within null genes and alternative genes, respectively. Within each block, gene expressions have an autoregressive order 1 (AR1) correlation structure with the correlation of \( \rho_{|i-j|} \) between the \( i \)th and \( j \)th elements. We set \( \rho = -0.8 \) so that we have both positive and negative correlations of varying magnitude within each block to mimic the genetic correlations within biological pathways.

The realized FDR and relative power are plotted in Figure 4 Panels (c) and (d), and results are very similar to that of set 3. We also tried a larger magnitude of correlation (\( \rho = -0.9 \)) and larger block size (50), and the results are not significantly different from this setting. The results in set 4 suggest that \( \text{CLfdr} \) is robust to the correlation structure that is common in the gene expression study.

Here we briefly summarize the simulation results for multiple covariates. \( \text{FDR}_{\text{reg}} \) can have significantly inflated FDR levels when the original test statistics are not \( z \)-values. Among procedures that maintain proper FDR control, \( \text{CLfdr} \) is the most powerful procedure, and its power stays the closest to \( \text{ORC} \).

3.3. Single Covariate. Here we focus on the settings with a single covariate (\( p = 1 \)). Throughout the single covariate simulation settings, we generated data under the normal mean model, where \( Y_i \sim N(0, 1) \) under true nulls, and the covariate \( X_i \sim \text{Unif}[0, 1], i = 1, \ldots, m \), and we set \( m = 10000 \). We also considered two other methods that can only handle a single covariate. We will compare our method with the independent hypothesis weighting (IHW) method proposed by Ignatiadis et al. (2016), which is implemented in the IHW R package v1.0.2 from Bioconductor (bioconductor.org). On the other hand, the procedure proposed by Ferkingstad et al. (2008) cannot be
Fig 4. Simulation results for simulation set 3 (left column) and set 4 (right column). First row, realized FDR; second row, power relative to ORC.
compared due to its tendency to abort with running errors, which were also observed in Ignatiadis et al. (2016).

3.3.1. Set 5: varying $\pi_0(\cdot)$ only. We generated data under Model II where only the true null probability is allowed to change with the covariate while the alternative density remains constant. More specifically, we set $\pi_0(x) = 0.8 + \delta \sin(2\pi x)$, where $\delta = 0, 0.05, 0.1, 0.15,$ and $0.2$. The alternative density is constant and symmetric with $Y_i \sim N(\mu, 1)$ and $\mu = 3$ or $-3$ with equal probabilities. When $\delta > 0$, the true null probability follows a full cycle of the sine function, and the magnitude of change increases with $\delta$. When $\delta = 0$, the true null probability is constant, and the model degenerates to Model I.

Figure 5(a) shows the bias of realized FDR as a function of $\delta$. The FDR levels of ORC are very close to the nominal level with little bias. IHW is the most conservative method, and the FDR levels of other methods are close to the target level. Figure 5(b) shows the power relative to the oracle procedure as a function of $\delta$, and we can roughly divide all procedures into three groups: CLfdr and FDRreg can fully adapt to varying $\pi_0(\cdot)$ and achieve almost the same power of ORC; Lfdr and BH-RB assume constant $\pi_0$ and lose power as $\delta$ increase; IHW is the least powerful method mainly because its conservativeness. IHW can partially adapt to varying $\pi_0(\cdot)$ by dividing $p$-values into bins according to their corresponding covariate values and assigning different weights to different bins. Through this weighting scheme, IHW adapts to changing $\pi_0(\cdot)$ in a piecewise constant fashion, and its power is less affected by the changing $\delta$ than Lfdr and BH-RB. As the alternative density is constant and symmetric, Lfdr doesn’t hold any power advantage over $p$-value based procedure like BH-RB, and their minor difference can only be explained by their different overall $\pi_0$-estimators.

We emphasize that CLfdr performs similarly as ORC, even though the true model is Model II and CLfdr is overfitting the simulated data with Model III. This is achieved through the automatic bandwidth selection described in Step 3. Across all replications in this set, the median of the bandwidth values for the covariate $X$ is about 80. With the covariate $X$ ranging between 0 and 1, the large bandwidth values effectively smooth out the covariate and lead to an estimation of the common alternative density. That is, the cross-validation method we employ in Step 3 automatically over-smooths irrelevant covariates. This phenomenon is similar to what has been observed in Hall et al. (2004) when choosing smoothing parameters for conditional overall density while ours is for conditional alternative density. Even when $\delta = 0$ and the true model is Model I, the performance of CLfdr stays close to that of ORC, indicating the robustness of our estimating procedures when
3.3.2. Set 6: varying $f_1(\cdot | \cdot)$ only. We set the true null probability to be constant at 0.8. The conditional alternative density is a mixture of two densities with mixing proportion depending on the covariate: $Y_i \sim N(\mu, 1)$ and $\mu \sim xN(2, 0.33^2) + (1-x)N(-2, 0.33^2)$, where the effect sizes $\mu$’s are generated from a mixture of two normal distributions. We evaluate all procedures at varying target FDR levels between 0.05 and 0.3.

Figure 5(c) shows the realized FDR as a function of the target FDR level $\alpha$. IHW is the most conservative procedure, and FDR levels of all other procedures are reasonably close to the target FDR levels. From Figure 5(d), CLfdr dominates all other procedures in terms of power because CLfdr is the only method that can adapt to the changing alternative distribution. The power differences between CLfdr and other procedures decrease when the target FDR level increases. This is simply because the relaxation of the target FDR level leads to less unrejected alternatives and there is less room for power improvement.

3.3.3. Set 7: varying $\pi_0(\cdot)$ and $f_1(\cdot | \cdot)$. We set $\pi_0(x) = 0.4 + 0.5 \sin(\pi x)$ with the conditional alternative density $f_1(y|x) = xN(-\mu, 1) + (1-x)N(\mu, 1)$. The effect size $\mu = 1.5, 2, 2.5,$ and 3 to represent moderate to strong signals.

From Figure 5(e), FDRreg can slightly lose FDR control while the FDR levels of other procedures are close to or below the nominal level. Figure 5(f) shows the relative power as a function of $\mu$: CLfdr stays closest to ORC and dominates all other procedures; FDRreg can adapt to varying $\pi_0(\cdot)$ and has the second best power; IHW and the methods that ignore the covariate, Lfdr and BH-RB, are the least powerful. The relative power differences between CLfdr and other methods increase when the effect size $\mu$ decreases. This is because the additional information from varying $\pi_0(\cdot)$ and $f_1(\cdot | \cdot)$ becomes more important when signals are weak.

To evaluate the ability to rank the false null hypotheses ahead of the true nulls, we draw the partial receiver operating characteristic (ROC) curves of all procedures in Figure 6 for $\mu = 1.5$, where the curves are separated the most. BH-RB and Lfdr have almost identical ROC curves, so only the ROC curve of BH-RB was plotted to avoid overlap. According to Theorems 1 and 2, CLfdr should be the best ranking statistics, and ORC and CLfdr are superior to other procedures. IHW performs the worst among procedures that take the covariate into account, possibly because of its inefficiency due to binning and the use of p-values. Not surprisingly, BH-RB and Lfdr perform the worst overall because they cannot utilize the covariate information. Without considering the covariate, the overall alternative density is symmetric in this model is over-specified.
Fig 5. Simulation results for simulation sets 5–7. Left column: realized average FDR. Right column: power relative to ORC. Reference line for target FDR levels, dotted line.
Fig 6. Partial ROC curves for simulation set 7 with $\mu = 1.5$.

set, and there is no difference between BH-RB and Lfdr in terms of ranking ability.

Here we summarize the simulation results for the single covariate. CLfdr is the most powerful procedure, and its power stays closest to ORC. IHW is the most conservative and least powerful procedure.

4. Application. We now return to the application introduced in Section 1. Recall that three pairs of lesional and nonlesional skin samples were compared in the RNA-seq study of psoriasis in Jabbari et al. (2012), and the read count data can be downloaded from the recount2 project (Collado-Torres et al., 2017) with accession number SRP016583. There are 18151 genes that have at least one non-zero count across all samples. We analyzed RNA-seq data using the limma-voom method (Law et al., 2014) which returns a $t$-statistic for each gene. We choose the limma-voom method because it has been shown to have better FDR control than competing methods for small sample size experiments such as ours, see simulation results in Law et al. (2014) and Benidt and Nettleton (2015). The effective degrees of freedom is 8.8, similar to the degrees of freedom used in simulation sets 2–4. The primary statistics of $z$-values were transformed from the $t$-statistics as in (1).

For each gene, the length of the coding region in the number of nucleotides
(len) can also be obtained from the recount2 project. In RNA-seq, the longer genes tend to have a higher number of observed counts and better power to detect differential expression (Oshlack and Wakefield, 2009). Furthermore, a related study using microarray has been conducted by Gudjonsson et al. (2010) with 58 pairs of lesional and nonlesional skin samples, and the dataset is publicly available in the Gene Expression Omnibus with accession number GSE13355. Only 16490 genes out of 18151 can be found in the microarray study because of the design of the microarray. We analyzed microarray data using the limma method (Smyth, 2004) and obtained t-statistics for the 16490 genes, and we will denote the microarray t-statistics as tval.

There could be scientific interest to perform a meta-analysis to identify genes that are DE in both RNA-seq and microarray studies because of their shared study design of paired lesional and nonlesional skin. However, meta-analysis across different technology platforms can be challenging. As only some of the microarray-measured genes are also measured via RNA-Seq, it is unclear what to do with the genes measured only via RNA-Seq. Even for the genes that are measured on both technology platforms, there could be systematic differences in their DE statuses and effect sizes. Skin samples from the two studies went through different biochemical processing protocols, and resulting data were subject to different statistical normalization steps before statistical analysis. Furthermore, microarray only measures expression levels of specific parts of genes that are designed on microarray chips, and microarray measurement may not be reliable for genes whose expression levels are either too high or low (Kukurba and Montgomery, 2015). Finally, the patient enrollment criteria show marked differences between the two studies. Patients in Jabbari et al. (2012) showed moderate-to-severe psoriasis (> 10% skin affected) and had been off treatment for at least 4 weeks before sample collection. On the other hand, Gudjonsson et al. (2010) enrolled all psoriasis patients with > 1% affected skin and only required 1 week off-treatment before the sample collection. Considering the systematic differences between the two studies, we decide to treat the microarray results as potentially biased and noisy auxiliary information to enhance our analysis of the current RNA-seq data.

In our model fitting, we used the normalized ranks of tval and len as our covariates, because that on the original scale, both covariates have a few extreme values which make the estimation of local conditional density difficult. More specifically, define the two covariates \(x_1 = \text{rank}(tval)/m\) and \(x_2 = \text{rank}(\text{len})/m\). This transformation is a one-to-one monotone transformation, and the covariates have the interpretation of the quantile of the original covariates.
Fig 7. Panel (a), contour plot of true null probabilities for psoriasis data. Panel (b), scatter plot of additional discoveries of CLfdr comparing to BH-RB. $x_1$ is the normalized rank of microarray t-statistic, and $x_2$ is the normalized rank of gene length.

Naturally, we divide the total of 18151 genes into two groups: one group contains the 16490 genes with matching microarray data and another group of 1661 genes without matches. At first, we will analyze the 16490 shared genes with $tval$ and $len$ as covariates. The $z$-values are assumed to follow a standard normal distribution under the null hypothesis of no differential expression, and we estimated the true null probability as $\hat{\pi}_0 = 0.6$. Panel (a) of Figure 7 shows the contour plot of the estimated true null probabilities from our method. The contour plot shows a complex surface of the true null probabilities, suggesting that the additive model may not be suitable. As mid-range of $x_1$ values correspond to small absolute values of $tval$, the contour plot indicates that a gene is most likely to be non-DE if the corresponding $|tval|$ is small and its coding region is short.

Finally, we plot the density histograms of $z$-values grouped according to 3-quantiles of covariate values in Figure 8. The number of genes in each combination is denoted as $n$ in the upper part of each panel. The estimated null densities are plotted as solid black curves, and the estimated alternative densities are plotted as blue dash curves. From top to bottom, the alternative density is clearly shifting from the negative side to the positive side as $tval$ increases. The impact of gene length is most evident on the true null
probability and is highly dependent on \texttt{tval}: when \texttt{tval} is positive (the bottom row), there is little gene length effect; on the other hand, when \texttt{tval} is negative (the top row), longer genes are more likely to be down-regulated. This observation suggests strong interaction between \texttt{tval} and \texttt{len}. We performed an informal test of additivity with the estimated CLfdr values as responses and compared the additive model versus the non-additive model. We obtained a \( p \)-value < 0.001 against the null hypothesis that additive model is adequate, and this result also suggests the strong interaction effect between covariates.

For the group of 1661 genes without matching microarray gene measurements, we analyzed them with only the gene length as the covariate and obtained CLfdr estimate for each gene. To combined results from two groups, we can rank the genes by their CLfdr estimates across groups and decide the rejection threshold according to the optimal procedure in Section 2.1. According to Cai and Sun (2009), this is the optimal combining strategy when hypotheses are grouped.

For comparison, we consider the following procedures: BH, the linear step-up procedure proposed by Benjamini and Hochberg (1995), and BH-RB, the adaptive BH procedure. In Section 1, we cast doubt about the suitability of FDRreg and the method of Qu et al. (2012) to our psoriasis dataset, and here we provide more concrete evidence. First, their assumption of constant alternative density is clearly violated according to Figures 2 and 8. Furthermore, in addition to the inadequacy of the additive model, the zero-mean normal model of noncentrality parameters in Qu et al. (2012) implies the alternative distribution and therefore the overall distribution is symmetric around zero. However, there are 142 \( z \)-values > 4 but only 20 \( z \)-values < \(-4\), and the \( p \)-value from a equal probability Binomial test is almost zero (< \(2 \times 10^{-16}\)). Similarly, FDRreg implements the normal mean model, which can be problematic for \( z \)-values derived from \( t \)-statistics as illustrated in our simulation sets 2–4.

### Table 2

The number of rejections

<table>
<thead>
<tr>
<th>FDR</th>
<th>CLfdr</th>
<th>BH-RB</th>
<th>BH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5627</td>
<td>3035</td>
<td>1836</td>
</tr>
</tbody>
</table>

Table 2 shows the number of rejections at the target FDR level of 0.1. Our CLfdr is the most powerful one and followed by BH-RB and BH. Panel (b) of Figure 7 shows the scatter plot of the covariate values for the additional discoveries made by CLfdr comparing to BH-RB. The data points in Panel (b) show an impressive matching pattern with the contour plot in Panels (a),
Fig 8. Density histograms of z-values across covariate value combinations in psoriasis data.
and it is evident that the additional discoveries are concentrated in the low true null probability area. As expected from our theoretical and simulation results, CLfdr promotes the significances of those genes that are more likely to be DE according to their covariate values.

In real applications, we usually do not know the true statuses of null hypotheses. After an extensive search of the literature, we found a closely related subsequent study of psoriasis by Tsoi et al. (2015), where 27 pairs of lesional and nonlesional skin samples were measured by RNA-seq. The patient enrollment criteria of the new study are very similar to those of the microarray study and are therefore less stringent than those of our psoriasis study. Despite their differences, the new study still has a reference value because of the similarities between the two studies, such as the same technology platform and the same study design of paired skin samples. We again analyzed the new RNA-seq data using the limma-voom method (Law et al., 2014) and obtained the $p$-values of the rejected 5627 genes at FDR level of 0.1 using CLfdr in our current study. Treating all $p$-values larger than 0.5 as coming from true null hypotheses, we can estimate the proportion of the true null hypotheses among these 5627 genes as 8.1% by using the $\pi_0$-estimator proposed in Storey et al. (2004). This estimate suggests that our CLfdr method increases the power by utilizing the covariate information while maintaining proper FDR control.

Our empirical Bayes method also provides for each gene the estimate of the posterior probability of no differential expression. To control the FDR at level 0.1, the average CLfdr estimate of rejected 5627 genes is no greater than 0.1. However, the highest CLfdr estimate of these 5627 genes can be as large as 0.32, which implies some rejected genes may have about one in three chance to be non-DE. This illustrates the drawback of controlling the FDR at a target level, which allows hypotheses with high CLfdr to be rejected. As suggested by Efron (2007), we would advocate a more sensible rejection threshold of CLfdr $\leq$ 0.1 or 0.2, depending on the desired stringency.

Among the genes that are declared significant by our method but not by BH-RB or BH, Interferon Regulatory Factor 7 (IRF7) offers an interesting example. By the RNA-seq data of Jabbari et al. (2012) alone, the $p$-value for IRF7 is 0.059, barely missing the typical single hypothesis cut-off of 0.05. Its significance will be even lower after multiplicity adjustment. On the other hand, our method estimates its CLfdr as 0.01, which means IRF7 is very likely to be DE. Our CLfdr estimate for IRF7 is primarily influenced by its corresponding microarray t-statistic from Gudjonsson et al. (2010), which is 23.8 with about 61 degrees of freedom and leads to a $p$-value of $2 \times 10^{-32}$. This microarray t-statistic translates to $x_1 = 0.995$, and the coding region length
of IRF7 is 2628 nucleotide bases and corresponds to $x_2 = 0.209$. From Figure 7 Panel (a), it can be seen that the covariate values of IRF7 lead to a small prior probability of IRF7 being non-DE. Furthermore, the $z$-value of IRF7 is 1.89, which indicates potentially up-regulated gene expression levels in lesional skin. This is in agreement with the perceived alternative distribution conditional on $t_{val} = 23.8$ from Figure 2, further lending support to the conclusion that IRF7 is up-regulated in lesional skin.

Interferon-α (IFN-α) plays an important role in psoriasis pathogenesis (van der Fits et al., 2004; Nestle et al., 2005). As a gene inducible by IFN-α, IRF7 has been studied extensively in psoriasis studies. For example, van der Fits et al. (2004) show increased IRF7 expression levels in lesional skin compared to nonlesional and healthy skin using reverse transcription polymerase chain reaction (RT-PCR). Yao et al. (2008) find increased IRF7 expression levels in lesional skin compared to paired nonlesional skin using microarray. Finally, IRF7 also shows significantly increased expression levels in lesional skin in the recent RNA-seq study from Tsoi et al. (2015). In summary, IRF7 has been repeatedly shown to have elevated expression levels in lesional skin compared to nonlesional skin. Despite the weak evidence of IRF7 over-expression in lesional skin from the Jabbari et al. (2012) data, we are able to adjust its significance level using known covariate information and arrive at a correct conclusion.

5. Discussion. Our model III implies that the primary statistics are conditionally independent given covariates. The conditional independence is less stringent and a more reasonable assumption than the marginal independence implied by the two-group model. In applications with spatial covariates, such as in neuroimaging and astronomy, our conditional model can provide a reasonable explanation of the spatial clustering of signals with similar strengths. On the other hand, the correlation among true null statistics depend heavily on the specific application under investigation and can pose a serious challenge. Through simulation, we demonstrate that the performance of our procedure is relatively robust to block dependence structure, which mimics the correlation among gene expression levels in pathways. However, this robustness may not generalize well to other types of dependence in applications such as GWAS. For recent investigations into the multiplicity adjustment for GWAS with consideration of linkage disequilibrium, see, for example, Stange et al. (2016) and Brzyski et al. (2017). Some recent studies of the FDR under dependence include Efron (2010), Schwartzman and Lin (2011), Fan et al. (2012), and Fan and Han (2017), among others. We will leave the exploration of the dependent case for future research.
In large-scale multiple testing applications, confounding factors could affect the validity of each individual test statistics and lead to spurious findings. Detection and adjustment for confounding factors should be carefully carried out before the multiplicity adjustment. Interested readers can refer to Leek and Storey (2007), Gagnon-Bartsch and Speed (2012), and Wang et al. (2017) for literature and methods. The confounding issue is less of concern for our psoriasis dataset because paired lesional and nonlesional skin samples were collected from each psoriasis patient in the study. Subject level confounding factors such as age and sex would be balanced by the paired design. Also, our result shows high concordance with a previous microarray study (Gudjonsson et al., 2010) and a subsequent RNA-seq study with a larger sample size (Tsoi et al., 2015).

Our method is a nonparametric empirical Bayes method that estimates the true null probability function using thin plate spline and the conditional alternative density using kernel density. Such methodology works best when the number of tests is large and the number of covariates is small, as in our psoriasis application. On the other hand, these nonparametric methods will encounter difficulty when the number of covariates is large due to the curse of dimensionality. Several adaptations may be employed, for example, the additive model in Qu et al. (2012) can be used for estimating the true null probability function, certain dimension reduction method such as principal component analysis may lead to a more stable result, and recently Tansey et al. (2018) extend the FDR regression framework of Scott et al. (2015) to the high-dimensional setting through a deep neural network. The details will depend on specific applications and will be left for future research.

References.


van der Fits, L., L. van der Wel, J. D. Laman, E. P. Prens, and M. C. Verschuren (2004). In psoriasis lesional skin the type I interferon signaling pathway is activated, whereas interferon-α sensitivity is unaltered. *Journal of Investigative Dermatology* 122(1), 51–60.


E-MAIL: kun.liang@uwaterloo.ca