



Covariate Adaptive False Discovery Rate Control With Applications to Omics-Wide Multiple Testing

Xianyang Zhang & Jun Chen

To cite this article: Xianyang Zhang & Jun Chen (2022) Covariate Adaptive False Discovery Rate Control With Applications to Omics-Wide Multiple Testing, Journal of the American Statistical Association, 117:537, 411-427, DOI: [10.1080/01621459.2020.1783273](https://doi.org/10.1080/01621459.2020.1783273)

To link to this article: <https://doi.org/10.1080/01621459.2020.1783273>



View supplementary material [↗](#)



Published online: 17 Aug 2020.



Submit your article to this journal [↗](#)



Article views: 598



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)



Covariate Adaptive False Discovery Rate Control With Applications to Omics-Wide Multiple Testing

Xianyang Zhang^a and Jun Chen^b

^aDepartment of Statistics, Texas A&M University, College Station, TX; ^bDivision of Biomedical Statistics and Informatics, and Center for Individualized Medicine, Mayo Clinic, Rochester, MN

ABSTRACT

Conventional multiple testing procedures often assume hypotheses for different features are exchangeable. However, in many scientific applications, additional covariate information regarding the patterns of signals and nulls are available. In this article, we introduce an FDR control procedure in large-scale inference problem that can incorporate covariate information. We develop a fast algorithm to implement the proposed procedure and prove its asymptotic validity even when the underlying likelihood ratio model is misspecified and the p -values are weakly dependent (e.g., strong mixing). Extensive simulations are conducted to study the finite sample performance of the proposed method and we demonstrate that the new approach improves over the state-of-the-art approaches by being flexible, robust, powerful, and computationally efficient. We finally apply the method to several omics datasets arising from genomics studies with the aim to identify omics features associated with some clinical and biological phenotypes. We show that the method is overall the most powerful among competing methods, especially when the signal is sparse. The proposed covariate adaptive multiple testing procedure is implemented in the R package CAMT. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received November 2018
Accepted June 2020

KEYWORDS

Covariates; EM-algorithm;
False discovery rate; Multiple
testing

1. Introduction

Multiple testing refers to simultaneous testing of more than one hypothesis. Given a set of hypotheses, multiple testing deals with deciding which hypotheses to reject while guaranteeing some notion of control on the number of false rejections. A traditional measure is the family-wise error rate (FWER), which is the probability of committing at least one Type I error. As the number of trials increases, FWER still measures the probability of at least one false discovery, which is overly stringent in many applications. This absolute control is in contrast to the proportionate control afforded by the false discovery rate (FDR).

Consider the problem of testing m distinct hypotheses. Suppose a multiple testing procedure rejects R hypotheses among which V hypotheses are null, that is, it commits V Type I errors. In the seminal paper by Benjamini and Hochberg, the authors introduced the concept of FDR defined as

$$\text{FDR} = E \left[\frac{V}{R \vee 1} \right],$$

where $a \vee b = \max\{a, b\}$ for $a, b \in \mathbb{R}$, and the expectation is with respect to the random quantities V and R . FDR has many advantageous features comparing to other existing error measures. Control of FDR is less stringent than the control of FWER especially when a large number of hypothesis tests are performed. FDR is also adaptive to the underlying signal structure in the data. The widespread use of FDR is believed to stem from and motivated by the modern technologies which

produce big datasets, with huge numbers of measurements on a comparatively small number of experimental units. Another reason for the popularity of FDR is the existence of a simple and easy-to-use procedure proposed in Benjamini and Hochberg (1995) (the BH procedure, hereafter) to control the FDR at a prespecified level.

Although the BH procedure is more powerful than procedures aiming to control the FWER, it assumes hypotheses for different features are exchangeable which could result in suboptimal power as demonstrated in recent literature when individual tests differ in their true effect size, signal-to-noise ratio or prior probability of being false. In many scientific applications, particularly those from genomics studies, there are rich covariates that are informative of either the statistical power or the prior null probability. These covariates can be roughly derived into two classes: statistical covariates and external covariates (Ignatiadis et al. 2016). Statistical covariates are derived from the data itself and could reflect the power or null probability. Generic statistical covariates include the sample variance, total sample size and sample size ratio (for two-group comparison), and the direction of the effects. There are also specific statistical covariates for particular applications. For example, in transcriptomics studies using RNA-Seq, the sum of read counts per gene across all samples is a statistical covariate informative of power since the low-count genes are subject to more sampling variability. Similarly, the minor allele frequency and the prevalence of the bacterial

species can be taken as statistical covariates for genome-wide association studies (GWAS) and microbiome-wide association studies (MWAS), respectively. Moreover, the average methylation level of a CpG site in epigenome-wide association studies (EWAS) can be a statistical covariate informative of the prior null probability due to the fact that differential methylation frequently occurs in highly or lowly methylated region depending on the biological context. Besides these statistical covariates, there are a plethora of covariates that are derived from external sources and are usually informative of the prior null probability. These external covariates include the deleteriousness of the genetic variants for GWAS, the location (island and shore) of CpG methylation variants for EWAS, and pathogenicity of the bacterial species for MWAS. Useful external covariates also include p -values from previous or related studies which suggest that some hypotheses are more likely to be nonnull than others. Exploiting such external covariates in multiple testing could lead to improved statistical power as well as enhanced interpretability of research results.

Accommodating covariates in multiple testing has recently been a very active research area. We briefly review some contributions that are most relevant to the current work. The basic idea of many existing works is to relax the p -value thresholds for hypotheses that are more likely to be nonnull and tighten the thresholds for the other hypotheses so that the overall FDR level can be controlled. For example, Genovese, Roeder, and Wasserman (2006) proposed to weight the p -values with different weights, and then apply the BH procedure to the weighted p -values. Hu, Zhao, and Zhou (2010) developed a group BH procedure by estimating the proportions of null hypotheses for each group separately, which extends the method in Storey (2002). Li and Barber (2019) generalized this idea by using the censored p -values (i.e., p -values that are greater than a prespecified threshold) to adaptively estimate the weights that can be designed to reflect any structure believed to be present. Ignatiadis et al. (2016) proposed the independent hypothesis weighting (IHW) for multiple testing with covariate information. Their idea is to bin the covariates into several groups and then apply the weighted BH procedure with piecewise constant weights. Boca and Leek (2018) extended the idea by using a regression approach to estimate weights. Another related method (named AdaPT) was proposed in Lei and Fithian (2018), which iteratively estimates the p -value thresholds using partially censored p -values. The above procedures can be viewed to some extent as different variants of the weighted BH procedure. Along a separate line, local FDR (LFDR) based approaches have been developed to accommodate various forms of auxiliary information.

For example, Cai and Sun (2009) considered multiple testing of grouped hypotheses using the pooled LFDR statistic. Sun et al. (2015) developed a LFDR-based procedure to incorporate spatial information. Scott et al. (2015) and Tansey et al. (2017) proposed EM-type algorithms to estimate the LFDR by taking into account covariate and spatial information, respectively. Other relevant methods include Ferkingstad et al. (2008), Dobriban (2017), Zablocki et al. (2014) and Zablocki et al. (2017).

Although the approaches mentioned above excel in certain aspects, a method that is flexible, robust, powerful, and computationally efficient is still lacking. For example, IHW developed in Ignatiadis et al. (2016) cannot handle multiple covariates. AdaPT in Lei and Fithian (2018) is computationally intensive and may suffer from significant power loss when the signal is sparse, and covariate is not very informative. Li and Barber's (2019) procedure is not Bayes optimal as shown in Lei and Fithian (2018) and thus could lead to suboptimal power as observed in our numerical studies. The FDR regression method proposed in Scott et al. (2015) lacks a rigorous FDR control theory. Table 1 provides a detailed comparison of these methods.

In this article, in addition to a thorough evaluation of these methods using comprehensive simulations covering different signal structures, we propose a new procedure to incorporate covariate information with generic applicability. The covariates can be any continuous or categorical variables that are thought to be informative of the statistical properties of the hypothesis tests. The main contributions of our article are 2-fold:

1. Given a sequence of p -values $\{p_1, \dots, p_m\}$, we introduce a general decision rule of the form

$$(1 - k_i)p_i^{-k_i} \geq \frac{(1 - t)\pi_i}{t(1 - \pi_i)}, \quad 0 < k_i < 1, \quad 1 \leq i \leq m, \quad (1)$$

which serves as a surrogate for the optimal decision rule derived under the two-component mixture model with varying mixing probabilities and alternative densities. Here π_i and k_i are parameters that can be estimated from the covariates and p -values, and t is a cutoff value to be determined by our FDR control method. We develop a new procedure to estimate (k_i, π_i) and find the optimal threshold value for t in (1). We show that (i) when π_i and k_i are chosen independently of the p -values, the proposed procedure provides finite sample FDR control; (ii) our procedure provides asymptotic FDR control when π_i and k_i are chosen to maximize a potentially misspecified likelihood based on the covariates and p -values; (iii) Similar to some recent works (e.g., Ignatiadis et al. 2016; Lei and Fithian 2018; Li and Barber 2019), our method allows

Table 1. Comparison of several covariate adaptive FDR control procedures in recent literature.

Procedure	π_0	f_1	FDR control	Dependent p -values	Misspec. robust	Multiple covariates	Computation
Ignatiadis et al. (2016)	Varying	Partially used	Asymptotic control	Unknown	Yes	No	++++
Li and Barber (2019)	Varying	Not used	Finite sample upper bound	Gaussian copula	Yes	No ^a	++++
Lei and Fithian (2018)	Varying	Varying	Finite sample control	Unknown	Yes	Yes	+
Scott et al. (2015)	Varying	Fixed	No guarantee	Unknown	Unknown	Yes	+++
Boca and Leek (2018)	Varying	Not used	Unknown	Unknown	Yes	Yes	+++
The proposed method	Varying	Varying	Asymptotic control	Asymptotic	Yes	Yes	+++

NOTE: The number of "+" represents the speed.

^aThe framework of Li and Barber (2019) allows accommodating multiple covariates, but the provided software did not implement.

the underlying likelihood ratio model to be misspecified. A distinctive feature is that our asymptotic analysis does not require the p -values to be marginally independent or conditionally independent given the covariates. More specifically, we allow the pairs of p -value and covariate across different hypotheses to be strongly mixing as specified in [Assumption 3.3](#).

2. We develop an efficient algorithm to estimate π_i and k_i . The developed algorithm is scalable to problems with millions of tests. Through extensive numerical studies, we show that our procedure is highly competitive to several existing approaches in the recent literature in terms of finite sample performance. The proposed procedure is implemented in the R package CAMT.

Our method is related to Lei and Fithian (2018), and it is worth highlighting the differences from their work. (i) Lei and Fithian (2018) used partially censored p -values to determine the threshold, which can discard useful information concerning the alternative distribution of p -values (i.e., $f_{1,i}$ in (3)) since small p -values that are likely to be generated from the alternative are censored. In contrast, we use all the p -values to determine the threshold. Our method is seen to exhibit more power as compared to Lei and Fithian (2018) when signal is (moderately) sparse. Although our method no longer offers theoretical finite sample FDR control, we show empirically that the power gain is not at the cost of FDR control. (ii) Different from Lei and Fithian (2018) which requires multiple stages for practitioners to make their final decision, our method is a single-stage procedure that only needs to be run one time; thus, the implementation of our method is faster and scalable to modern big datasets. (iii) Our theoretical analysis is entirely different from those in Lei and Fithian (2018). In particular, we show that our method achieves asymptotic FDR control even when the p -values are dependent.

2. Methodology

2.1. Rejection Rule

We consider simultaneous testing of m hypotheses H_i for $i = 1, 2, \dots, m$. Let p_i be the p -value associated with the i th hypothesis, and with some abuse of notation, let H_i indicate the underlying truth of the i th hypothesis. In other words, $H_i = 0$ if the i th hypothesis is true and $H_i = 1$ otherwise. For each hypothesis, we observe a covariate x_i lying in some space $\mathcal{X} \subseteq \mathbb{R}^q$ with $q \geq 1$. From a Bayesian viewpoint, we can model H_i given x_i as a Bernoulli random variable with success probability $1 - \pi_{0i}$, where π_{0i} denotes the prior probability that the i th hypothesis is under the null when conditioning on x_i . One approach to model the p -value distribution is via a two-component mixture model,

$$H_i|x_i \sim \text{Bernoulli}(1 - \pi_{0i}), \quad (2)$$

$$p_i|x_i, H_i \sim (1 - H_i)f_0 + H_i f_{1,i}, \quad (3)$$

where f_0 and $f_{1,i}$ are the density functions corresponding to the null and alternative hypotheses, respectively. In the following discussions, we shall assume that f_0 satisfies the following condition: for any $a \in [0, 1]$

$$\int_0^a f_0(x)dx \leq \int_{1-a}^1 f_0(x)dx. \quad (4)$$

This condition relaxes the assumption of uniform distribution on the unit interval. It is fulfilled when f_0 is nondecreasing or f_0 is symmetric about 0.5 (in which case the equality holds in (4)). We demonstrate that this relaxation is capable of describing plausible data generating processes that would create a nonuniform null distribution. Let T be a test statistic such that under the null its z -score $Z = (T - \mu_0)/\sigma_0$ is standard normal. In practice, one uses $\hat{\mu}$ and $\hat{\sigma}$ to estimate μ_0 and σ_0 , respectively. Let Φ be the standard normal CDF. The corresponding one-sided p -value is given by $\Phi((T - \hat{\mu})/\hat{\sigma})$ whose distribution function is $P(\Phi((T - \hat{\mu})/\hat{\sigma}) \leq x) = \Phi((\Phi^{-1}(x)\hat{\sigma} + \hat{\mu} - \mu_0)/\sigma_0)$. When $\mu_0 \geq \hat{\mu}$ (i.e., we underestimate the mean), one can verify that f_0 is a non-decreasing. In the case of $\mu_0 = \hat{\mu}$ and $\sigma_0 \neq \hat{\sigma}$, f_0 is non-uniformly symmetric about 0.5.

Compared to the classical two-component mixture model, the varying null probability reflects the relative importance of each hypothesis given the external covariate information x_i and the varying alternative density $f_{1,i}$ emphasizes the heterogeneity among signals. In the context without covariate information, it is well known that the optimal rejection is based on the LFDR (see, e.g., Efron 2004; Sun and Cai 2007). The result has been generalized to the setups with group or covariate information (see, e.g., Cai and Sun 2009; Lei and Fithian 2018). Based on these insights, one can indeed show that the optimal rejection rule that controls the expected number of false positives while maximizes the expected number of true positives takes the form of

$$\frac{f_{1,i}(p_i)}{f_0(p_i)} \geq \frac{(1-t)\pi_{0i}}{t(1-\pi_{0i})}, \quad (5)$$

where $t \in (0, 1)$ is a cutoff value. This decision rule is generally unobtainable because $f_{1,i}$ is unidentifiable without extra assumptions on its form. Moreover, consistent estimation of the decision rule (5) is difficult, and even with the use of additional approximations, such as splines or piecewise constant functions. In this work, we do not aim to estimate the optimal rejection rule directly. Instead, we try to find a rejection rule that can mimic some useful operational characteristics of the optimal rule. Our idea is to first replace $f_{1,i}/f_0$ by a surrogate function h_i . We emphasize that h_i needs not agree with the likelihood ratio $f_{1,i}/f_0$ for our method to be valid. In fact, the validity of our method does not rely on the correct specification of models (2) and (3). We require h_i to satisfy (i) $h_i(p) \geq 0$ for $p \in [0, 1]$; (ii) $\int_0^1 h_i(p)dp = 1$; (iii) h is decreasing. Requirement (iii) is imposed to mimic the common likelihood ratio assumption in the literature (see, e.g., Sun and Cai 2007). In this article, we suggest to use the beta density,

$$h_i(p) = (1 - k_i)p^{-k_i}, \quad 0 < k_i < 1, \quad (6)$$

where k_i is a parameter that depends on x_i . Suppose that under the null hypothesis, p_i is uniformly distributed, whereas under the alternative, it follows a beta distribution with parameters $(1 - k_i, 1)$, then the true likelihood ratio would take exactly the form given in (6). To demonstrate the approximation of the proposed surrogate likelihood ratio to the actual likelihood ratio for realistic problems, we simulated two binary variables and generated four alternative distributions $f_{1,i}$ depending on the four levels of the two variables (details in the legend of

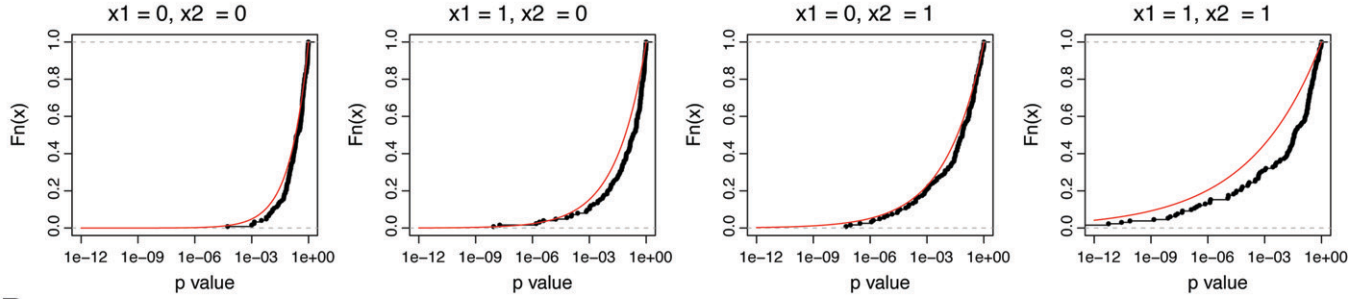
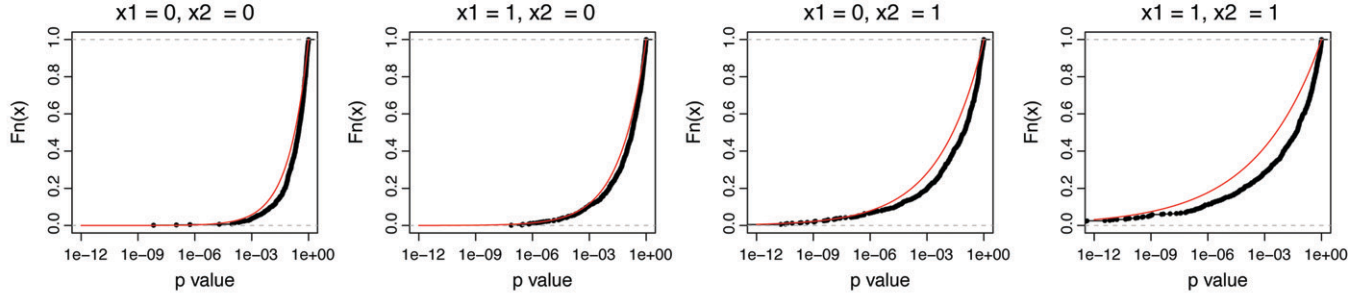
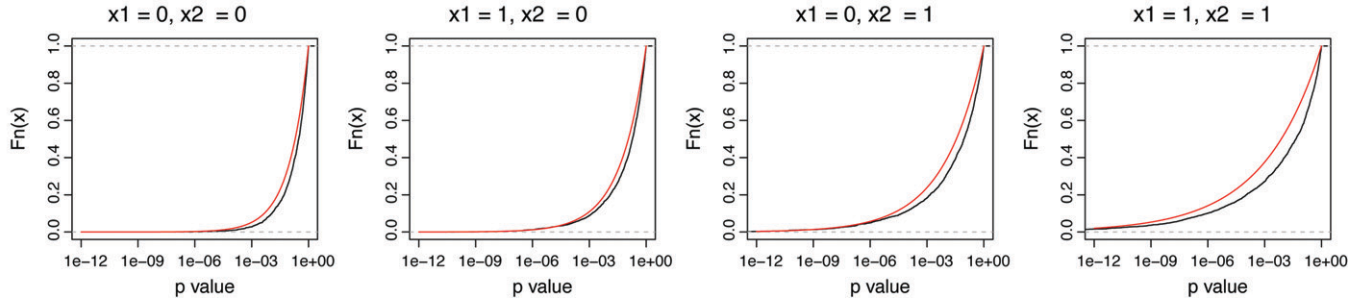
A 99% null, $m=50,000$, four-component f1**B** 95% null, $m=50,000$, four-component f1**C** 80% null, $m=50,000$, four-component f1

Figure 1. Illustration of the approximation to the true likelihood ratio by the surrogate likelihood ratio based on a beta distribution. Two binary covariates x_1 and x_2 were simulated. The z-score under the alternative was drawn from $N(0, 1.5 + 0.5x_1 + x_2)$. Three levels of null proportions (A: 99%, B: 95%, and C: 80%) were simulated, where the null z-score was drawn from $N(0, 1)$. Two-sided p -values were calculated based on the z-score and the parameter k_i of the beta distribution was estimated by CAMT. The CDF of the empirical distribution of the p -value under the alternative (black) was compared to CDF of the fitted beta-distribution (red).

Figure 1). We used the proposed procedure to find the best k_i and compared the CDF of the empirical distribution (reflecting the actual likelihood ratio) to that of the fitted beta distribution (reflecting the surrogate likelihood ratio). We can see from Figure 1 the approximation was reasonably well and the accuracy increases with the signal density.

Based on the surrogate likelihood ratio, the corresponding rejection rule is given by

$$h_i(p_i) \geq w_i(t) := \frac{(1-t)\pi_i}{t(1-\pi_i)}, \quad (7)$$

for some weights π_i to be determined later. See Section 2.3 for more details about the estimation of k_i and π_i .

2.2. Adaptive Procedure

We first note that the false discovery proportion (FDP) associated with the rejection rule (7) is equal to

$$\text{FDP}(t) := \frac{\sum_{i=1}^m (1 - H_i) \mathbf{1}\{h_i(p_i) \geq w_i(t)\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}}.$$

Then for a cutoff value t , we have

$$\begin{aligned} \text{FDP}(t) &= \frac{\sum_{i=1}^m (1 - H_i) \mathbf{1}\{p_i \leq h_i^{-1}(w_i(t))\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} \\ &\approx \frac{\sum_{i=1}^m (1 - H_i) P(p_i \leq h_i^{-1}(w_i(t)))}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} \\ &\leq \frac{\sum_{i=1}^m (1 - H_i) P(1 - p_i \leq h_i^{-1}(w_i(t)))}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} \\ &\approx \frac{1 + \sum_{i=1}^m (1 - H_i) \mathbf{1}\{h_i(1 - p_i) \geq w_i(t)\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} \\ &\leq \frac{1 + \sum_{i=1}^m \mathbf{1}\{h_i(1 - p_i) \geq w_i(t)\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} := \text{FDP}_{\text{up}}(t), \end{aligned}$$

where the approximations are due to the law of large numbers and the inequality follows from Condition (4).¹ This strategy is partly motivated by the recent distribution-free method proposed in Barber and Candès (2015). We refer any FDR estimator constructed using this strategy as the BC-type estimator. Both

¹Rigorous theoretical justifications are provided in Theorems 2.1 and 3.1.

the adaptive procedure in Lei and Fithian (2018) and the proposed method fall into this category. A natural idea is to select the largest threshold such that $\text{FDP}_{\text{up}}(t)$ is less or equal to a prespecified FDR level α . Specifically, we define

$$t^* = \max \left\{ t \in [0, t_{\text{up}}] : \text{FDP}_{\text{up}}(t) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{h_i(1 - p_i) \geq w_i(t)\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{h_i(p_i) \geq w_i(t)\}} \leq \alpha \right\},$$

where t_{up} satisfies that $w_i(t_{\text{up}}) \geq h_i(0.5)$ for all i , and we reject all hypotheses such that $h_i(p_i) \geq w_i(t^*)$. The following theorem establishes the finite sample control of the above procedure when π_i and h_i are prespecified and thus independent of the p -values. For example, π_i and h_i are estimated based on data from an independent but related study.

Theorem 2.1. Suppose h_i is strictly decreasing for each i and f_0 satisfies Condition (4). If the p -values are independent and the choice of h_i and π_i is independent of the p -values, then the adaptive procedure provides finite sample FDR control at level α .

2.3. An Algorithm

The optimal choices of π_i and k_i are rarely known in practice, and a generally applicable data-driven method is desirable. In this subsection, we propose an EM-type algorithm to estimate π_i and k_i . In particular, we model both π_i and k_i as functions of the covariate x_i . As an illustration, we provide the following example.

Example 2.1. Suppose

$$p_i | x_i, H_i \sim (1 - H_i)f_0 + H_i f_{1,i}, \\ x_i | H_i \sim (1 - H_i)g_0 + H_i g_1,$$

where $H_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(1 - \pi_0)$. Using the Bayes rule, we have

$$f(p_i | x_i) = \frac{f(p_i, x_i | H_i = 0)\pi_0 + f(p_i, x_i | H_i = 1)(1 - \pi_0)}{f(x_i | H_i = 0)\pi_0 + f(x_i | H_i = 1)(1 - \pi_0)} \\ = \frac{f(p_i | x_i, H_i = 0)f(x_i | H_i = 0)\pi_0 + f(p_i | x_i, H_i = 1)f(x_i | H_i = 1)(1 - \pi_0)}{f(x_i | H_i = 0)\pi_0 + f(x_i | H_i = 1)(1 - \pi_0)} \\ = \pi(x_i)f_0(p_i) + (1 - \pi(x_i))f_{1,i}(p_i),$$

where $\pi(x) = g_0(x)\pi_0 / \{g_0(x)\pi_0 + g_1(x)(1 - \pi_0)\} = f(H_i = 0 | x_i = x)$. Therefore, π_i is the conditional probability that the i th hypothesis is under the null given the covariate x_i .

To motivate our estimation procedure for π_i and k_i , let us define $\pi_\theta(x) = 1/(1 + e^{-\theta_0 - \theta_1'x})$ and $k_\beta(x) = 1/(1 + e^{-\beta_0 - \beta_1'x})$ for $x \in \mathbb{R}^q$, where $\theta = (\theta_0, \theta_1)$ and $\beta = (\beta_0, \beta_1)$. Suppose that conditional on x_i and marginalizing over H_i ,

$$f(p_i | x_i) = \pi_\theta(x_i)f_0(p_i) + (1 - \pi_\theta(x_i))f_{1,i}(p_i) \\ = f_0(p_i) \left\{ \pi_\theta(x_i) + (1 - \pi_\theta(x_i)) \frac{f_{1,i}(p_i)}{f_0(p_i)} \right\}.$$

Replacing $f_{1,i}/f_0$ by the surrogate likelihood ratio whose parameters k_i depend on x_i , we obtain

$$\tilde{f}(p_i | x_i) = f_0(p_i) \left\{ \pi_\theta(x_i) + (1 - \pi_\theta(x_i))(1 - k_\beta(x_i))p_i^{-k_\beta(x_i)} \right\}.$$

Moving to a log scale and summing up the individual log-likelihoods, we see that the null density is a nuisance parameter that does not depend on θ and β :

$$\sum_{i=1}^m \log \tilde{f}(p_i | x_i) \\ = \sum_{i=1}^m \log \left\{ \pi_\theta(x_i) + (1 - \pi_\theta(x_i))(1 - k_\beta(x_i))p_i^{-k_\beta(x_i)} \right\} + C_0,$$

where $C_0 = \sum_{i=1}^m \log f_0(p_i)$. The above discussions thus motivate the following optimization problem for estimating the unknown parameters:

$$\max_{\theta=(\theta_0, \theta_1)' \in \Theta, \beta=(\beta_0, \beta_1)' \in \mathcal{B}} \sum_{i=1}^m \log \{ \pi_i + (1 - \pi_i)(1 - k_i)p^{-k_i} \}, \quad (8)$$

where

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \theta_0 + \theta_1'x_i, \quad \log \left(\frac{k_i}{1 - k_i} \right) = \beta_0 + \beta_1'x_i, \quad (9)$$

and $\Theta, \mathcal{B} \subseteq \mathbb{R}^{q+1}$ are some compact parameter spaces. This problem can be solved using the EM-algorithm together with the Newton's method in its M-step. Let $\hat{\theta}$ and $\hat{\beta}$ be the maximizer from (8). Define

$$\hat{\pi}_i = W(1/(1 + e^{-\tilde{x}_i'\hat{\theta}}), \epsilon_1, \epsilon_2) \\ := \begin{cases} \epsilon_1, & \text{if } 1/(1 + e^{-\tilde{x}_i'\hat{\theta}}) \leq \epsilon_1, \\ 1/(1 + e^{-\tilde{x}_i'\hat{\theta}}), & \text{if } \epsilon_1 < 1/(1 + e^{-\tilde{x}_i'\hat{\theta}}) < 1 - \epsilon_2, \\ 1 - \epsilon_2, & \text{otherwise,} \end{cases}$$

and $\hat{k}_i = 1/(1 + e^{-\tilde{x}_i'\hat{\beta}})$ with $\tilde{x}_i = (1, x_i)'$, and

$$\hat{w}_i(t) = \frac{(1 - t)\hat{\pi}_i}{t(1 - \hat{\pi}_i)}.$$

We use winsorization to prevent $\hat{\pi}_i$ from being too close to zero. In numerical studies, we found the choices of $\epsilon_1 = 0.1$ and $\epsilon_2 = 10^{-5}$ perform reasonably well. Further denote

$$\hat{t} = \max \left\{ t \in [0, 1] : \frac{1 + \sum_{i=1}^m \mathbf{1}\{(1 - \hat{k}_i)(1 - p_i)^{-\hat{k}_i} > \hat{w}_i(t)\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{(1 - \hat{k}_i)p_i^{-\hat{k}_i} \geq \hat{w}_i(t)\}} \leq \alpha \right\}.$$

Then we reject the i th hypothesis if

$$(1 - \hat{k}_i)p_i^{-\hat{k}_i} \geq \hat{w}_i(\hat{t}).$$

Remark 2.1. We can replace $x_i \in \mathbb{R}^q$ by $(g_1(x_i), \dots, g_{q_0}(x_i)) \in \mathbb{R}^{q_0}$ for some transformations (g_1, \dots, g_{q_0}) to allow nonlinearity in the logistic regressions. In numerical studies, we shall consider the spline transformation.

3. Asymptotic Results

3.1. FDR Control

In this subsection, we provide asymptotic justification for the proposed procedure. Note that

$$\begin{aligned} \mathbf{1}\{(1 - \hat{k}_i)p^{-\hat{k}_i} \geq \hat{w}_i(t)\} &= \mathbf{1}\{p \leq c(t, \hat{\pi}_i, \hat{k}_i)\} \text{ for } c(t, \hat{\pi}_i, \hat{k}_i) \\ &= 1 \wedge \left\{ \frac{t(1 - \hat{k}_i)(1 - \hat{\pi}_i)}{(1 - t)\hat{\pi}_i} \right\}^{1/\hat{k}_i}. \end{aligned}$$

Define

$$\text{FDR}(t, \Pi, K) = E \left[\frac{\sum_{i=1}^m (1 - H_i) \mathbf{1}\{p_i \leq c(t, \pi_i, k_i)\}}{\sum_{i=1}^m \mathbf{1}\{p_i \leq c(t, \pi_i, k_i)\}} \right]$$

with $\Pi = (\pi_1, \dots, \pi_m)$ and $K = (k_1, \dots, k_m)$. We make the following assumptions to facilitate our theoretical derivations.

Assumption 3.1. Suppose the parameter spaces Θ and \mathcal{B} are both compact.

Assumption 3.2. Suppose

$$\lim_m \frac{1}{m} \sum_{i=1}^m E \log \{ \pi_\theta(x_i) + (1 - \pi_\theta(x_i))(1 - k_\beta(x_i)) p_i^{-k_\beta(x_i)} \}$$

converges uniformly over $\theta \in \Theta$ and $\beta \in \mathcal{B}$ to $R(\theta, \beta)$, which has a unique maximum at (θ^*, β^*) in $\Theta \times \mathcal{B}$.

Let $\mathcal{F}_a^b = \sigma((x_i, p_i), a \leq i \leq b)$ be the Borel σ -algebra generated by the random variables (x_i, p_i) for $a \leq i \leq b$. Define the α -mixing and ϕ -mixing coefficients, respectively, as

$$\begin{aligned} \alpha(\nu) &= \sup_b \sup_{A \in \mathcal{F}_{-\infty}^b, B \in \mathcal{F}_{b+\nu}^{+\infty}} |P(AB) - P(A)P(B)|, \\ \phi(\nu) &= \sup_b \sup_{A \in \mathcal{F}_{-\infty}^b, B \in \mathcal{F}_{b+\nu}^{+\infty}, P(B) > 0} |P(A|B) - P(A)|. \end{aligned}$$

Assumption 3.3. Suppose (x_i, p_i) is α -mixing with $\alpha(\nu) = O(\nu^{-\xi})$ for $\xi > r/(r-1)$ and $r > 1$ (or ϕ -mixing with $\phi(\nu) = O(\nu^{-\xi})$ for $\xi > r/(2r-1)$ and $r \geq 1$). Further assume $\sup_i E |\log(p_i)|^{r+\delta} < \infty$ and $\max_i \|x_i\|_\infty < C$, where $\|\cdot\|_\infty$ denotes the l_∞ norm of a vector and $C, \delta > 0$.

Assumption 3.1 is standard. **Assumption 3.2** is a typical condition in the literature of maximum likelihood estimation for misspecified models (see, e.g., White 1982). **Assumption 3.3** relaxes the usual independence assumption by allowing (x_i, p_i) to be weakly dependent. It is needed to establish the uniform strong law of large numbers for the process $R_m(\theta, \beta)$ defined in the proof of **Lemma 3.1** which establishes the uniform strong consistency for $\hat{\pi}_i$ and \hat{k}_i . The boundedness assumption on x_i could be relaxed with a more delicate analysis to control its tail behavior and study the convergence rate of $\hat{\theta}$ and $\hat{\beta}$. Denote by $\|\cdot\|$ the l_2 norm of a vector. An essential condition required in our proof of **Lemma 3.1** is $\|\hat{\theta} - \theta^*\| \max_{1 \leq i \leq n} \|x_i\| = o_{a.s.}(1)$. If $\|\hat{\theta} - \theta^*\| = O_{a.s.}(n^{-a})$ for some $a > 0$, then by the Borel–Cantelli lemma, we require $\max_{1 \leq i \leq n} E \|x_i\|^k < \infty$ for some k with $ak > 2$, that is, x_i should have a sufficiently light polynomial tail. We remark that **Assumption 3.3** can be

replaced by more primitive conditions which allow other weak dependence conditions (see, e.g., Pötscher and Prucha 1989). Let $\pi_i^* = W(1/(1 + e^{-\tilde{x}_i^{\theta^*}}), \epsilon_1, \epsilon_2)$ and $k_i^* = 1/(1 + e^{-\tilde{x}_i^{\beta^*}})$.

Lemma 3.1. Under **Assumptions 3.1–3.3**, we have

$$\max_{1 \leq i \leq m} |\hat{\pi}_i - \pi_i^*| \xrightarrow{a.s.} 0, \quad \max_{1 \leq i \leq m} |\hat{k}_i - k_i^*| \xrightarrow{a.s.} 0.$$

We impose some additional assumptions to study the asymptotic FDR control.

Assumption 3.4. For two sequences $a_i, b_i \in [\epsilon, 1]$ with small enough ϵ and large enough m ,

$$\left| \frac{1}{m} \sum_{i=1}^m \{P(p_i \leq a_i | x_i) - P(p_i \leq b_i | x_i)\} \right| \leq c_0 \max_{1 \leq i \leq m} |a_i - b_i|,$$

where c_0 depends on ϵ but is independent of m, x_i, a_i , and b_i .

Assumption 3.5. Assume that

$$\frac{1}{m} \sum_{i=1}^m P(p_i \leq c(t, \pi_i^*, k_i^*)) \rightarrow G_0(t), \quad (10)$$

$$\frac{1}{m} \sum_{i=1}^m P(1 - p_i < c(t, \pi_i^*, k_i^*)) \rightarrow G_1(t), \quad (11)$$

$$\frac{1}{m} \sum_{H_i=0} P(p_i \leq c(t, \pi_i^*, k_i^*)) \rightarrow \tilde{G}_1(t), \quad (12)$$

for any $t \geq t_0$ with $t_0 > 0$, where $G_0(t)$, $G_1(t)$, and $\tilde{G}_1(t)$ are all continuous functions of t . Note that the probability here is taken with respect to the joint distribution of (p_i, x_i) .

Let $U(t) = G_1(t)/G_0(t)$, where G_1 and G_0 are defined in **Assumption 3.5**.

Assumption 3.6. There exists a $t' > t_0 > 0$ such that $U(t') < \alpha$.

Assumption 3.4 is fulfilled if the conditional density of p_i given x_i is bounded uniformly across i on $[\epsilon, 1]$. This assumption is not very strong as we still allow the density to be unbounded around zero. **Assumptions 3.5** and **3.6** are similar to those in Theorem 4 of Storey, Taylor, and Siegmund (2004). In particular, **Assumption 3.6** ensures the existence of a cutoff to control the FDR at level α .

We are now in position to state the main result of this section which shows that the proposed procedure provides asymptotic FDR control. The proof is deferred to the supplementary materials.

Theorem 3.1. Suppose **Assumptions 3.1–3.6** hold and f_0 satisfies Condition (4). Then we have

$$\limsup_m \text{FDR}(\hat{t}, \hat{\Pi}, \hat{K}) \leq \alpha,$$

where $\hat{\Pi} = (\hat{\pi}_1, \dots, \hat{\pi}_m)$ and $\hat{K} = (\hat{k}_1, \dots, \hat{k}_m)$.

It is worth mentioning that the validity of our method does not rely on the mixture model assumptions (2) and (3). In this sense, our method is misspecification robust as the classical BH procedure does.

3.2. Power Analysis

We study the asymptotic power of the oracle procedure. Suppose the mixture models (2) and (3) hold with $\pi_{0i} = \pi_0(x_i)$ and $f_{1,i}(\cdot) = f_1(\cdot; x_i)$, where $f_1(\cdot; x)$ is a density function for any fixed $x \in \mathcal{X}$. Denote by $F_1(\cdot; x)$ and $\bar{F}_1(\cdot; x)$ the distribution and survival functions associated with $f_1(\cdot; x)$, respectively. Suppose the empirical distribution of x_i 's converges to the probability law \mathcal{P} . Consider the oracle procedure with $\pi_i = \pi_0(x_i)$ and $k_i = k_0(x_i)$. Here $k_0(\cdot)$ minimizes the integrated Kullback–Leibler divergence, that is,

$$k_0 = \operatorname{argmin}_{k \in \mathcal{K}} \int \mathrm{D}_{\mathrm{KL}}(f(\cdot; x) \| g(\cdot; k(x))) \mathcal{P}(dx),$$

$$\mathrm{D}_{\mathrm{KL}}(f(\cdot; x) \| g(\cdot; k(x))) = \int_0^1 f(p; x) \log \frac{f(p; x)}{g(p; k(x))} dp,$$

with $f(p; x) = \pi_0(x)f_0(p) + (1 - \pi_0(x))f_1(p; x)$ and $g(p; k(x)) = \pi_0(x) + (1 - \pi_0(x))(1 - k(x))p^{-k(x)}$, and $\mathcal{K} = \left\{k(x) : \log \left(\frac{k(x)}{1-k(x)} \right) = \beta_0 + \beta_1'x, (\beta_0, \beta_1) \in \mathcal{B}\right\}$. Write $c(t, x) = c(t, \pi_0(x), k_0(x))$. By the law of large numbers, the realized power of the oracle procedure has the approximation

$$\begin{aligned} \text{Power} &= \frac{\sum_{i=1}^m \mathbf{1}\{i : H_i = 1, p \leq c(t, x_i)\}}{\sum_{i=1}^m \mathbf{1}\{i : H_i = 1\}} \\ &\approx \frac{\int (1 - \pi(x))F_1(c(t_{\text{opt}}, x); x) \mathcal{P}(dx)}{\int (1 - \pi(x)) \mathcal{P}(dx)}, \end{aligned}$$

where t_{opt} is the largest positive number such that

$$\frac{\int \{\pi_0(x)F_0(c(t, x)) + (1 - \pi_0(x))\bar{F}_1(1 - c(t, x); x)\} \mathcal{P}(dx)}{\int \{\pi_0(x)F_0(c(t, x)) + (1 - \pi_0(x))F_1(c(t, x); x)\} \mathcal{P}(dx)} \leq \alpha. \quad (13)$$

We remark that when

$$\frac{\int (1 - \pi_0(x))\bar{F}_1(1 - c(t_{\text{opt}}, x); x) \mathcal{P}(dx)}{\int \{\pi_0(x)F_0(c(t_{\text{opt}}, x)) + (1 - \pi_0(x))F_1(c(t_{\text{opt}}, x); x)\} \mathcal{P}(dx)} \approx 0, \quad (14)$$

the asymptotic power of the proposed procedure is closed to the oracle procedure based on the LFDR given by

$$\text{LFDR}_i(p_i) = \frac{\pi_{0i}f_0(p_i)}{\pi_{0i}f_0(p_i) + (1 - \pi_{0i})f_{1,i}(p_i)}. \quad (15)$$

4. Simulation Studies

We conduct comprehensive simulations to evaluate the finite-sample performance of the proposed method and compare it to competing methods. For genome-scale multiple testing, the numbers of hypotheses could range from thousands to millions. For demonstration purpose, we start with $m = 10,000$ hypotheses. To study the impact of signal density and strength, we simulate three levels of signal density (sparse, medium, and dense signals) and six levels of signal strength (from very weak to very strong). To demonstrate the power improvement by using external covariates, we simulate covariates of varying informativeness (non-informative, moderately informative, and

strongly informative). For simplicity, we simulate one covariate $x_i \sim N(0, 1)$ for $i = 1, \dots, m$. Given x_i , we let

$$\pi_{0i} = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad \eta_i = \eta_0 + k_d x_i,$$

where η_0 and k_d determine the baseline signal density and the informativeness of the covariate, respectively. For each simulated dataset, we fix the value of η_0 and k_d . We set $\eta_0 \in \{3.5, 2.5, 1.5\}$, which achieves a signal density around 3%, 8%, and 18%, respectively, at the baseline (i.e., no covariate effect), representing sparse, medium and dense signals. We set $k_d \in \{0, 1, 1.5\}$, representing a noninformative, moderately informative and strongly informative covariate, respectively. Thus, we have a total of $3 \times 3 = 9$ parameter settings. Based on π_{0i} , the underlying truth H_i is simulated from

$$H_i \sim \text{Bernoulli}(1 - \pi_{0i}).$$

Finally, we simulate independent z -scores using

$$z_i \sim N(k_s H_i, 1),$$

where k_s controls the signal strength (effect size) and we use values equally spaced on $[2, 2.8]$. Z -scores are converted into p -values using the one-sided formula $1 - \Phi(z_i)$. p -values together with x_i are used as the input for the proposed method.

In addition to the basic setting (denoted as Setup S0), we investigate other settings to study the robustness of the proposed method. Specifically, we study

Setup S1. Additional f_1 distribution. Instead of simulating normal z -scores under f_1 , we simulate z -scores from a noncentral gamma distribution with the shape parameter $k = 2$. The scale/noncentrality parameters of the noncentral gamma distribution are chosen to match the variance and mean of the normal distribution under S0.

Setup S2. Covariate-dependent π_{0i} and $f_{1,i}$. On top of the basic setup S0, we simulate another covariate $x'_i \sim N(0, 1)$ and let x'_i affect $f_{1,i}$. Specifically, we scale k_s by $\frac{2 \exp(k_f x'_i)}{1 + \exp(k_f x'_i)}$, where we set $k_f \in \{0, 0.25, 0.5\}$ for noninformative, moderately informative and strongly informative covariate scenarios, respectively.

Setup S3. Dependent hypotheses. We further investigate the effect of dependency among hypotheses by simulating correlated multivariate normal z -scores. Four correlation structures, including two block correlation structures and two AR(1) correlation structures, are investigated. For the block correlation structure, we divide the 10,000 hypotheses into 500 equal-sized blocks. Within each block, we simulate equal positive correlations ($\rho = 0.5$) (S3.1). We also further divide the block into 2 by 2 sub-blocks, and simulate negative correlations ($\rho = -0.5$) between the two sub-blocks (S3.2). For AR(1) structure, we investigate both $\rho = 0.75^{|i-j|}$ (S3.3) and $\rho = (-0.75)^{|i-j|}$ (S3.4).

Setup S4. Heavy-tail covariate. In this variant, we generate x_i from the t distribution with 5 degrees of freedom.

Setup S5. *Nontheoretical null distribution.* We simulate both increasing and decreasing f_0 . For an increasing f_0 (S5.1), we generate null z -score $z_i|H_0 \sim N(-0.15, 1)$. For a decreasing f_0 (S5.2), we generate null z -score $z_i|H_0 \sim N(0.15, 1)$.

We present the simulation results for the Setup S0–S2 in the main text and the results for the Setup S3–S5 in the supplementary materials. To allow users to conveniently implement our method and reproduce the numerical results reported here, we make our code and data publicly available at <https://github.com/jchen1981/CAMT>.

4.1. Competing Methods

We label our method as covariate adaptive multiple testing (CAMT) and compare it to the following competing methods:

- Oracle: Oracle procedure based on LFDR (see, e.g., (15)) with simulated π_{0i} and $f_{1,i}$, which theoretically has the optimal performance;
- BH: Benjamini–Hochberg procedure (Benjamini et al. 1995, *p.adjust* in R 3.4.2);
- ST: Storey’s BH procedure (Storey 2002, *qvalue* package, v2.10.0);
- BL: Boca and Leek procedure (Boca and Leek 2018, *swfdr* package, v1.4.0);
- IHW: Independent hypothesis weighting (Ignatiadis et al. 2016, *IHW* package, v1.6.0);
- FDRreg: False discovery rate regression (Scott et al. 2015, *FDRreg* package, v0.2, <https://github.com/jgscott/FDRreg>), FDRreg(T) and FDRreg(E) represent FDRreg with the theoretical null and empirical null, respectively;
- SABHA: Structure adaptive BH procedure (Li and Barber 2019, $\tau = 0.5, \epsilon = 0.1$ and stepwise constant weights, https://www.stat.uchicago.edu/~rina/sabha/All_q_est_functions.R);
- AdaPT: Adaptive p -value thresholding procedure (Lei and Fithian 2018, *adaptMT* package, v1.0.0).

We evaluate the performance based on FDR control (false discovery proportion) and power (true positive rate) with a target FDR level of 5%. Results are averaged over 100 simulation runs.

4.2. Simulation Results

We first study the performance of the proposed method under the basic setup (S0, Figure 2). All compared methods generally controlled the FDR around/under the nominal level of 0.05 and no serious FDR inflation was observed at any of the parameter setting (Figure 2(A)). However, FDRreg exhibited a slight FDR inflation under some parameter settings and the inflation seemed to increase with the informativeness of the covariate and signal density. Conservativeness was also observed for some methods in some cases. As expected, the BH procedure, which did not take into account π_0 , was conservative when the signal was dense. IHW procedure was generally more conservative than BH and the conservativeness increased with the informativeness of the covariate. CAMT, the proposed method, was

conservative when the signal was sparse and the covariate was less informative. The conservativeness was more evident when the effect size was small but decreased as the effect size became larger. AdaPT was more conservative than CAMT under sparse signal/weak covariate. In terms of power (Figure 2(B)), there were several interesting observations. First, as the covariate became more informative, all the covariate adaptive methods became more powerful than ST and BH. The power differences between these methods also increased. Second, FDRreg was the most powerful across settings. Under a highly informative covariate, it was even slightly above the oracle procedure, which theoretically had an optimal power. The superior power of FDRreg could be partly explained by a less well controlled FDR. The IHW was more powerful than BL/SABHA when the signal was sparse; but the trend reversed when the signal was dense. Third, AdaPT was very powerful when the signal was dense and the covariate was highly informative. However, the power decreased as the signal became more sparse and the covariate became less informative. In fact, when the signal was sparse and the covariate was not informative or moderately informative, AdaPT had the lowest power. In contrast, the proposed method CAMT was close to the oracle procedure. It was comparable to AdaPT when AdaPT was the most powerful, but was significantly more powerful than AdaPT in its unfavorable scenarios. CAMT had a clear edge when the covariate was informative and signal was sparse. Similar to AdaPT, CAMT had some power loss under sparse signal and non-informative covariate, probably due to the discretization effect from the BC-type estimator.

We conducted more evaluations on Type I error control under S0. We investigated the FDR control across different target levels. Figure 3 showed excellent FDR control across target levels for all methods except FDRreg. The actual FDR level of BH and IHW was usually below the target level. CAMT was slightly conservative at a small target level under the scenario of sparse signal and less informative covariate, but it became less conservative at larger target levels. We also simulated a complete null, where no signal was included (Figure 4). In such case, FDR was reduced to FWER. Interestingly, FDRreg was as conservative as CAMT and AdaPT under the complete null.

It is interesting to study the performance of the competing methods under a much larger feature size, less signal density, and weaker signal strength, representing the most challenging scenario in real problems. To achieve this end, we simulated $m = 100,000$ features with a signal density of 0.5% at the baseline (no covariate effect). Under a moderately informative covariate, we observed a substantial power improvement of CAMT over all other methods including FDRreg while controlling the FDR adequately at different target levels (Figure 5). We further reduced the feature size to 1000 (Figure A1 in the supplementary materials) to study the robustness of the methods to a much smaller feature size. Although CAMT and AdaPT were still more powerful than the competing methods when the signal was dense and the covariate was informative, a significant power loss was observed in other parameter settings, particularly under sparse signal and a less informative covariate. As we further decreased the feature size to 200, CAMT and AdaPT became universally less powerful than ST across parameter settings

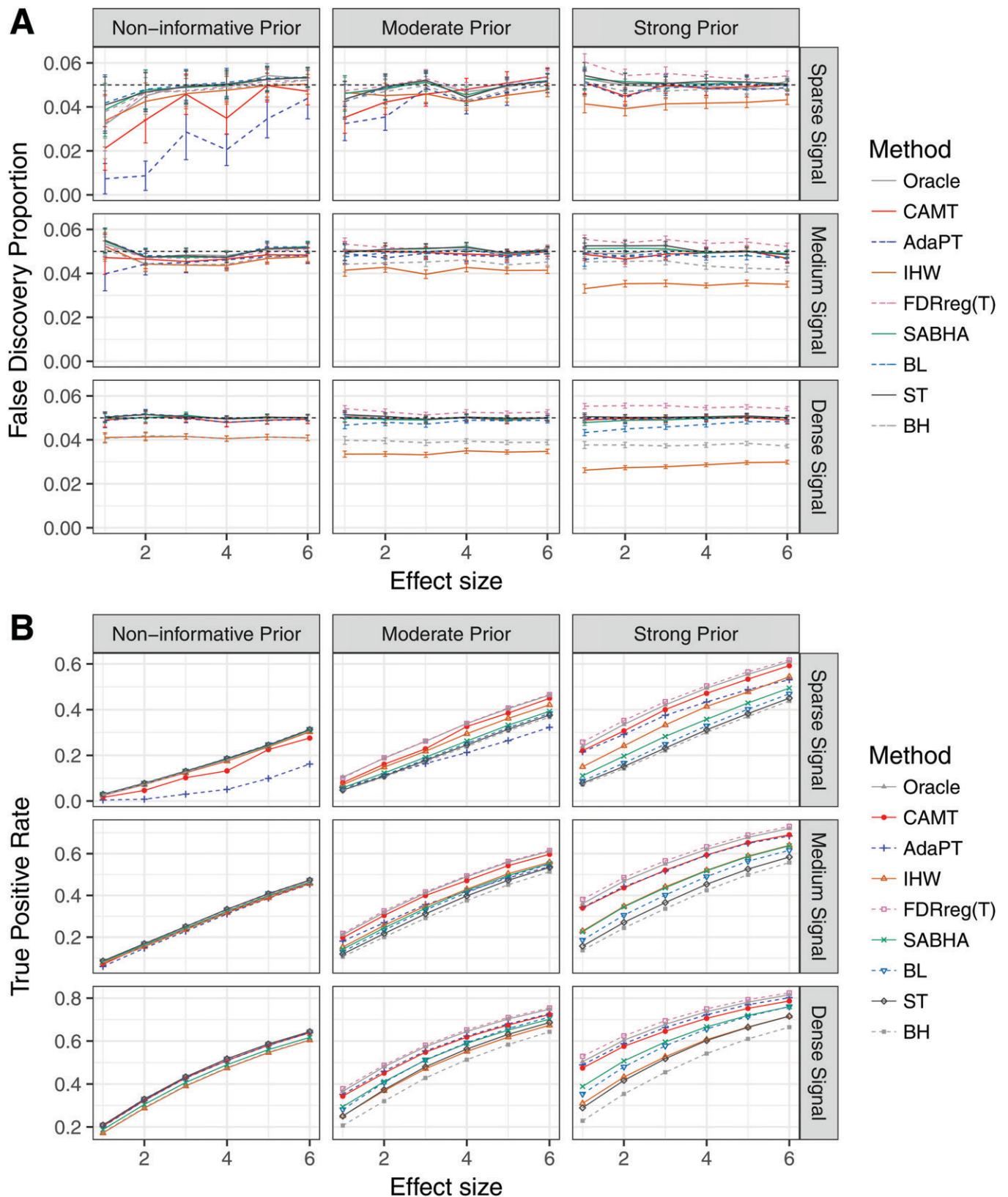


Figure 2. Performance comparison under the basic setting (S0). False discovery proportions (A) and true positive rates (B) were averaged over 100 simulation runs. Error bars (A) represent the 95% CIs and the dashed horizontal line indicates the target FDR level of 0.05.

(data not shown). Therefore, application of CAMT or AdaPT to datasets with small numbers of features was not recommended unless the signal was dense and the covariate was highly informative.

We also simulated datasets, where the z -scores under the alternative were drawn from a noncentral gamma distribution (Setup S1). Under such setting, the trend remained almost the same as the basic setup (Figure 6), but FDRreg had a more

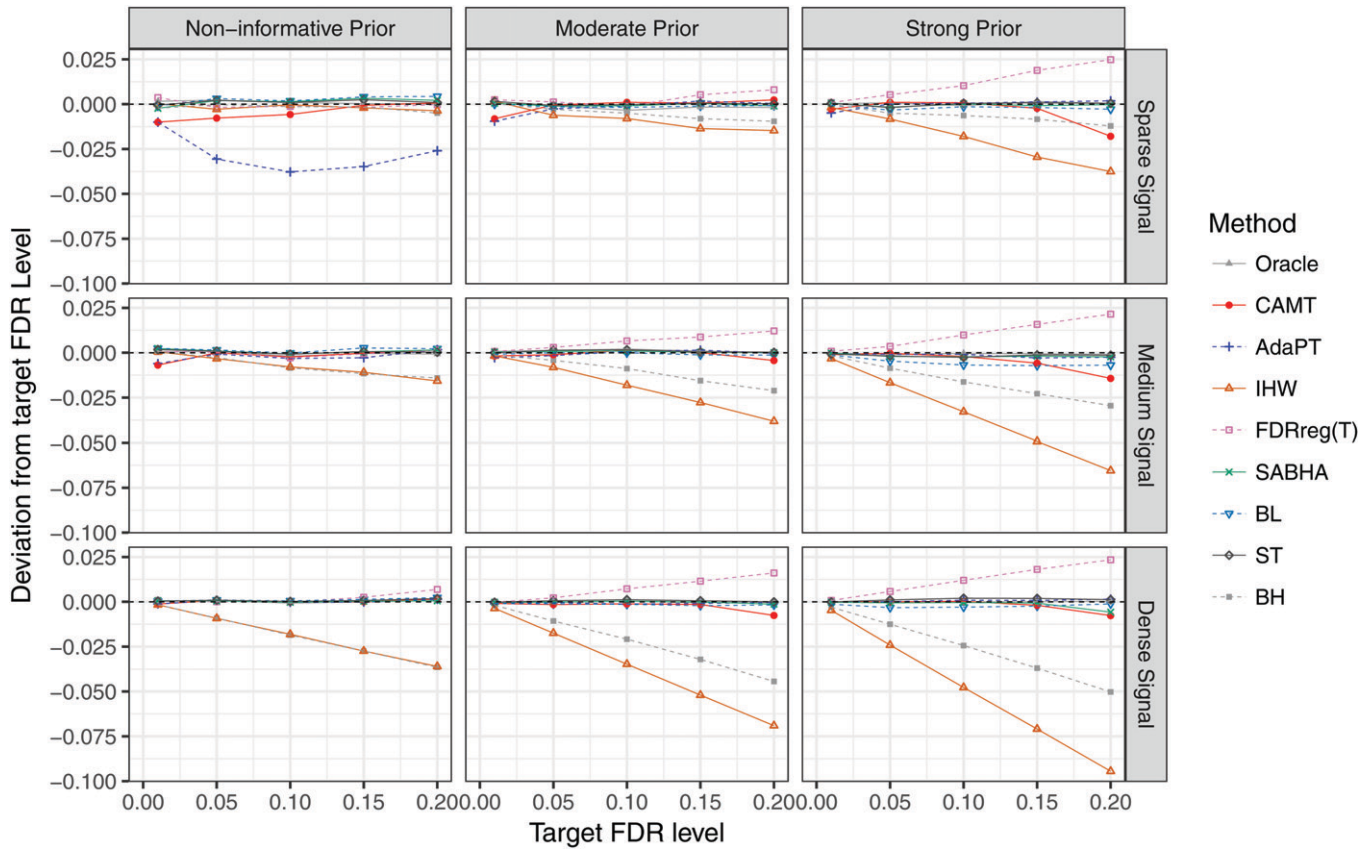


Figure 3. FDR control at various target levels (0.01–0.20) under the basic setting (S0) and a medium signal strength. False discovery proportions were averaged over 100 simulation runs and the deviation from the target level (y-axis) was plotted.

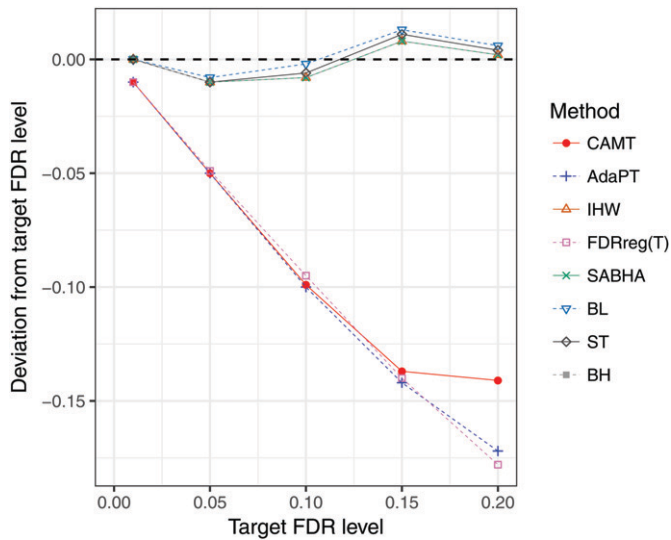


Figure 4. FDR control at various target levels (0.01–0.20) under the complete null (no signal was simulated). False discovery proportions were averaged over 1000 simulation runs and the deviation from the target level (y-axis) was plotted.

marked FDR inflation. When both π_{0i} and $f_{1,i}$ depended on the covariate (Setup S2), CAMT became slightly more powerful without affecting the FDR control, especially when the covariate was highly informative (Figure 7). Meanwhile, the performance of FDRreg was also remarkable with a very small FDR inflation. However, if we increased the effect on $f_{1,i}$ by

reducing the standard deviation of the z-score under the alternative, FDRreg was no longer robust and the observed FDP was substantially above the target level when the signal strength was weak, indicating the benefit of modeling covariate-dependent f_1 (Figure A2 in the supplementary materials). CAMT was also robust to different correlation structures (Setup S3.1, S3.2, S3.3, S3.4) and we observed similar performance under these correlation structures (Figures A3–A6 in the supplementary materials). The performance of CAMT was also robust to a heavy-tail covariate (Setup S4, Figure A7 in the supplementary materials). In an unreported numerical study, we added different levels of perturbation to the covariate by multiplying random small values drawn from $\text{Unif}(0.95, 1.05)$, $\text{Unif}(0.9, 1.1)$, and $\text{Unif}(0.8, 1.2)$, respectively. We observed that the π_0 estimates under perturbation are highly correlated with the π_0 estimates without perturbation, which showed the stability of our method against data perturbations.

We also examined the robustness of CAMT to the deviation from the theoretical null (Setup S5). Specifically, we simulated both decreasing and increasing f_0 . The new results were presented in Figures A8 and A9 in the supplementary materials. We observed that, for an increasing f_0 , all the methods other than FDRreg were conservative and had substantial less power than the oracle procedure. FDRreg using a theoretical null was conservative when the covariate was less informative but was anti-conservative under a highly informative covariate. On the other hand, FDRreg using an empirical null had an improved power and controlled the FDR closer to the target level for most

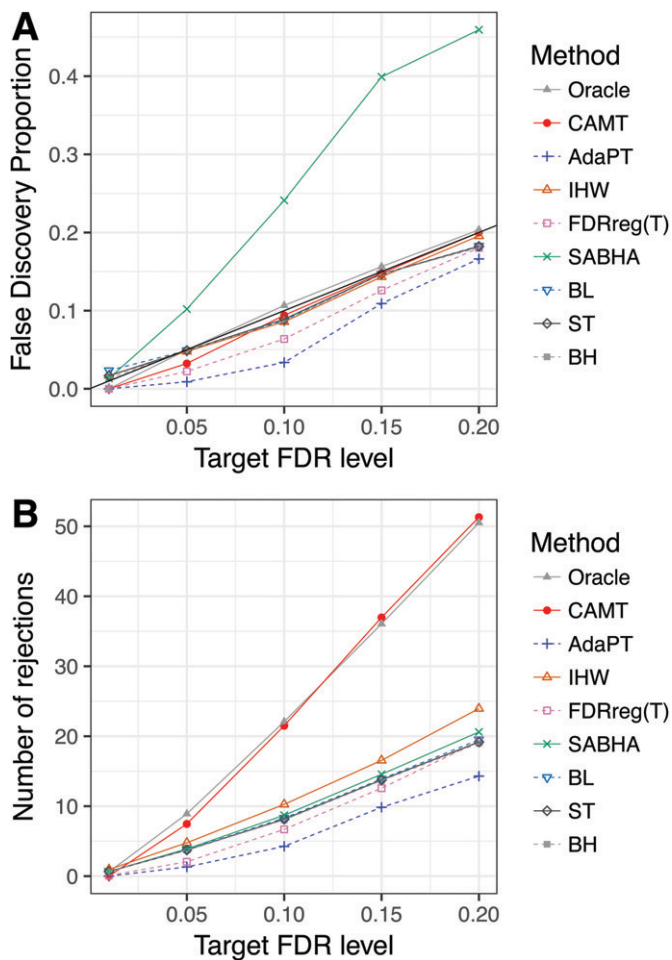


Figure 5. Performance comparison with $m = 100,000$ under the basic setting (S0). Extremely low signal density ($>99\%$), moderate covariate strength and low signal strength were simulated. False discovery proportions (A) and number of rejections (B) were averaged over 100 simulation runs and were plotted against various FDR target levels (0.01–0.20).

settings. However, it did not control the FDR well when the signal was dense and the prior information was strong. When f_0 was decreasing, all the methods without using the empirical null failed to control the FDR. FDRreg with an empirical null improved the FDR control substantially for most settings but still could not control the FDR well under the dense-signal and strong-prior setting. Therefore, there is still room for improvement to address the empirical null problem.

Finally, we compared the computational efficiency of these competing methods (Figure 8). SABHA (step function) and IHW were computationally the most efficient and they completed the analysis for one million p -values in less than two minutes. CAMT and the new version of FDRreg (v0.2) were also computationally efficient, followed by BL, and they all could complete the computation in minutes for one million p -values under S0. AdaPT was computationally the most intensive and completed the analysis in hours for one million p -values. We note that all the methods including AdaPT are computationally feasible for a typical omics dataset.

In summary, CAMT improves over existing covariate adaptive multiple testing procedures, and is a powerful, robust and computationally efficient tool for large-scale multiple testing.

5. Application to Omics-Wide Multiple Testing

To demonstrate the use of the proposed method for real-world applications, we applied CAMT to several omics datasets from transcriptomics, proteomics, epigenomics, and metagenomics studies with the aim to identify omics features associated with the phenotype of interest. Since AdaPT is the most start-of-the-art method, we focused our comparison to it. To make a fair comparison, we first run the analyses on the four omics datasets, which were also evaluated by AdaPT (Lei and Fithian 2018), including Bottomly (Bottomly et al. 2011), Pasilla (Brooks et al. 2011), Airway (Himes et al. 2014), and Yeast Protein dataset (Dephore et al. 2012). The Bottomly, Pasilla, and Airway were three transcriptomics datasets from RNA-seq experiments with a feature size of 13,932, 11,836, and 33,469, respectively. The yeast protein dataset was a proteomics dataset from with a feature size of 2666. We used the same methods to calculate the p -values for these datasets as described in Lei and Fithian (2018). The distributions of the p -values for these four datasets all exhibited a spike in the low p -value region, indicating that the signal was dense. The logarithm of normalized count (averaged across all samples) was used as the univariate covariate for the three RNA-seq data (Bottomly, Pasilla and Airway). The logarithm of the total number of peptides across all samples was used as the univariate covariate for the yeast protein data. Following AdaPT, we used a spline basis with six equiquantile knots for $\pi_{0i,f_{1,i}}$ (CAMT and AdaPT) and for π_{0i} (FDRreg, BL) to account for potential complex nonlinear effects. Since IHW and SABHA could only take univariate covariate, we used the univariate covariate directly. We summarized the results in Figure 9. We were able to reproduce the results in Lei and Fithian (2018). Indeed, AdaPT was more powerful than SABHA, IHW, ST, and BH on the four datasets. FDRreg and BL, which were not compared in Lei and Fithian (2018), also performed well and made more rejections than other methods on the Yeast dataset and the Bottomly dataset, respectively. The performance of the proposed method, CAMT, was almost identical to AdaPT, which was consistent with the simulation results in the scenario of dense signal and informative covariate (Figure 2).

We next applied to two additional omics datasets from an EWAS of congenital heart disease (CHD) (Wijnands et al. 2017) and a microbiome-wide association study (MWAS) of sex effect (McDonald et al. 2018).

- **EWAS data.** The aim of the EWAS of CHD was to identify the CpG loci in the human genome that were differentially methylated between healthy ($n = 196$) and CHD ($n = 84$) children. The methylation levels of 455,741 CpGs were measured by the Illumina 450K methylation beadchip and was normalized properly before analysis. The p -values were produced by running a linear regression to the methylation outcome for each CpG, adjusting for potential confounders such as age, sex, and blood cell mixtures as described in Wijnands et al. (2017). Since widespread hyper-methylation (increased methylation in low-methylation regions) or hypo-methylation (decreased methylation in high-methylation regions) are common in many diseases (Robertson 2005), we use the mean methylation across samples as the univariate covariate.

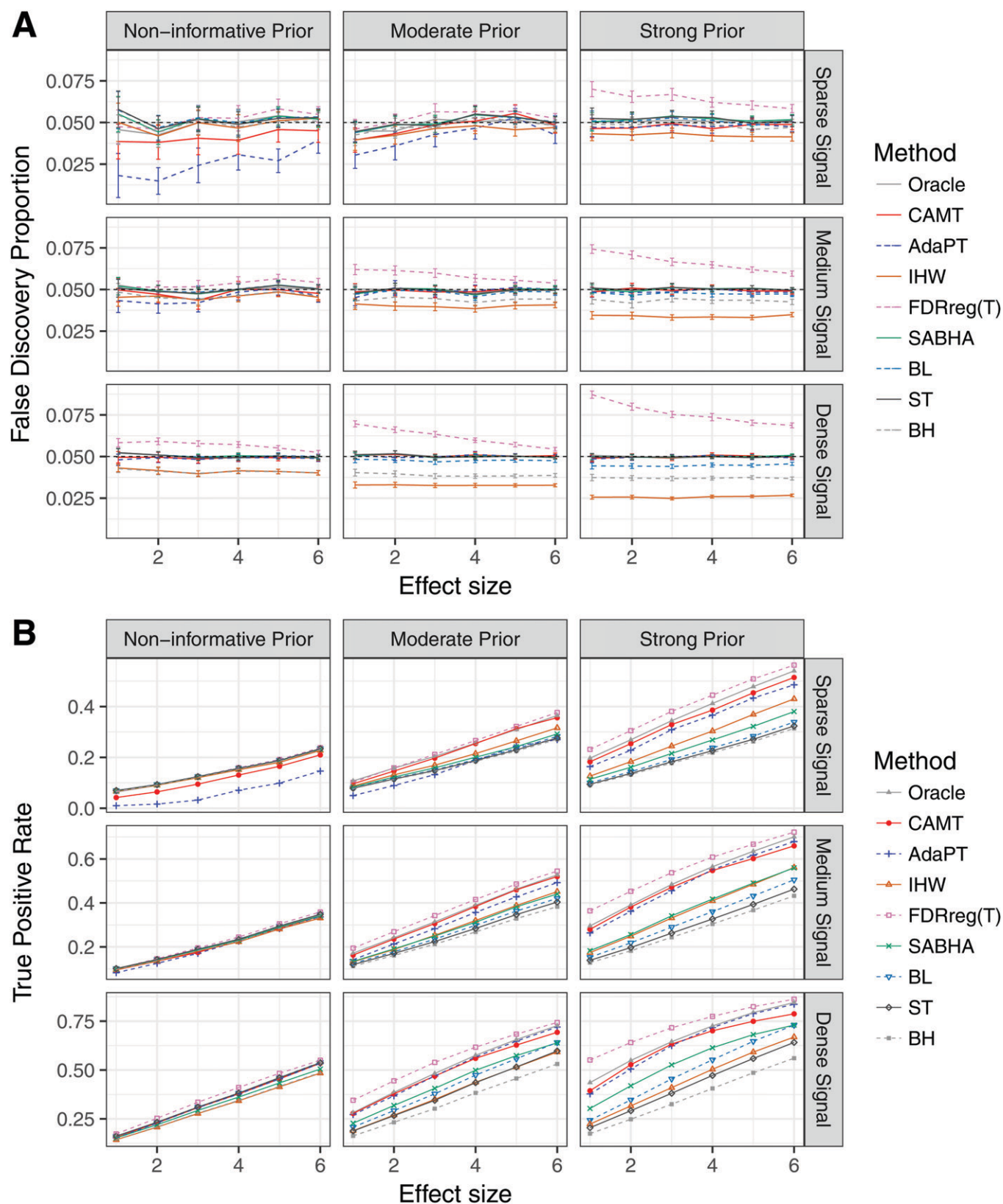


Figure 6. Performance comparison under $S1$ ($f_{1,j}$; noncentral gamma distribution). False discovery proportions (A) and true positive rates (B) were averaged over 100 simulation runs. Error bars (A) represent the 95% CIs and the dashed horizontal line indicates the target FDR level of 0.05.

• *MWAS data.* The aim of the MWAS of sex was to identify differentially abundant bacteria in the gut microbiome between males and females, where the abundances of the gut bacteria were determined by sequencing a fingerprint gene

in the bacteria 16S rRNA gene. We used the publicly available data from the AmericanGut project (McDonald et al. 2018), where more than the gut microbiome from more than 10,000 subjects were sequenced. We focused our analysis on

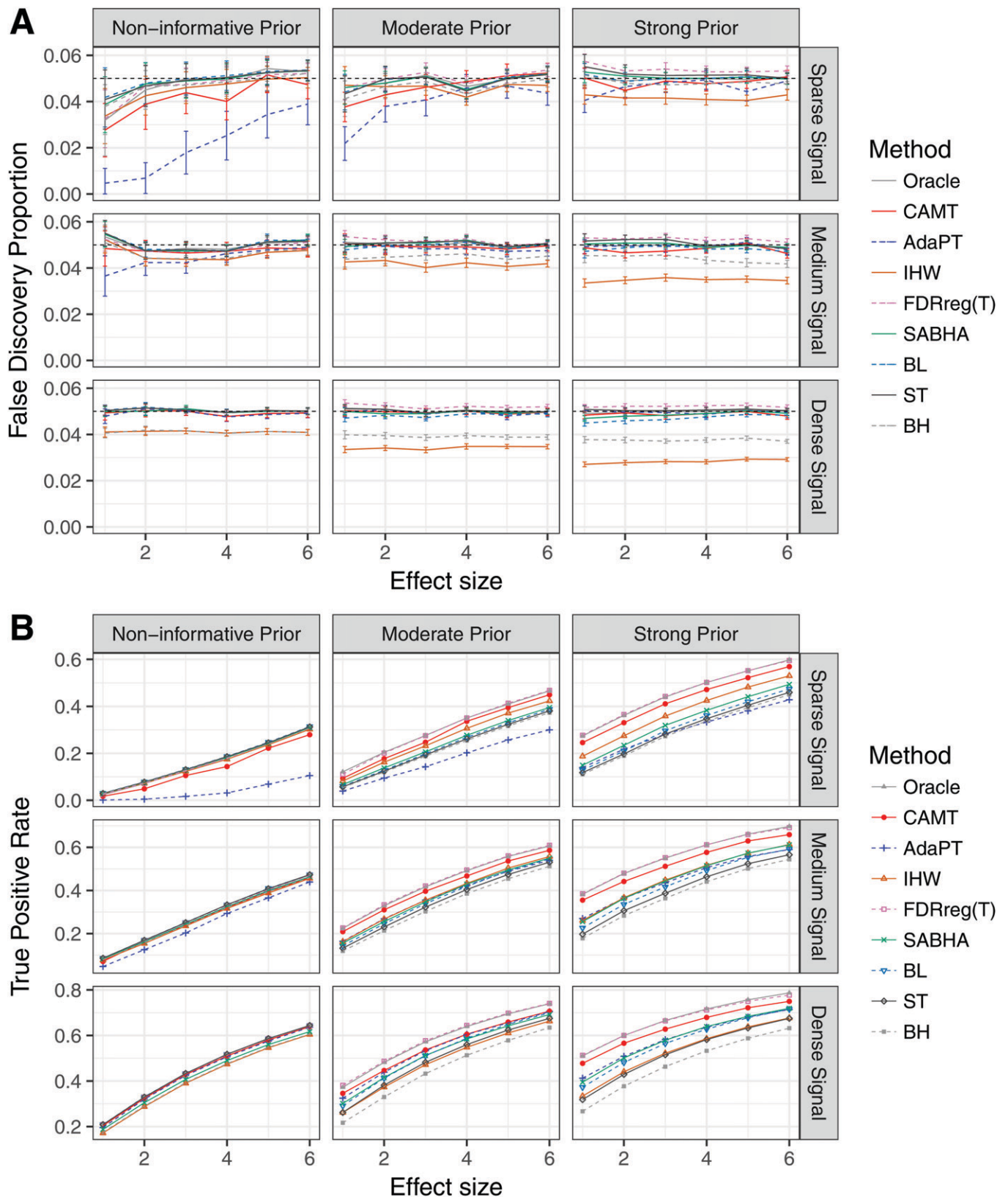


Figure 7. Performance comparison under S2 (covariate-dependent $\pi_{0,i}$ and $f_{1,i}$). False discovery proportions (A) and true positive rates (B) were averaged over 100 simulation runs. Error bars (A) represent the 95% CIs and the dashed horizontal line indicates the target FDR level of 0.05.

a relatively homogenous subset consisting of 481 males and 335 males (age between 13 and 70, normal BMI, from United States). We removed OTUs (clustered sequencing units representing bacteria species) observed in less than 5 subjects,

and a total of 2492 OTUs were tested using Wilcoxon rank sum test on the normalized abundances. We use the percentage of zeros across samples as the univariate covariate since we expect a much lower power for OTUs with excessive zeros.

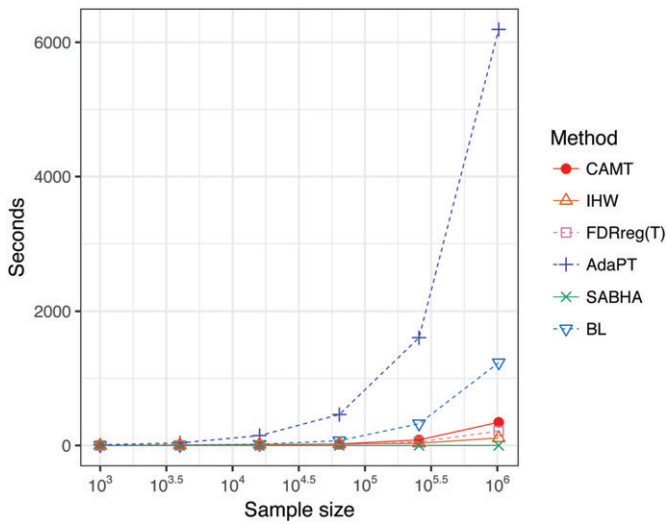


Figure 8. Comparison of runtime under the basic setting (S0). Medium signal density and strength, and a moderately informative covariate was simulated. The number of features varied from 10^3 to $4^5 \times 10^3$. The average runtime over three replications was plotted against the feature size on a log scale. The computation was performed on an AMD Opteron CPU with 256GB RAM and 16 MB available cache.

The results for these two datasets were summarized in [Figure 10](#). For the EWAS data, the signal density was very sparse ($\hat{\pi}_0 = 0.99$, *qvalue* package). CAMT identified far more loci than the other methods at various FDR levels. The performance was consistent with the simulation results in the scenario of extremely sparse signal and informative covariate, where CAMT was substantially more powerful than the competing methods ([Figure 5](#)). At an FDR of 20%, we identified 55 differentially methylated CpGs, compared to 19 for AdaPT. These 55 CpG loci were mainly located in CpG islands and the gene promoter regions, which were known for their important role in gene expression regulation ([Robertson 2005](#)). Interestingly, all but one CpG loci had low levels of methylation, indicating the methylation level was indeed informative to help identify differential CpGs. We also did gene set enrichment analysis for the genes where the identified CpGs were located (<https://david.ncifcrf.gov/>). Based on the GO terms annotated to biological processes (BP_DIRECT), three GO terms were found to be significant (unadjusted *p*-value < 0.05) including one term “embryonic heart tube development,” which was very relevant to the CHD under study ([Wijnands et al. 2017](#)). As a sanity check, we randomized the covariate and reanalyzed the data

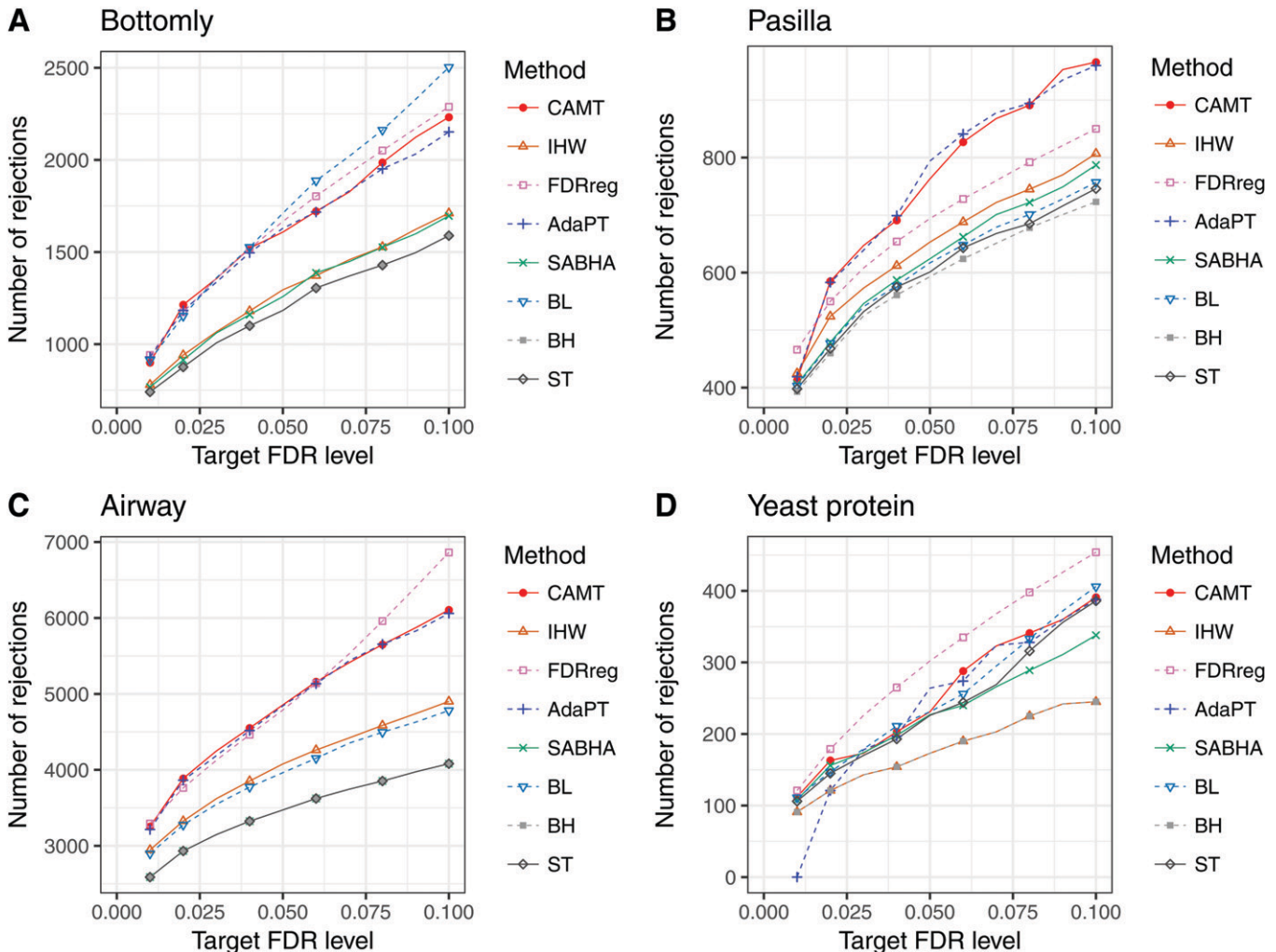


Figure 9. The number of rejections at different target FDR levels on four real datasets used to demonstrate the performance of AdaPT. The Bottomly (A), Pasilla (B), and Airway (C) datasets were three transcriptomics datasets from RAN-seq experiments with a feature size of 13,932, 11,836, and 33,469, respectively. The yeast protein dataset (D) was a proteomics dataset with a feature size of 2666.

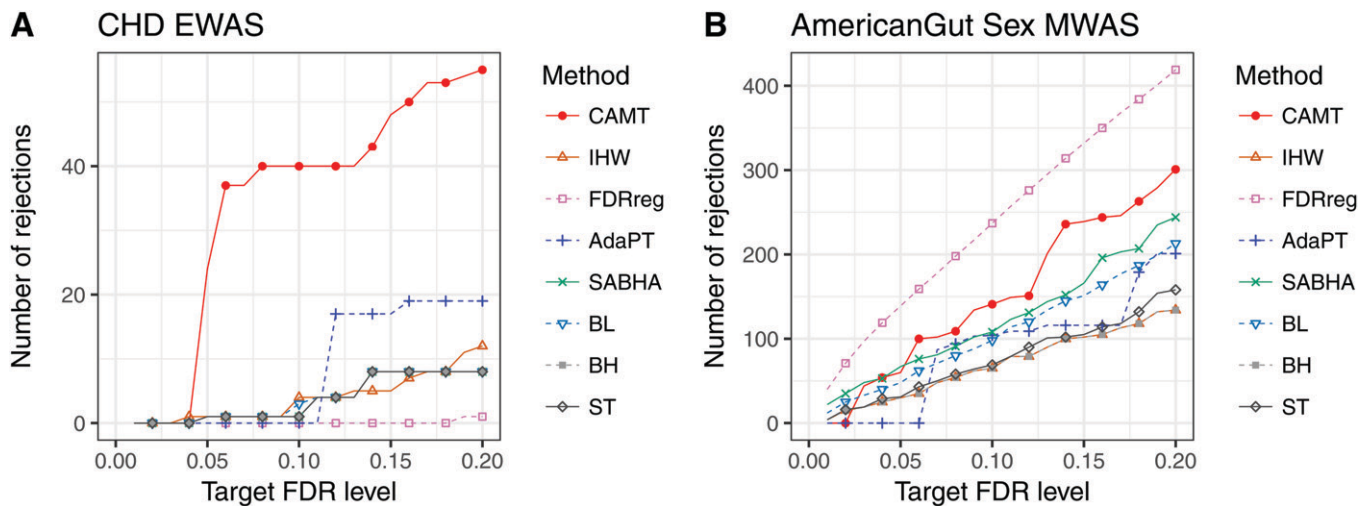


Figure 10. The number of rejections at different target FDR levels on two real datasets: EWAS of congenital heart disease (A) and MWAS of sex effect (B). The EWAS dataset was produced by the Illumina 450K methylation beadchip ($m = 455,741$) and the MWAS dataset was produced by the 16S rRNA gene amplicon sequencing ($m = 2492$).

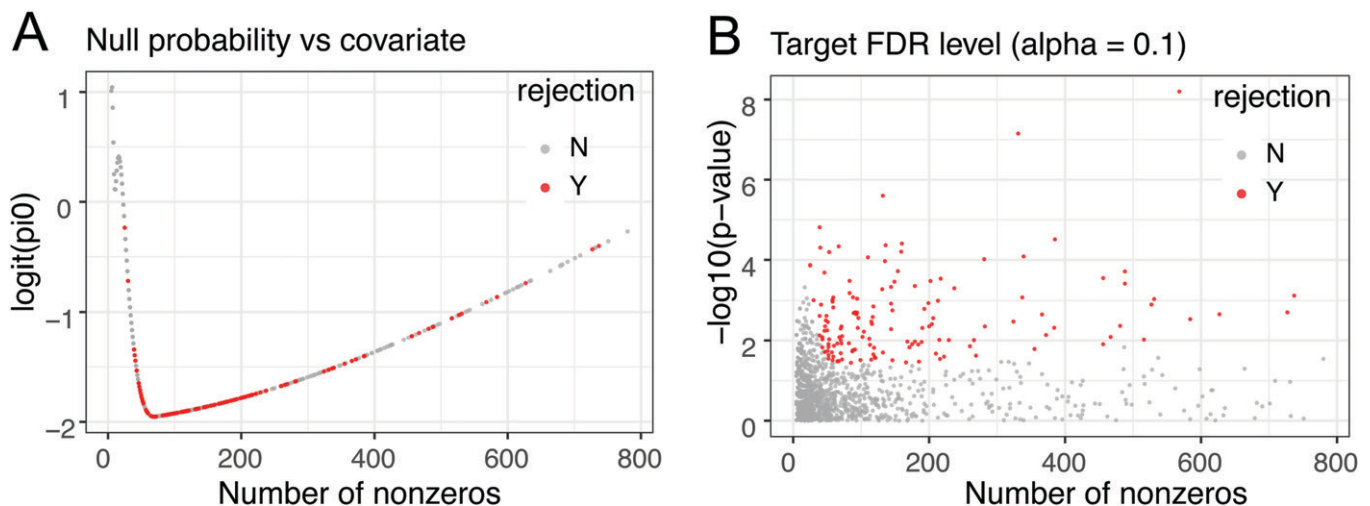


Figure 11. Performance on the MWAS dataset. (A) The fitted π_{0i} (logit scale) versus the covariate (number of nonzeros). (B) p -value (log scale) versus the covariate (number of nonzeros). Rejected hypotheses at FDR 10% were in red.

using CAMT. As expected, CAMT became similar to BH/ST and identified the same eight CpGs at 20% FDR level.

For the MWAS data, although the difference was not as striking as the EWAS data, CAMT was still overall more powerful than other competing methods except FDRreg. However, given the fact that FDRreg was not robust under certain scenarios, the interpretation of the increased power should be cautious. The relationship between the fitted π_{0i} and the covariate (number of nonzeros) was very interesting: $\hat{\pi}_{0i}$ first decreased, reached a minimum at around 70 and then increased (Figure 11). When the OTU was rare (e.g., a small number of nonzeros, only a few subjects had it), it was either very individualized or we had limited power to reject it, leading to a large π_{0i} . In the other extreme where the OTU was very prevalent (e.g., a large number of nonzeros, most of the subjects had it), it was probability not sex-specific either. Therefore, taking into account the sparsity level could increase the power of MWAS. It is also informative to compare CAMT to the traditional filtering-based procedure for MWAS. In practice, we usually apply a prevalence-based filter before performing multiple testing correction, based on the idea

that rare OTUs are less likely to be significant and including them will increase the multiple testing burden. A subjective filtering criterion has to be determined beforehand. For this MWAS dataset, if we removed OTUs present in less than 10% of the subjects, ST and BH recovered 116 and 85 significant OTUs at an FDR of 10%, compared to 69 and 65 on the original dataset, indicating that filtering did improve the statistical power of traditional FDR control procedures. However, if we removed OTUs present in less than 20% of the subjects, the numbers of significant OTUs by ST and BH reduced to 71 and 50, respectively. Therefore, filtering could potentially leave out biologically important OTUs. In contrast, CAMT did not require an explicit filtering criterion, and was much more powerful (141 significant OTUs at 10% FDR) than the filtering-based method.

6. Discussions

There are generally two strategies for estimating the number of false rejections $\sum_{i=1}^m (1 - H_i) \mathbf{1}\{h_i(p_i) \geq w_i(t)\}$ given the form of the rejection rule $h_i(p_i) \geq w_i(t)$. The first approach

(called BH-type estimator) is to replace the number of false rejections by its expectation assuming that p_i follows the uniform distribution on $[0, 1]$ under the null, which leads to the quantity $\sum_{i=1}^m \pi_{0i} c(t, \pi_{0i}, k_i)$ for $c(\cdot)$ defined in Section 3. The second approach (called BC-type estimator) estimates the false rejection conservatively by $\xi + \sum_{i=1}^m \mathbf{1}\{h_i(1 - p_i) \geq w_i(t)\}$ for a nonnegative constant ξ under the assumption that the null distribution of p -values is symmetric about 0.5. Both procedures enjoy optimality in some asymptotic sense (see, e.g., Arias-Castro and Chen 2017). The advantage of the BC-type procedure lies on that its estimation of the number of false rejections is asymptotically conservative when the rejection rule converges to a nonrandom limit (which holds even under a misspecified model, see, e.g., White 1982) and f_0 is mirror conservative (see Lei and Fithian 2018, eq. (3)). This fact allows us to estimate the rejection rule by maximizing a potentially misspecified likelihood as the resulting rejection rule has a nonrandom limit under suitable conditions. This is not necessarily the case for the BH-type estimator without imposing additional constraint when estimating π_{0i} and k_i . Specific restriction on the estimators of π_{0i} is required for the BH-type estimator to achieve FDR control (see, e.g., Li and Barber 2019, eq. (3)).

On the other hand, as the BC-type estimator uses a counting approach to estimate the number of false rejections, it suffers from the discretization issue (i.e., the BC-type estimator is a step function of t while the BH-type estimator is continuous), which may result in a large variance for the FDR estimate. This is especially the case when the FDR level is small. For small FDR level, the number of rejections is usually small, and thus both the denominator and numerator of the FDR estimate become small and more variable. Another issue with the BC-type estimator is the selection of ξ . We follow the idea of knockoff+ in Barber and Candès (2015) by setting $\xi = 1$. This choice could make the procedure rather conservative when the signal is very sparse, and the target FDR level is small. A choice of smaller ξ (e.g., $\xi = 0$) often leads to inflated FDR in our unreported simulation studies. To alleviate this issue, one may consider a mixed strategy by using

$$\max \left\{ \sum_{i=1}^m \pi_{0i} c(t, \pi_{0i}, k_i), \sum_{i=1}^m \mathbf{1}\{h_i(1 - p_i) \geq w_i(t)\} \right\}$$

as a conservative estimate for the number of false rejections when t is relatively small. Our numerical results in Figure A10 in the supplementary materials show that the resulting method can successfully reduce the power loss in the case of sparse signals (or small FDR levels) and less informative covariates while maintaining the good power performance in other cases. A serious investigation of this mixed procedure and the BH-type estimator is left for future research.

Since our method is not robust to a decreasing f_0 , some diagnostics are needed before running CAMT. To detect a decreasing f_0 , the genomic inflation factor (GIF) can be employed (Devlin and Roeder 1999). GIF is defined as the ratio of the median of the observed test statistic to the expected median based on the theoretical null distribution. GIF has been widely used in GWAS to assess the deviation of the empirical distribution of the null p -values from the theoretical uniform distribution. To accommodate potential dense signals for some

genomics studies, we recommend to confine the GIF calculation to p -values between 0.5 and 1. If the GIF is larger, using CAMT may result in excess false positives. In such case, the user should not trust the results and may consider recalculating the p -values by adjusting potential confounding factors, either known or estimated based on some latent variable approach such as surrogate variable analysis (Leek and Storey 2007), or using the simple genomic control approach based on p -values (Devlin and Roeder 1999).

Supplementary Materials

The supplementary materials provide proofs of the results in Section 3 and additional numerical results.

Funding

Zhang acknowledges partial support from NSF DMS-1830392 and NSF DMS-1811747. Chen acknowledges support from Mayo Clinic Center for Individualized Medicine.

References

- Arias-Castro, E., and Chen, S. (2017), "Distribution-Free Multiple Testing," *Electronic Journal of Statistics*, 11, 1983–2001. [426]
- Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *Annals of Statistics*, 43, 2055–2085. [414,426]
- 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. [411,418]
- Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011), "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays," *PLoS One*, 6, e17820. [421]
- Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. (2011), "Conservation of an RNA Regulatory Map Between *Drosophila* and Mammals," *Genome Research*, 21, 193–202. [421]
- Boca, S. M., and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 6, e6035. [412,418]
- Cai, T. T., and Sun, W. (2009), "Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks," *Journal of the American Statistical Association*, 104, 1467–1481. [412,413]
- Dephoure, N., and Gygi, S. P. (2012), "Hyperplexing: A Method for Higher-Order Multiplexed Quantitative Proteomics Provides a Map of the Dynamic Response to Rapamycin in Yeast," *Science Signaling*, 5, rs2. [421]
- Devlin, B., and Roeder, K. (1999), "Genomic Control for Association Studies," *Biometrics*, 55, 997–1004. [426]
- Dobriban, E. (2017), "Weighted Mining of Massive Collections of p -values by Convex Optimization," *Information and Inference: A Journal of the IMA*, 7, 251–275. [412]
- Efron, B. (2004), "Local False Discovery Rate," Technical Report, Department of Statistics, Stanford University. [413]
- Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008), "Unsupervised Empirical Bayesian Testing with External Covariates," *Annals of Applied Statistics*, 2, 714–735. [412]
- Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False Discovery Control With p -Value Weighting," *Biometrika*, 93, 509–524. [412]
- Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderma, B., Whitaker, R. M., Duan, Q., Lasky-Su, J., Nikolos, C., and Jester, W. (2014), "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene That Modulates Cytokine Function in Airway Smooth Muscle Cells," *PLoS One*, 9, e99625. [421]
- Hu, J. X., Zhao, H., and Zhou, H. H. (2010), "False Discovery Rate Control With Groups," *Journal of the American Statistical Association*, 105, 1215–1227. [412]

- Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016), "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing," *Nature Methods*, 13, 577–580. [411,412,418]
- Leek, J. T., and Storey, J. D. (2007), "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis," *PLoS Genetics*, 3, e161. [426]
- Lei, L., and Fithian, W. (2018), "AdaPT: An Interactive Procedure for Multiple Testing With Side Information," *Journal of the Royal Statistical Society, Series B*, 80, 649–679. [412,413,415,418,421,426]
- Li, A., and Barber, R. F. (2019), "Multiple Testing with the Structure Adaptive Benjamini-Hochberg Algorithm," *Journal of the Royal Statistical Society, Series B*, 81, 45–74. [412,418,426]
- McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., and Goldasich, L. D. (2018), "American Gut: An Open Platform for Citizen Science Microbiome Research," *mSystems*, 3, e00031-18. [421,422]
- Pötscher, B. M., and Prucha, I. R. (1989), "A Uniform Law of Large Numbers for Dependent and Heterogeneous Data Processes," *Econometrica*, 57, 675–683. [416]
- Robertson, K. D. (2005), "DNA Methylation and Human Disease," *Nature Reviews Genetics*, 6, 597. [421,424]
- Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015), "False Discovery Rate Regression: An Application to Neural Synchrony Detection in Primary Visual Cortex," *Journal of the American Statistical Association*, 110, 459–471. [412,418]
- Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498. [412,418]
- Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society, Series B*, 66, 187–205. [416]
- Sun, W., and Cai, T. T. (2007), "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal of the American Statistical Association*, 102, 901–912. [413]
- Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015), "False Discovery Control in Large-Scale Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 77, 59–83. [412]
- Tansey, W., Koyejo, O., Poldrack, R. A., and Scott, J. G. (2017), "False Discovery Rate Smoothing," arXiv no. 1411.6144. [412]
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [416,426]
- Wijnands, K. P., Chen, J., Liang, L., Verbiest, M. M., Lin, X., Helbing, W. A., Gittenberger-de Groot, A. C., Van Der Spek, P. J., Uitterlinden, A. G., and Steegers-Theunissen, R. P. (2017), "Genome-Wide Methylation Analysis Identifies Novel CpG Loci for Perimembranous Ventricular Septal Defects in Human," *Epigenomics*, 9, 241–251. [421,424]
- Zablocki, R. W., Levine, R. A., Schork, A. J., Xu, S., Wang, Y., Fan, C. C., and Thompson, W. K. (2017), "Semiparametric Covariate-Modulated Local False Discovery Rate for Genome-Wide Association Studies," *The Annals of Applied Statistics*, 11(4), 2252–2269. [412]
- Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M., and Thompson, W. K. (2014), "Covariate-Modulated Local False Discovery Rate for Genome-Wide Association Studies," *Bioinformatics*, 30(15), 2098–2104. [412]