AdaFDR: a Fast, Powerful and Covariate-Adaptive Approach to Multiple Hypothesis Testing

Martin J. Zhang[1], **Fei Xia**[1], **and James Zou**[1,2,3*]

[1]Department of Electrical Engineering, Stanford University, Palo Alto, 94304 USA
[2]Department of Biomedical Data Science, Stanford University, Palo Alto, 94304 USA
[3]Chan-Zuckerberg Biohub, San Francisco, 94158 USA
[*]Corresponding author

ABSTRACT

Multiple hypothesis testing is an essential component of modern data science. Its goal is to maximize the number of discoveries while controlling the fraction of false discoveries. In many settings, in addition to the p-value, additional information/covariates for each hypothesis are available. For example, in eQTL studies, each hypothesis tests the correlation between a variant and the expression of a gene. We also have additional covariates such as the location, conservation and chromatin status of the variant, which could inform how likely the association is to be due to noise. However, popular multiple hypothesis testing approaches, such as Benjamini-Hochberg procedure (BH) and independent hypothesis weighting (IHW), either ignore these covariates or assume the covariate to be univariate. We introduce `AdaFDR`, a fast and flexible method that adaptively learns the optimal p-value threshold from covariates to significantly improve detection power. On eQTL analysis of the GTEx data, `AdaFDR` discovers 32% and 27% more associations than BH and IHW, respectively, at the same false discovery rate. We prove that `AdaFDR` controls false discovery proportion, and show that it makes substantially more discoveries while controlling FDR in extensive experiments. `AdaFDR` is computationally efficient and can process more than 100 million hypotheses within an hour and allows multi-dimensional covariates with both numeric and categorical values. It also provides exploratory plots for the user to interpret how each covariate affects the significance of hypotheses, making it broadly useful across many applications.

Introduction

Multiple hypothesis testing is an essential component in many modern data analysis workflows. A very common objective is to maximize the number of discoveries while controlling the fraction of false discoveries. For example, we may want to identify as many genes as possible that are differentially expressed between two populations such that less than, say, 10% of these identified genes are false positives.

In the standard setting, the data for each hypothesis is summarized by a p-value, with a smaller value presenting stronger evidence against the null hypothesis that there is no association. Commonly-used procedures such as Benjamini-Hochberg (BH)[1] works solely with this list of p-values[2–6]. Despite being widely used, these multiple testing procedures fail to utilize additional information that is often available in modern applications that are not directly captured by the p-value.

For example, in expression quantitative trait loci (eQTL) mapping or genome-wide association studies (GWAS), single nucleotide polymorphism (SNP) in active chromatin state are more likely to be significantly associated with the phenotype[7]. Such chromatin information is readily available in public databases[8], but is not used by standard multiple hypothesis testing procedures—it is sometimes used for post-hoc biological interpretation. Similarly, the location of the SNP, its conservation score, etc., can alter the likelihood for the SNP to be an eQTL. Together such additional information, called covariates, forms a feature representation of the hypothesis; this feature vector is ignored by the standard multiple hypothesis testing procedures.

In this paper, we present `AdaFDR`, a fast and flexible method that adaptively learns the decision threshold from covariates to significantly improve the detection power while having the false discovery proportion (FDP) controlled at a user-specified level. A schematic diagram for `AdaFDR` is shown in Figure 1. `AdaFDR` takes as input a list of hypotheses, each with a p-value and a covariate vector. Conventional methods like BH use only p-values and have the same p-value threshold for all hypotheses (Figure 1 top right). However, as illustrated in the bottom-left panel, the data may have an enrichment of small p-values for certain values of the covariate, which suggests an enrichment of alternative hypotheses around these covariate values. Intuitively, allocating more FDR budget to hypothesis with such covariates could increase the detection power. `AdaFDR` adaptively learns such pattern using both p-values and covariates, resulting in a covariate-dependent threshold that makes more discoveries under the same FDP constraint (Figure 1 bottom right).

Overview. `AdaFDR` extends conventional procedures like BH and Storey-BH (SBH)[2,3] by considering multiple hypothesis testing with side information on the hypotheses. The input of `AdaFDR` is a set of hypotheses each with a p-value and a vector

of covariates, whereas the output is a set of selected (also called rejected) hypotheses. For eQTL analysis, each hypothesis is one pair of SNP and gene, and the p-value tests for association between their values across samples. The covariate can be the location, conservation, and chromatin status at the SNP and the gene. The standard assumption of `AdaFDR` and all the related methods is that the covariates should not affect the p-values under the null hypothesis (see the Methods section for more discussion of this). `AdaFDR` learns the covariate-dependent p-value selection threshold by first fitting a mixture model using expectation maximization (EM) algorithm, where the mixture model is a combination of a generalized linear model (GLM) and Gaussian mixtures[9–11]. Then it makes local adjustments in the p-value threshold by optimizing for more discoveries. We prove that `AdaFDR` controls FDP under standard statistical assumptions in Theorem 1. `AdaFDR` is designed to be fast and flexible — it can simultaneously process more than 100 million hypotheses within an hour and allows multi-dimensional covariates with both numeric and categorical values. In addition, `AdaFDR` provides exploratory plots visualizing how each covariate is related to the significance of hypotheses, allowing users to interpret its findings. We also provide a much faster but slightly less powerful version, `AdaFDR-fast`, which uses only the EM step and skips the subsequent optimization. It can process more than 100 million hypotheses in around 5 minutes on a standard laptop.

We systematically evaluate the performance of `AdaFDR` across multiple datasets. We first consider the problem of eQTL discovery using the data from the Genotype-Tissue Expression (GTEx) project[7]. As covariates, we consider the distance between the SNP and the gene, the gene expression level, the alternative allele frequency as well as the chromatin states of the SNP. Across all 17 tissues considered in the study, `AdaFDR` has an improvement of 32% over BH and 27% over the state-of-art covariate-adaptive method independent hypothesis weighting (IHW)[12,13]. We next consider other applications, including three RNA-Seq datasets[14–16] with the gene expression level as the covariate, two microbiome datasets[17,18] with ubiquity (proportion of samples where the feature is detected) and the mean nonzero abundance as covariates, a proteomics dataset[12,19] with the peptides level as the covariate, and two fMRI datasets[20,21] with the Brodmann area label[22] as the covariate that represents different functional regions of human brain. In all experiments, `AdaFDR` shows a similar improvement. Finally, we perform extensive simulations, including ones from a very recent benchmark paper (Oct 31st 2018)[18], to demonstrate that `AdaFDR` has the highest detection power while controlling the false discovery proportion (FDP) in various cases where the p-values may be either independent or dependent. The default parameters of `AdaFDR` are used for every experiment in this paper, both real data analysis and simulations, without any tuning. In addition to the experiments, we theoretically prove that `AdaFDR` controls FDP with high probability when the null p-values, conditional on the covariates, are independently distributed and stochastically greater than the uniform distribution, a standard assumption also made by related literature[13,23,24].

Related works. The problem of multiple hypothesis testing with covariates has recently been actively explored[12,13,24–28]. These works assume that for each hypothesis, we observe not only a p-value P_i but also a general covariate $\mathbf{x}_i$ which is meant to capture the information on the significance of the hypothesis. However, the nature of this relationship is not known ahead of time and must be learned from the data. IHW[12,13] groups the hypotheses into a pre-specified number of bins and applies a constant threshold for each bin to maximize the discoveries. It is practical, well-received by the community, and can scale up to 1 billion hypotheses. Yet it only supports the covariate to be univariate and uses a stepwise-constant function for the threshold, which limits its detection power. AdaPT[24] cleverly uses a p-value masking procedure to control FDR. While IHW and `AdaFDR` need to split the hypotheses into multiple folds for FDR control, AdaPT can learn the threshold using virtually the entire data. However, such p-value masking procedure takes many iterations of optimization, and can be computationally expensive. Hence, while having high detection power, AdaPT usually takes a long time to run. `AdaFDR` is designed to achieve the best of both worlds: it has a speed comparable to IHW while using a flexible modeling strategy to have greater detection power than AdaPT.

There are also other methods in the field tailored for specific applications, where the domain knowledge can be used to increase the detection power. For example, gene set enrichment analysis (GSEA)[29] uses the gene pathway information to identify classes of genes that are over-represented in a given set of genes. A recent work integrates genomic annotations into a Bayes hierarchical model to increase detection power in eQTL study[30]. Another incorporates phylogenetic tree information into a Bayesian model to increase the detection power in microbiome-wide multiple testing[31]. Compared to these methods, `AdaFDR` does not assume any prior knowledge about the covariates and learns the decision threshold in a completely data-driven manner. Hence, it is a more general approach that has a wider range of applications. Some other related works include non-adaptive p-value weighting[32–34,34,35], estimation of the covariate-dependent null proportion[36–38], and estimation of the local false discovery rate[39–43].

`AdaFDR` is the mature development of and subsumes a previous, preliminary method that we called `NeuralFDR`[27]. Instead of using a neural network to model the discovery threshold as in `NeuralFDR`, `AdaFDR` uses a mixture model that lacks some flexibility but is much faster to optimize — for the GTEx data used in the `NeuralFDR` paper, it takes `NeuralFDR` 10+ hours to process but only 9 minutes for `AdaFDR`. Yet, `AdaFDR` maintains similar discovery power on the benchmark data used to test `NeuralFDR` (Supplementary Figure 3b). We systematically evaluated `AdaFDR` on many more settings and experiments than what was done for `NeuralFDR`.

Results

Discovering eQTLs in GTEx

We first consider detecting eQTLs using data from GTEx[7]. The GTEx project has collected both genetic variation data (SNPs) and gene expression data (RNA-Seq) from 44 human tissues, with sample sizes ranging from 70 (uterus) to 361 (muscle skeletal). Its goal is to study the associations between genotype and gene expression across humans. Each hypothesis test is to test if there is a significant association between a SNP and a gene, also referred to as an eQTL. A standard caveat is that a selected eQTL (either through small p-value or a FDR procedure) may not be a true *causal* SNP — it could tag a nearby causal SNP due to linkage disequilibrium. We should interpret the selected eQTLs with care; nonetheless, it is still valuable to discover candidate associations and local regions with strong associations while controlling FDR[12].

We focus on cis-eQTLs where the SNP and the gene are close to each other on the genome ($<$ 1 million base pairs). Previous works provide evidence that various covariates could be associated with the significance of cis-eQTLs[7,30,44,45]. In this study, we consider four covariates for each SNP-gene pair: 1) the distance from SNP to gene transcription start site (TSS); 2) the log10 gene expression level; 3) the alternative allele frequency (AAF) of the SNP; 4) the chromatin state of the SNP. Out of 44 tissues, we selected 17 whose chromatin state information is available[8] and have more than 100 samples. For each tissue, p-values for all associations are tested simultaneously with numbers of hypotheses ranging from 140 to 180 million for different tissues, imposing a very-large-scale multiple hypothesis testing problem. We use a nominal FDR level of 0.01. Such experiments of testing all SNP-gene pairs simultaneously are also performed in[12,13]. An alternative analysis workflow is to first discover significant genes (eGenes) and then match significant SNPs (eVariants) for each eGene[30].

As shown in Figure 2a, `AdaFDR` and its fast version consistently make more discoveries than other methods in every tissue. On average, it has an improvement of 32% over BH and 27% over IHW. Next we investigate whether using the eQTL p-values of an existing tissue could boost the power of discovering eQTLs in a new tissue. To simulate this scenario, we consider specifically the two adipose tissues, Adipose_Subcutaneous and Adipose_Visceral_Omentum. For each of them, we use the -log10 p-values from the other tissue as an additional covariate— e.g. for Adipose_Subcutaneous, the -log10 p-value of Adipose_Visceral_Omentum is used as an extra covariate. Leveraging previous eQTL results substantially increases discovery power (Figure 2b); the p-value augmentation (AdaFDR (aug)) yields 56% and 83% more discoveries for the two adipose tissues compared to BH. We then perform a control experiment, where the augmented p-values, instead of coming from the other adipose tissue that is similar to the one under investigation, are from a brain tissue (Brain_Caudate_basal_ganglia) that is very different from the adipose tissue (Figure 2 in the GTEx paper[7]). In this case, the improvement in the number of discoveries due to the extra covariate vanishes for the two tissues (AdaFDR (ctrl), which is consistent with the idea that `AdaFDR` learns to leverage shared genetic architecture in closely related tissues to improve power. This analysis suggests that we can potentially greatly improve eQTL discovery by leveraging related tissues during multiple hypothesis testing. We provide additional supporting experiments for the two colon tissues in Supplementary Figure 1a.

`AdaFDR` also characterizes how each covariate affect the significance level of the hypotheses. The results for Adipose_Subcutaneous are shown in Figure 2c as an example. We first consider the distance from TSS and the top-left panel provides a simple visualization, where for each hypothesis (downsampled to 10k), the p-values are plotted against the distances from TSS. There is a strong enrichment of small p-values when the distance is close to 0, indicating that the SNP and gene are more likely to have a significant association if they are close to each other. In the top-center panel, `AdaFDR` characterizes such relationship by providing estimates of the null hypothesis distribution (blue) and the alternative hypothesis distribution (orange), with respect to the distance from TSS. It learns that an alternative hypothesis is more likely to appear at the center, where the distance from TSS is small, consistent with previous works[7,44].

`AdaFDR` interprets other covariates in a similar fashion. Figure 2c top-right panel indicates that genes with higher expression levels are more likely to have significant associations, in agreement with previous observations[12,24]. SNPs with AAF close to 0.5 are also more likely to have significant associations. In addition, the bottom-center panel indicates that SNPs with active chromatin states—Tx (strong transcription), TxWk (weak transcription), TssA (active TSS)—are more likely to have significant associations as compared to SNPs with inactive states—Quies (quiescent), ReprPC (repressed PolyComb) ReprPCWk (weak repressed PolyComb). Finally, the bottom-right panel shows that p-values from the augmented tissue Adipose_Visceral_Omentum are positively correlated with the significance of the associations. See Supplementary Figure 1b for analogous results on the Colon_Sigmoid tissue.

We use adipose eQTL data from the Multiple Tissue Human Expression Resource (MuTHER) project[46] to validate our GTEx eQTL discoveries. The participants in MuTHER are disjoint from the GTEx participants, making MuTHER an independent dataset. For this analysis, we compare the testing results of `AdaFDR` on Adipose_Subcutaneous with that of Storey-BH (SBH), which is known to be a better baseline than BH. As shown in the top panel of Figure 2d, `AdaFDR` detects almost all discoveries made by SBH while having 26% more discoveries. The p-values of these discoveries are shown in the middle panel of Figure 2d, where x-axis is the p-value quantile and y-axis is the -log10 p-value. Hypotheses discovered by both methods have significantly smaller GTEx p-values while SBH-only p-values are smaller than `AdaFDR`-only p-values in the GTEx data; the latter is due to

the fact that SBH uses the same threshold for all p-values. On the MuTHER validation data, the eQTLs discovered only by AdaFDR have more significant p-values than the eQTLs discovered only by SBH. This reveals a counter-intuitive behaviour of AdaFDR: it rejects some hypotheses with larger p-values if these SNPs have covariates that indicate a higher likelihood of eQTL. The MuTHER data validates this strategy—AdaFDR is able to discover more eQTLs on GTEx and the discovered eQTLs have more significant replication results on MuTHER.

AdaFDR can be broadly applied to any multiple testing problem where we have covariates for the hypotheses. This includes many highthroughput biological studies beyond eQTL. Here we evaluate its applications to RNA-seq, microbiome, proteomics and fMRI imaging data. In all cases, AdaFDR significantly outperforms current state-of-the-art methods.

Small GTEx data

AdaPT cannot be run on the full GTEx data due to its computational limitations. In order to perform a direct comparison between AdaFDR and AdaPT, we created a small GTEx data that contains the first 300k associations from chromosome 21 for the two adipose tissues. Even this small data takes AdaPT around 15 hours to process compared to less than 20 minutes for other methods. As shown in Figure 3a, AdaFDR has most number of discoveries in both experiments while AdaPT has slightly less. In addition, all covariate-adaptive methods have significant improvement over the non-adaptive methods (BH, SBH).

RNA-Seq data

We considered three RNA-Seq datasets that were used for differential expression analysis in AdaPT and IHW, i.e. the Bottomly data[15], the Pasilla data[16] and the airway data[14]. Here, the log expression level is used as the covariate, and the FDR level is set to be 0.1. The results are shown in Figure 3a, where AdaFDR and AdaPT have a similar number of discoveries (AdaFDR is consistently higher), and both are substantially more powerful than others. All covariate-adaptive methods make significantly more discoveries than the non-adaptive methods. In addition, the covariate patterns learned by AdaFDR are shown in Figure 3b for the Bottomly data and the Pasilla data, and in Supplementary Figure 2c for the airway data. The alternative hypotheses are more likely to occur when the expression levels are high, consistent with previous findings[12,13,24].

Microbiome data

We considered a subset of microbiome data from the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA), where samples were acquired from monitoring wells in a site contaminated by former waste disposal ponds and all sampled wells have various geochemical and physical measurements[17,18]. Following the original study, we performed two experiments to test for correlations between the operational taxonomic units (OTUs) and the pH, Al respectively. Ubiquity and the mean nonzero abundance are used are covariates, where the ubiquity is defined as the proportion of samples in which the OTU is present. The FDR level is set to be 0.2 for more discoveries and the fast version of AdaFDR is used due to the small sample size. As shown in Figure 3a, AdaFDR is significantly more powerful than other methods. The covariates are visualized in Figure 3c for the pH test and Supplementary Figure 2b for the Al test. The alternative hypotheses are more likely to occur when both the ubiquity and the mean nonzero abundance are high. This may be because that a higher level of these two quantities improves the detection power similar to the expression level in the RNA-Seq case.

Proteomics data

We considered a proteomics dataset where yeast cells treated with rapamycin were compared to yeast cells treated with dimethyl sulfoxide (2×6 biological replicates)[12,19]. Differential abundance of 2,666 proteins is evaluated using Welch's t-test. The total number of peptides is used as covariate that is quantified across all samples for each protein. The FDR level is set to be 0.1 and the fast version of AdaFDR is used due to the small sample size. As shown in Figure 3a, AdaFDR is significantly more powerful than other methods. The covariate is visualized in Figure 3d where a higher level of peptides increases the likelihood for the alternative hypotheses to occur. This is expected since the peptides level is similar to the expression level in the RNA-Seq data.

fMRI data

We considered two functional magnetic resonance imaging (fMRI) experiments where the human brain is divided spatially into isotropic voxels and the null hypothesis for each voxel is that there is no response to the stimulus[20]. The first experiment was done on a single participant with auditory stimulus and the second was done on a healthy adult female participant where the stimulus was to ask the person to imagine playing tennis[21]. We use the Brodmann area label, which represents different functional regions of the human brain[22], as covariate for each voxel. The FDR level is set to be 0.1 and the fast version of AdaFDR is used due to the inflation of p-values at 1. As shown in Figure 3a, AdaFDR is significantly more powerful than other methods. The result of AdaPT is omitted since it does not support categorical covariates, and directly running the GAM model yields a result much worse than BH. The covariate is visualized in Figure 3e. For the auditory experiment, the Brodmann areas

corresponding to auditory cortices, namely 41, 42, 22, are among areas where the alternative hypotheses are most likely to occur. For the tennis imagination experiment, multiple cortices seem to respond to this stimulus, including auditory cortex (42), visual cortices (18,19), and motor cortices (4,6,7).

Simulation studies

In order to systematically quantify the FDP and power of all the methods, we conducted extensive analysis of synthetic data where we know the ground truth. Each experiment is repeated 10 times and 95% confidence intervals are provided. In Figure 4a, the top two panels correspond to a simulated data with one covariate while the bottom two panels correspond to a simulated data with weakly-dependent p-values generated according to a previous paper[4]. In both simulations, all methods control FDR while `AdaFDR` has significantly larger power. Additional simulation experiments with strongly-dependent p-values and higher dimensional covariates can be found in Supplementary Figure 4a, where similar results are observed. Detailed descriptions of the synthetic data can be found in Supplementary Section 3.

We also investigate the running time of different methods. In Figure 4b, all experiments are repeated 5 times and the 95% confidence intervals are provided. The top panel uses a simulated dataset with 2d covariate, with the number of hypotheses varying from 20k to 100k. `AdaFDR-fast` takes 10s to run while both `AdaFDR` and IHW finished within a reasonable time of around 100s. AdaPT, however, needs a few hours to finish, significantly slower than other methods. In the bottom panel, the number of hypotheses is fixed to be 50k and the covariate dimension varies from 2 to 8; a similar result is observed.

After we have finished our initial paper, a very recent work[18] on the bioRxiv (October 31, 2018) proposed a new set of benchmark experiments to compare state-of-the-art multiple testing methods including IHW, AdaPT and an additional method of Boca-Leek (BL) that is on the bioRxiv[47]. We use their main simulation benchmark that includes two RNA-Seq *in silico* experiments, one experiment with uninformative covariate, and another two experiments that vary the number of hypotheses and the null proportion respectively. We run `AdaFDR` on this benchmark without any modification or tuning; `AdaFDR` achieves greater power than all other methods while controlling FDR (Supplementary Figure 4, 5). `AdaFDR` reduces to SBH when the covariate is not informative, indicating that it is not overfitting the uninformative covariate (Supplementary Figure 4e).

Discussion

Here we propose `AdaFDR`, a fast and flexible method that efficiently utilizes covariate information to increase detection power. Extensive experiments show that `AdaFDR` has greater power than existing approaches while controlling FDR. We discuss some of its characteristics and limitations.

Our theory proves that `AdaFDR` controls FDP in the setting when the null hypotheses are independent (the alternative hypotheses can have arbitrary correlations, see Theorem 1). This is a standard assumption also used in BH, SBH, IHW and AdaPT. To check the robustness of `AdaFDR` when there is model mismatch, we have performed systematic simulations with different p-value correlation structures to demonstrate that `AdaFDR` still controls FDP even when the null hypotheses are not independent. Moreover, although there are correlations among SNPs in the eQTL study, we show that the discoveries made by `AdaFDR` on the GTEX data replicate well on the independent MuTHER data with a different cohort. These suggest that `AdaFDR` behaves well when there is a dependency between null p-values. Since none of the other methods popular methods—BH, SBH, IHW, AdaPT—provides FDR control under arbitrary dependency, our comparison experiments are fair. `AdaFDR` can potentially be extended to allow arbitrary dependency using a similar idea as discussed in IHW[13]. Specifically, hypotheses should be split in such a way that the p-values from the two folds are independent, though they may have dependency within each fold. As a result, the learned threshold is independent of the fold it is applied onto. Then ideas discussed in the Benjamini-Yekutieli paper[6] can be used to scale the threshold to allow arbitrary dependency[13].

The typical use-case for `AdaFDR` is when there are many hypotheses to be tested simultaneously — ideally more than 10k. This is because `AdaFDR` needs many data to learn the covariate-adaptive threshold and to have an accurate estimate of FDP. A similar recommendation on the number of hypotheses is also made for IHW. When we have a smaller number of hypotheses, the discoveries are still valid but need to be treated with precaution — ideally with some orthogonal validations.

The scalability of `AdaFDR` and its ability to handle multivariate discrete and continuous covariates makes it broadly applicable to any multiple testing applications where additional information is available. While we focus on genomics experiments in this paper—because most of the previous methods were also evaluated on genomics experiments — it would be interesting to also apply `AdaFDR` to other domains such as imaging association analysis.

Methods

Definitions and notations

Suppose we have N hypothesis tests and each of them can be characterized by a p-value P_i, a d-dimensional covariate $\mathbf{x}_i$, and a indicator variable h_i with $h_i = 1$ representing the hypothesis to be true alternative. Then the set of true null hypotheses $\mathscr{H}_0$

and the set of true alternative hypotheses $\mathscr{H}_1$ can be written as $\mathscr{H}_0 = \{i : i \in [N], h_i = 0\}$ and $\mathscr{H}_1 = \{i : i \in [N], h_i = 1\}$, where we adopt the notation $[N] \overset{def}{=} \{1, 2, \cdots, N\}$. Given a threshold function $t(\mathbf{x})$, we reject the ith null hypothesis if $P_i \leq t(\mathbf{x_i})$. The number of discoveries $\mathrm{D}(t)$ and the number of false discoveries $\mathrm{FD}(t)$ can be written as $\mathrm{D}(t) \overset{def}{=} \sum_{i\in[N]} \mathbb{I}_{\{P_i \leq t(\mathbf{x}_i)\}}$ and $\mathrm{FD}(t) \overset{def}{=} \sum_{i\in\mathscr{H}_0} \mathbb{I}_{\{P_i \leq t(\mathbf{x}_i)\}}$. The false discovery proportion (FDP) is defined as $\mathrm{FDP}(t) \overset{def}{=} \frac{\mathrm{FD}(t)}{\mathrm{D}(t)\vee 1}$, where $a \vee b \overset{def}{=} \max(a, b)$. The expected value of FDP is the false discovery rate (FDR): $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$[23].

Multiple testing via `AdaFDR`

`AdaFDR` can take as input multi-dimensional covariates $\mathbf{x}$. The key assumption is that the null p-values remain uniform regardless of the covariate value while others, including the alternative p-values and the likelihood for the hypotheses to be true null/alternative, may have arbitrary dependencies on the covariate. This is a standard assumption in the literature[12,23,24]. For example, in the case of AAF, the null p-values are uniformly distributed independent of AAF since the gene expression has no association with the SNP under the null hypothesis. However, the alternative p-values may depend on AAF since the associations are easier to detect/yield smaller p-values if the AAF is close to 0.5.

`AdaFDR` aims to optimize over a set of decision rules $t(\mathbf{x}) \in \mathscr{T}$ to maximize the number of discoveries, subject to the constraint that the FDP is less than a user-specified nominal level α. Conceptually, this optimization problem can be written as

$$\underset{t\in\mathscr{T}}{\text{maximize}}\ \mathrm{D}(t),\ \ s.t.\ \ \mathrm{FDP}(t) \leq \alpha. \tag{1}$$

There are three challenges in this optimization problem: 1. the set of decision thresholds $\mathscr{T}$ needs to be parameterized in such a way that both captures the covariate information and scales well with the covariate dimension; 2. the actual FDP is not directly available from the data; 3. direct optimization of (1) may cause overfitting and hence lose FDR control.

For the first challenge, intuitively, the decision threshold should have large values where the alternative hypotheses are enriched. Such enrichment pattern, as discussed the `NeuralFDR` paper[27], usually consists of local "bumps" at certain covariate locations and a global "slope" that represents generic monotonic relationships. For example, the distance from TSS and the AAF in Figure 2c correspond to the bump structure (at 0 and 0.5 respectively) whereas the rest of the covariates correspond to the slope structure. `AdaFDR` addresses these two structures by using a mixture of generalized linear model (GLM) and K-component Gaussian mixture (with diagonal covariance matrices), i.e.,

$$t(\mathbf{x}) = \exp(\mathbf{a}^T\mathbf{x} + b) + \sum_{k=1}^{K} \exp\left[w_k - (\mathbf{x} - \boldsymbol{\mu}_k)^T \mathrm{diag}(\boldsymbol{\sigma}_k)(\mathbf{x} - \boldsymbol{\mu}_k)\right], \tag{2}$$

where $\mathrm{diag}(\boldsymbol{\sigma}_k)$ represents a diagonal matrix with diagonal elements specified by the d-dimensional vector $\boldsymbol{\sigma}_k$. The set of parameters to optimize can be written as $\{\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}, \{w_k \in \mathbb{R}, \boldsymbol{\mu}_k \in \mathbb{R}^d, \boldsymbol{\sigma}_k \in \mathbb{R}^d\}_{k=1}^K\}$. We choose to use the diagonal covariance matrices for Gaussian mixture to speed up the optimization. As a result, the number of parameters grows linearly with respect to the covariate dimension d, and the parameters can be easily initialized via EM algorithm, as described below.

For the second challenge, we use a "mirror estimator" to estimate the number of false discoveries of a given threshold function t,

$$\text{mirror estimator:}\quad \widehat{\mathrm{FD}}(t) \overset{def}{=} \sum_{i=1}^{N} \mathbb{I}_{\{P_i \geq 1 - t(\mathbf{x}_i)\}}.$$

Such estimator has been used in recent works[24,27,48,49] and yields a conservative estimate of the true number of false discoveries (FD), in the sense that its expected value is larger than that of the true FD under mild assumptions (Lemma 1 in Supplementary Materials). Furthermore, FDP can be simply estimated as $\widehat{\mathrm{FDP}}(t) = \frac{\widehat{\mathrm{FD}}(t)}{\mathrm{D}(t)}$.

For the third challenge, `AdaFDR` controls FDP with high probability via hypothesis splitting. The hypotheses are randomly split into two folds; a separate decision threshold is learned on each fold and applied on the other. Since the learned threshold does not depend on the fold of data onto which it is applied, FDP can be controlled with high probability — such statement is made formal in Theorem 1. We note that in multiple testing by `AdaFDR`, the learning-and-testing process is repeated twice, with each fold being the training set at one time and the testing set at the other. Figure 5 shows one of such process with fold 1 being the training set.

The full algorithm is described in Algorithm 1. Here, for example, $\mathrm{D}_{\mathrm{train}}(t)$, $\mathrm{D}_{\mathrm{test}}(t)$ are understood as the number of discoveries on the training set and the testing set respectively. Similar notations are used for other quantities like $\mathrm{FDP}(t)$ and the mirror estimate $\widehat{\mathrm{FDP}}(t)$ without explicit definition.

`AdaFDR` follows a similar strategy as our preliminary work `NeuralFDR`[27], which it subsumes: both methods use the mirror estimator to estimate FDP and use hypothesis splitting for FDP control. The main difference is on the modeling of the

Algorithm 1 `AdaFDR` for multiple hypothesis testing

1: Randomly split the data $\mathscr{D} = \{(P_i, \mathbf{x}_i)\}_{i=1}^N$ into two folds $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2$ of equal size.
2: **for** $(j, j') = (1,2), (2,1)$ **do**
3: Set $\mathscr{D}_j$ to be the training set and $\mathscr{D}_{j'}$ the testing set.
4: Learn the decision threshold $t^*(\mathbf{x})$ on the training set by optimizing

$$\underset{t}{\text{maximize}} \ \ \mathrm{D}_{\text{train}}(t) \ \ s.t. \ \ \widehat{\mathrm{FDP}}_{\text{train}}(t) \le \alpha. \tag{3}$$

5: Compute the best rescale factor γ^* on the testing set

$$\gamma^* = \sup_{\gamma > 0} \{\gamma : \widehat{\mathrm{FDP}}_{\text{test}}(\gamma t^*) \le \alpha\}. \tag{4}$$

6: Reject the hypotheses $\mathscr{R}_{j'} = \{i : i \in \mathscr{D}_{j'}, P_i \le \gamma^* t^*(\mathbf{x}_i)\}$.
7: Report discoveries on both folds $\mathscr{R} = \mathscr{R}_1 \cup \mathscr{R}_2$.

decision threshold t: `NeuralFDR` uses a neural network, which is flexible enough but hard to optimize. `AdaFDR`, in contrast, adopts the simpler mixture model that may lack certain flexibility but is much easier to optimize. This change of modeling, however, does not seems to reduce much of the detection power for `AdaFDR`. As shown in Supplementary Figure 3b, the performance of `AdaFDR` is similar to that of `NeuralFDR`, while `AdaFDR` is orders of magnitude faster.

Optimization

Recall that the optimization is done solely on the training set $\mathscr{D}_{\text{train}}$. Substituting FDP in (1) with its mirror estimate we can rewrite the optimization problem as

$$\underset{t \in \mathscr{T}}{\text{maximize}} \ \mathrm{D}_{\text{train}}(t), \ \ s.t. \ \ \frac{\widehat{\mathrm{FD}}_{\text{train}}(t)}{\mathrm{D}_{\text{train}}(t)} \le \alpha, \tag{5}$$

where $\mathscr{T}$, the set of decision thresholds to optimize over, corresponds to the mixture model (2). Our strategy is to first compute a good initialization point and then perform optimization by gradient descent on a relaxed problem. We note that a better solution to the optimization problem will give a better detection power. However, the FDP control guarantee holds *regardless* of the decision threshold we come up with.

- **Initialization:** Let $\pi_0(\mathbf{x})$ and $\pi_1(\mathbf{x})$ be the distributions for the null hypotheses and the alternative hypotheses, over the covariate $\mathbf{x}$, respectively. Following the intuition that the threshold $t(\mathbf{x})$ should be large when the number of alternative hypotheses is high and the number of null hypotheses is low, it is a good heuristic to let

$$t(\mathbf{x}) \propto \frac{\pi_1(\mathbf{x})}{\pi_0(\mathbf{x})}.$$

  This is done in `AdaFDR` as follows. First, covariates with p-values larger than 0.75, i.e. $\{\mathbf{x}_i : i \in \mathscr{D}_{\text{train}}, P_i \ge 0.75\}$, are treated as an approximate ensemble of the null hypotheses, and those with p-values smaller than the BH threshold, i.e. $\{\mathbf{x}_i : i \in \mathscr{D}_{\text{train}}, P_i \le t_{\text{BH}}\}$, are treated as an approximate ensemble of the alternative hypotheses. Then first, a mixture model same as (2) is fitted on the null ensemble $\{\mathbf{x}_i : i \in \mathscr{D}_{\text{train}}, P_i \ge 0.75\}$ using EM algorithm, resulting in an estimate of the null hypothesis distribution $\hat{\pi}_0(\mathbf{x})$. Second, each point in the alternative ensemble $\{\mathbf{x}_i : i \in \mathscr{D}_{\text{train}}, P_i \le t_{\text{BH}}\}$ receives a sample weight $1/\hat{\pi}_0(\mathbf{x})$. Last, the mixture model (2) is fitted on the weighted alternative ensemble using EM algorithm to obtain the final initialization threshold. The details of the EM algorithm can be found in Supplementary SubSection 2.3.

- **Optimization:** First, a Lagrangian multiplier is used to deal with the constraint:

$$\underset{t \in \mathscr{T}}{\text{minimize}} \ -\mathrm{D}_{\text{train}}(t) + \lambda_1 [\widehat{\mathrm{FD}}_{\text{train}}(t) - \alpha \mathrm{D}_{\text{train}}(t)] \vee 0, \tag{6}$$

  where λ_1 is chosen heuristically to be $10/\alpha$. Second, the sigmoid function is used to deal with the discontinuity of the

indicator functions in $\mathrm{D}_{\text{train}}(t)$ and $\widehat{\mathrm{FD}}_{\text{train}}(t)$:

$$\mathrm{D}_{\text{train}}(t) = \sum_{i\in\mathscr{D}_{\text{train}}} \mathbb{I}_{\{P_i\le t(\mathbf{x}_i)\}} \approx \sum_{i\in\mathscr{D}_{\text{train}}} S[\lambda_0(t(\mathbf{x}_i)-P_i)],$$
$$\widehat{\mathrm{FD}}_{\text{train}}(t) = \sum_{i\in\mathscr{D}_{\text{train}}} \mathbb{I}_{\{P_i\ge 1-t(\mathbf{x}_i)\}} \approx \sum_{i\in\mathscr{D}_{\text{train}}} S[\lambda_0(P_i-1+t(\mathbf{x}_i))],$$

where $S(\cdot)=\frac{1}{1+e^{-x}}$ is the sigmoid function and λ_0 is automatically chosen at the beginning of the optimization such that the smoothed versions are good approximations to the original ones. Finally, the Adam optimizer[50] is used for gradient descent.

FDP control

We would like to point out that the mirror estimate is more accurate when its value is large. Hence, when the number of rejections is small (<100), the result should be treated with precaution. However, this should not be a major concern since in the target applications of `AdaFDR`, usually thousands to millions of hypotheses are tested simultaneously, and hundreds to thousands of hypotheses are rejected. In those cases, the mirror estimate is accurate. Hence, we further require that for each fold, the best scale factor γ^* should have a number of discoveries exceeding c_0N for some pre-specified small proportion c_0; failing to satisfy this condition will result in no rejection in this fold. In other words, we consider a modified version of Alg. 1 with (4) substituted by setting

$$\gamma^* = \sup_{\gamma>0}\{\gamma : \widehat{\mathrm{FDP}}_{\text{test}}(\gamma t^*)\le\alpha, \mathrm{D}_{\text{test}}(\gamma t^*)\ge c_0N\}\cup\{0\}. \tag{7}$$

Our FDP control on this modified version can be stated as follows.

Theorem 1. *(FDP control) Assume that all null p-values $P_i\in\mathscr{H}_0$, conditional on the covariates, are independently and identically distributed (i.i.d.) following* Unif$[0,1]$. *Then with probability at least* 1-δ, `AdaFDR` *with the modification* (7) *controls FDP at level* $(1+\varepsilon)\alpha$, *where* $\varepsilon = O\left(\sqrt{\frac{\log\frac{1}{\delta}}{\alpha N}}\right)$.

The assumption made in Theorem 1 is standard in the literature[13,24] and can be easily relaxed to the assumption that the null p-values, conditional on the covariates, are independently distributed and stochastically greater than Unif$[0,1]$ (Supplementary SubSection 4.1). In addition, Theorem 1 is strictly stronger than the one for `NeuralFDR` (Supplementary SubSection 2.2).

Covariate visualization via `AdaFDR_explore`

`AdaFDR` also provides a `FeatureExplore` function that can visualize the relationship between each covariate and the significance of hypotheses, in terms of estimated distributions for the null hypothesis and the alternative hypothesis with respect to each covariate, as those shown in Figure 2c and Figure 4b. This is done as follows. First, for the entire dataset, covariates with p-values greater than 0.75, i.e. $\{\mathbf{x}_i : i\in[N], P_i\ge 0.75\}$, are treated as an approximate ensemble of the null hypotheses, and those with p-values less than the BH threshold, i.e. $\{\mathbf{x}_i : i\in[N], P_i\le t_{\text{BH}}\}$, are treated as an approximate ensemble of the alternative hypotheses. Then, the null hypothesis distribution and the alternative hypothesis distribution are estimated from these two ensembles using kernel density estimation (KDE) for continuous covariates and simple count estimator for categorical covariates. In addition, for categorical covariates, the categories are reordered based on the ratio between the estimated alternative probability and null probability $\hat{\pi}_1(\mathbf{x})/\hat{\pi}_0(\mathbf{x})$.

References

1. Benjamini, Y. & Hochberg, Y. Multiple hypotheses testing with weights. *Scand. J. Stat.* **24**, 407–418 (1997).
2. Dunn, O. J. Multiple comparisons among means. *J. Am. statistical association* **56**, 52–64 (1961).
3. Storey, J. D. A direct approach to false discovery rates. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **64**, 479–498 (2002).
4. Storey, J. D., Taylor, J. E. & Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **66**, 187–205 (2004).
5. Efron, B. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1 (Cambridge University Press, 2012).

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

7. Consortium, G. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).

8. Bernstein, B. E. *et al.* The nih roadmap epigenomics mapping consortium. *Nat. biotechnology* **28**, 1045 (2010).

9. McCullagh, P. & Nelder, J. A. *Generalized linear models*, vol. 37 (CRC press, 1989).

10. Pregibon, D. & Hastie, T. J. Generalized linear models. In *Statistical Models in S*, 195–247 (Routledge, 2017).

11. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, NY, USA:, 2001).

12. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. methods* **13**, 577–580 (2016).

13. Ignatiadis, N. & Huber, W. Covariate-powered weighted multiple testing with false discovery rate control. *arXiv preprint arXiv:1701.05179* (2017).

14. Himes, B. E. *et al.* Rna-seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PloS one* **9**, e99625 (2014).

15. Bottomly, D. *et al.* Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using rna-seq and microarrays. *PloS one* **6**, e17820 (2011).

16. Brooks, A. N. *et al.* Conservation of an rna regulatory map between drosophila and mammals. *Genome research* **21**, 193–202 (2011).

17. Smith, M. B. *et al.* Natural bacterial communities serve as quantitative geochemical biosensors. *MBio* **6**, e00326–15 (2015).

18. Korthauer, K. *et al.* A practical guide to methods controlling false discoveries in computational biology. *bioRxiv* DOI: 10.1101/458786 (2018). https://www.biorxiv.org/content/early/2018/10/31/458786.full.pdf.

19. Dephoure, N. & Gygi, S. P. Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci. Signal.* **5**, rs2–rs2 (2012).

20. Schildknecht, K., Tabelow, K. & Dickhaus, T. More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses. *PloS one* **11**, e0149016 (2016).

21. Tabelow, K., Polzehl, J. *et al. Statistical parametric maps for functional MRI experiments in R: The package fmri* (WIAS, 2010).

22. Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues* (Barth, 1909).

23. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. royal statistical society. Ser. B (Methodological)* 289–300 (1995).

24. Lei, L. & Fithian, W. Adapt: an interactive procedure for multiple testing with side information. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **80**, 649–679 (2018).

25. Li, A. & Barber, R. F. Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926* (2016).

26. Lei, L., Ramdas, A. & Fithian, W. Star: A general interactive framework for fdr control under structural constraints. *arXiv preprint arXiv:1710.02776* (2017).

27. Xia, F., Zhang, M. J., Zou, J. Y. & Tse, D. Neuralfdr: Learning discovery thresholds from hypothesis features. In *https://arxiv.org/pdf/1711.01312.pdf* (2017).

28. Tansey, W., Wang, Y., Blei, D. M. & Rabadan, R. Black box fdr. *arXiv preprint arXiv:1806.03143* (2018).

29. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

30. Wen, X. *et al.* Molecular qtl discovery incorporating genomic annotations using bayesian false discovery rate control. *The Annals Appl. Stat.* **10**, 1619–1638 (2016).

31. Xiao, J., Cao, H. & Chen, J. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* **33**, 2873–2881 (2017).

32. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. journal statistics* 65–70 (1979).

33. Genovese, C. R., Roeder, K. & Wasserman, L. False discovery control with p-value weighting. *Biometrika* 509–524 (2006).

34. Roeder, K. & Wasserman, L. Genome-wide significance levels and weighted hypothesis testing. *Stat. science: a review journal Inst. Math. Stat.* **24**, 398 (2009).

35. Dobriban, E., Fortney, K., Kim, S. K. & Owen, A. B. Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika* **102**, 753–766 (2015).

36. Hu, J. X., Zhao, H. & Zhou, H. H. False discovery rate control with groups. *J. Am. Stat. Assoc.* **105**, 1215–1227 (2010).

37. Sankaran, K. & Holmes, S. structssi: simultaneous and selective inference for grouped or hierarchically structured data. *J. statistical software* **59**, 1 (2014).

38. Boca, S. M. & Leek, J. T. A regression framework for the proportion of true null hypotheses. *bioRxiv* 035675 (2015).

39. Efron, B. Simultaneous inference: When should hypothesis testing problems be combined? *The annals applied statistics* 197–223 (2008).

40. Cai, T. T. & Sun, W. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Am. Stat. Assoc.* **104**, 1467–1481 (2009).

41. Ferkingstad, E. *et al.* Unsupervised empirical bayesian multiple testing with external covariates. *The Annals Appl. Stat.* **2**, 714–735 (2008).

42. Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. & Kass, R. E. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Am. Stat. Assoc.* **110**, 459–471 (2015).

43. Zablocki, R. W. *et al.* Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30**, 2098–2104 (2014).

44. Consortium, G. *et al.* The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

45. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506 (2013).

46. Grundberg, E. *et al.* Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nat. genetics* **44**, 1084 (2012).

47. Boca, S. M. & Leek, J. T. A direct approach to estimating false discovery rates conditional on covariates. *bioRxiv* 035675 (2018).

48. Lei, L. & Fithian, W. Power of ordered hypothesis testing. In *International Conference on Machine Learning*, 2924–2932 (2016).

49. Arias-Castro, E., Chen, S. *et al.* Distribution-free multiple testing. *Electron. J. Stat.* **11**, 1983–2001 (2017).

50. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

51. Sherry, S. T. *et al.* dbsnp: the ncbi database of genetic variation. *Nucleic acids research* **29**, 308–311 (2001).

52. Huber, W. Reyes, a. pasilla: Data package with per-exon and per-gene read counts of rna-seq samples of pasilla knock-down by brooks et al. *Genome Res.* (2011).

53. Ignatiadis, N. *IHWpaper: Reproduce figures in IHW paper* (2018). R package version 1.7.0.

Acknowledgements

We would like to thank David Tse, Liuhua Lei, Nikolaos Ignatiadis and Vivek Bagaria for helpful discussions. MZ and FX are partially supported by Stanford Graduate Fellowship. JZ is supported by the Chan-Zuckerberg Initiative and National Science Foundation (NSF) Grant CRII 1657155.

Author contributions statement

MZ designed the algorithm and conducted the experiments with the help of FX. MZ performed the theoretical analysis. MZ and JZ wrote the manuscript. JZ supervised the research. All authors reviewed the manuscript.

Additional information

Code availability

- The code for the paper is available at https://github.com/martinjzhang/AdaFDRpaper
- The software is available at https://github.com/martinjzhang/adafdr

Data availability The GTEx data for the two adipose tissues and all other data are deposited into the online repository. See the github repository AdaFDRpaper for more information.

Competing interests. The authors declare that they have no competing financial interests.

Correspondence. Requests for materials should be addressed to JZ (e-mail: jamesz@stanford.edu).

Figures

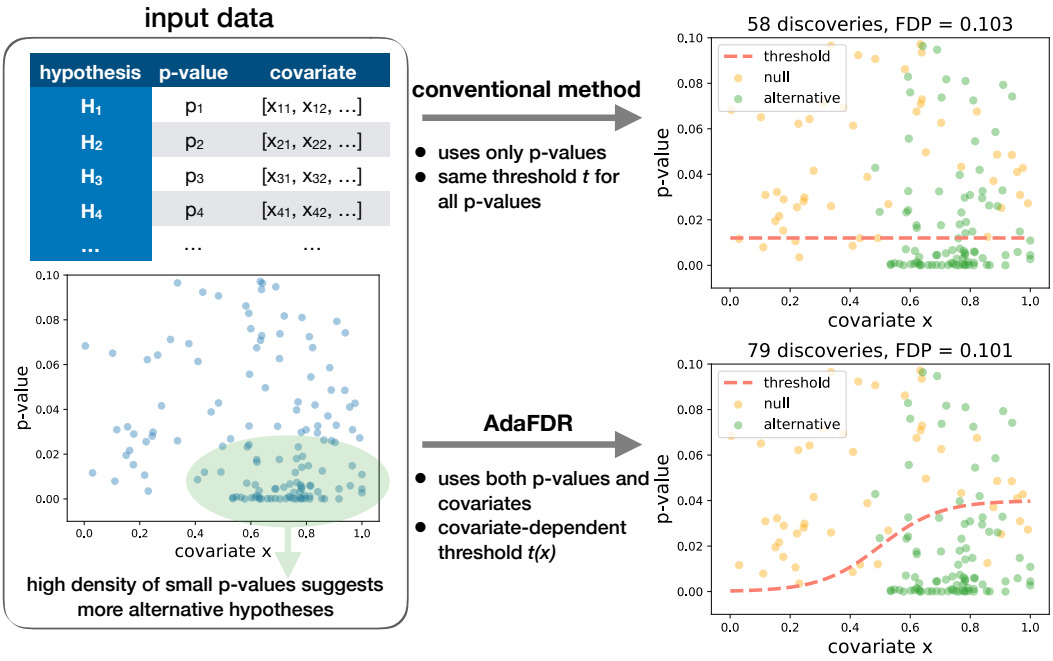
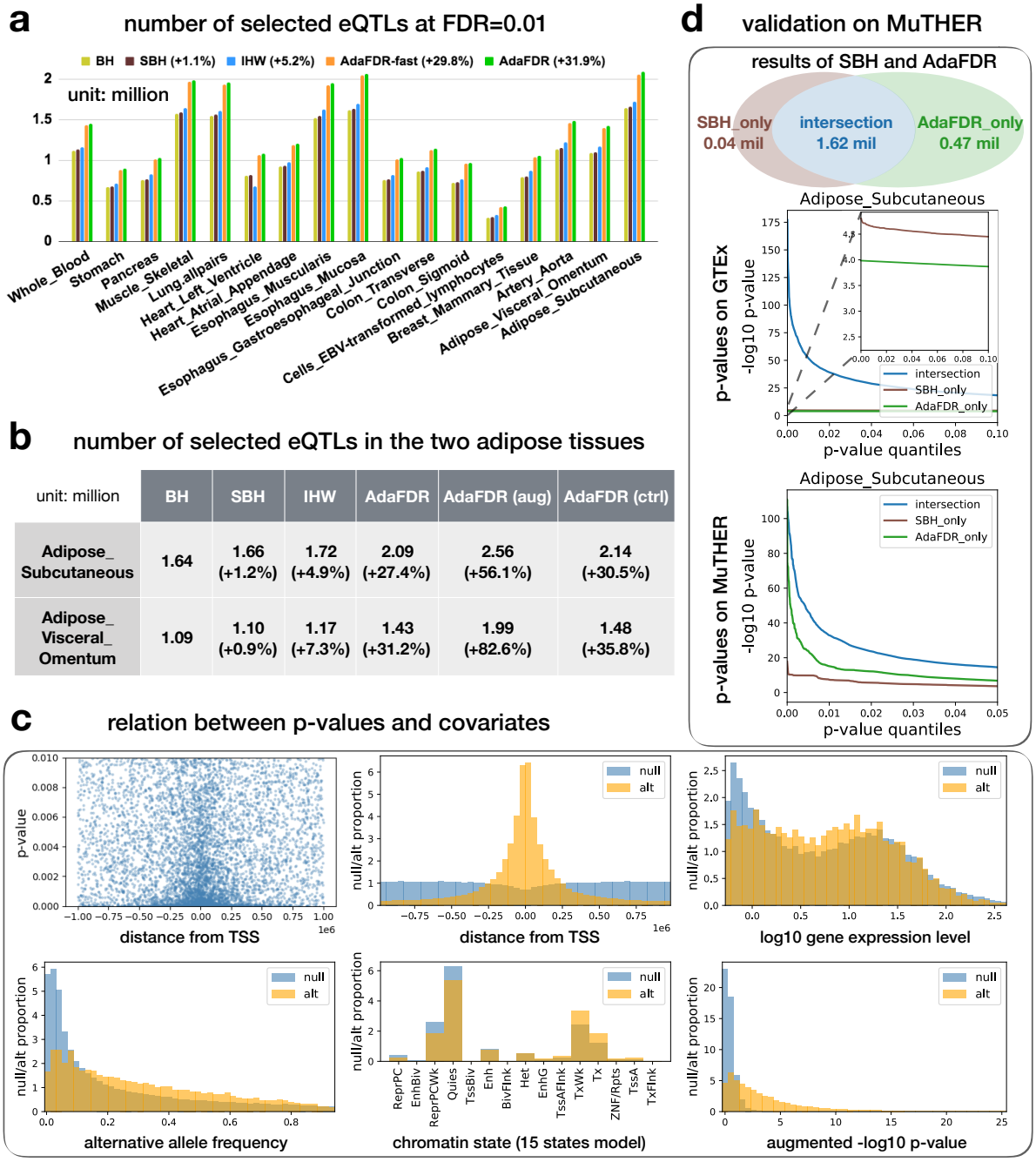


Figure 1. Intuition of `AdaFDR`. Top-left: As input, `AdaFDR` takes a list of hypotheses, each with a p-value and a covariate that could be multi-dimensional. Bottom-left: A toy example with a univariate covariate. The enrichment of small p-values in the bottom-right corner suggests that there are more alternative hypotheses there. Leveraging this structure can lead to more discoveries. Top-right: Conventional method uses only p-values and has the same p-value threshold for all hypotheses. Bottom-right: `AdaFDR` adaptively learns the uneven distribution of the alternative hypotheses, and makes more discoveries while controlling the false discovery proportion (FDP) at the desired level (0.1 in this case).

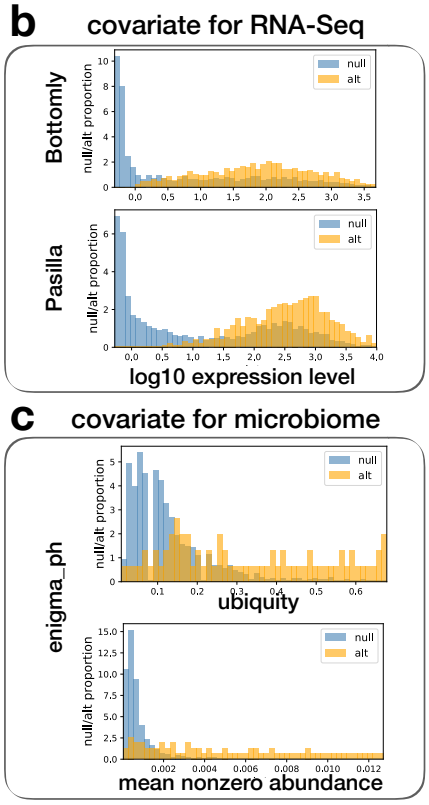


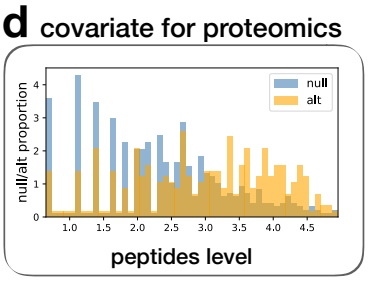
unit: million	BH	SBH	IHW	AdaFDR	AdaFDR (aug)	AdaFDR (ctrl)
Adipose_ Subcutaneous	**1.64**	**1.66 (+1.2%)**	**1.72 (+4.9%)**	**2.09 (+27.4%)**	**2.56 (+56.1%)**	**2.14 (+30.5%)**
Adipose_ Visceral_ Omentum	**1.09**	**1.10 (+0.9%)**	**1.17 (+7.3%)**	**1.43 (+31.2%)**	**1.99 (+82.6%)**	**1.48 (+35.8%)**

Figure 2. Analysis of the GTEx data. (a) Results of 17 tissues considered in the study. `AdaFDR` and its fast version consistently make more discoveries than other methods. (b) Results on the two adipose tissues where the -log10 p-value from another tissue was added as an extra covariate. Using p-values from a similar tissue (AdaFDR (aug)) yields significantly more discoveries than using p-values from an unrelated tissue (AdaFDR (ctrl)). (c) Top-left: P-values (y-axis) plotted against the distances from TSS (x-axis); each dot corresponds to one SNP-gene pair. Small p-values at the center suggest that hypotheses with smaller distances from TSS are more likely to be significant. Other panels: `AdaFDR`-estimated null hypothesis distribution (blue) and alternative hypothesis distribution (orange) with respect to each covariate. Higher values of the orange distribution suggest an enrichment of alternative hypotheses. (d) Top: Discoveries made by SBH and `AdaFDR`. Middle: The p-values of these discoveries—SBH-only p-values are smaller than `AdaFDR`-only p-values on GTEx. Bottom: The p-values of the same set of discoveries on the independent MuTHER data, where `AdaFDR`-only p-values are smaller than SBH-only p-values, suggesting that `AdaFDR`-only discoveries are more likely to be true discoveries.

a results in other applications

	BH	SBH	AdaPT	IHW	AdaFDR
small_GTEx: Adipose_ Subcutaneous	1182	1188 (+0.5%)	1333 (+12.8%)	1333 (+12.8%)	**1469 (+24.3%)**
small_GTEx: Adipose_ Visceral_Omentum	549	553 (+0.7%)	1037 (+88.9%)	724 (+31.9%)	**1360 (+148%)**
RNA-Seq: Bottomly	1583	1693 (+6.9)	2109 (+33.2%)	1714 (+8.3%)	**2144 (+35.4%)**
RNA-Seq: Pasilla	687	687 (+0.0%)	853 (+24.2%)	785 (+14.3%)	**856 (+24.6%)**
RNA-Seq: airway	4079	4079 (+0.0%)	6045 (+48.2%)	4862 (+19.2%)	**6050 (+48.3%)**
microbiome: enigma_ph	61	65 (+6.6%)	96 (+57.4%)	89 (+45.9%)	**124 (+103.3%)**
microbiome: enigma_al	206	437 (+112.1%)	496 (+140.8%)	243 (+18.0%)	**503 (+144.2%)**
proteomics	244	358 (+46.7%)	384 (+57.4%)	245 (+0.4%)	**402 (+64.8%)**
fMRI: auditory	888	888 (+0%)	-	1015 (+14.3%)	**1058 (+19.1%)**
fMRI: imagination	2141	2228 (+4.1%)	-	2151 (+0.5%)	**2239 (+4.6%)**





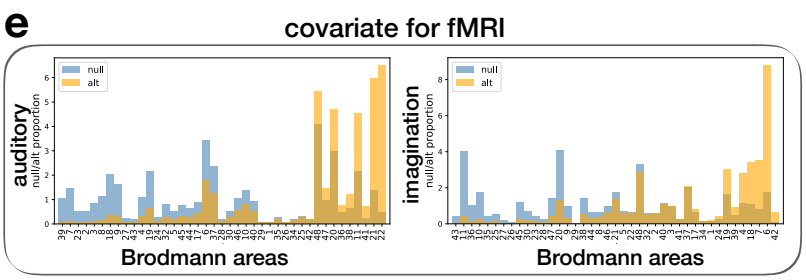


Figure 3. (a) The number of discoveries of various methods on two small GTEx eQTL datasets, three RNA-Seq differential expression datasets, two microbiome datasets, one proteomics dataset, and two fMRI datasets. The fMRI results for AdaPT are omitted since the AdaPT software does not support categorical covariates. (b) Covariate visualization for RNA-Seq datasets. (c) Covariate visualization for microbiome dataset. (d) Covariate visualization for proteomics dataset. (e) Covariate visualization for fMRI datasets.

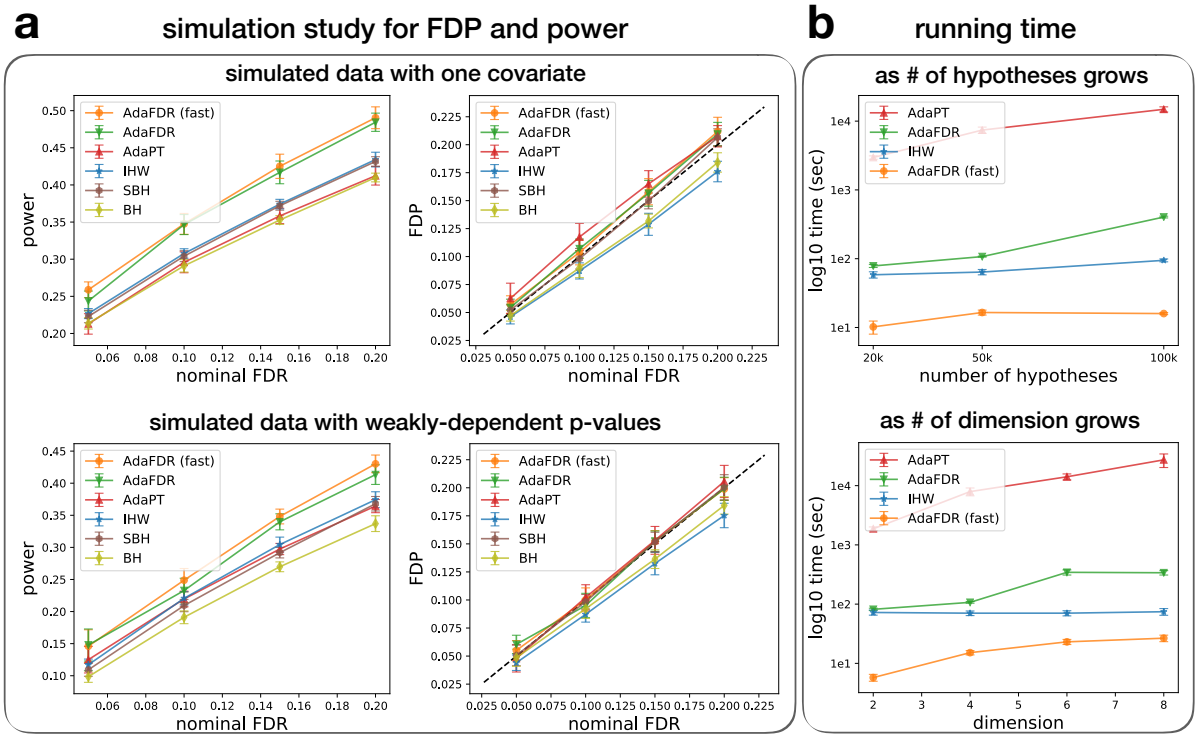


Figure 4. (a) The number of discoveries of various methods on three RNA-Seq differential expression datasets and two small GTEx eQTL datasets. (b) Covariate visualization for RNA-Seq datasets. Note that the AdaPT software does not support categorical covariates and can not be ran on the two fMRI datasets. (c) Simulation of FDP and power on an independent case (top) and a weakly-dependent case (bottom). (d) Running time analysis. Top: as the number of hypotheses grows. Bottom: as the covariate dimension grows.

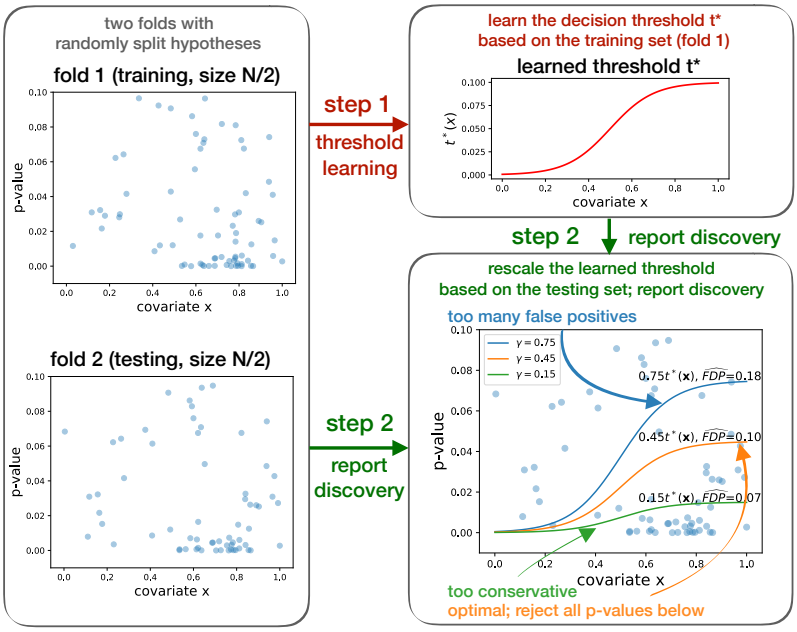


Figure 5. Schematic of the `AdaFDR` learning and testing process. Fold 1 is the training set and fold 2 is the testing set (left panel). In step 1, a decision threshold $t^*(\mathbf{x})$ is learned on the training set via solving the optimization problem (1) (upper-right panel). In step 2, as shown in the bottom-right panel, this learned threshold $t^*(\mathbf{x})$ is first rescaled by a factor γ^*, defined as the largest number whose corresponding mirror-estimated FDP on the testing set is less than α (orange). Then all p-values on the testing set below the rescaled threshold are rejected. Here the nominal FDP is $\alpha = 0.1$.

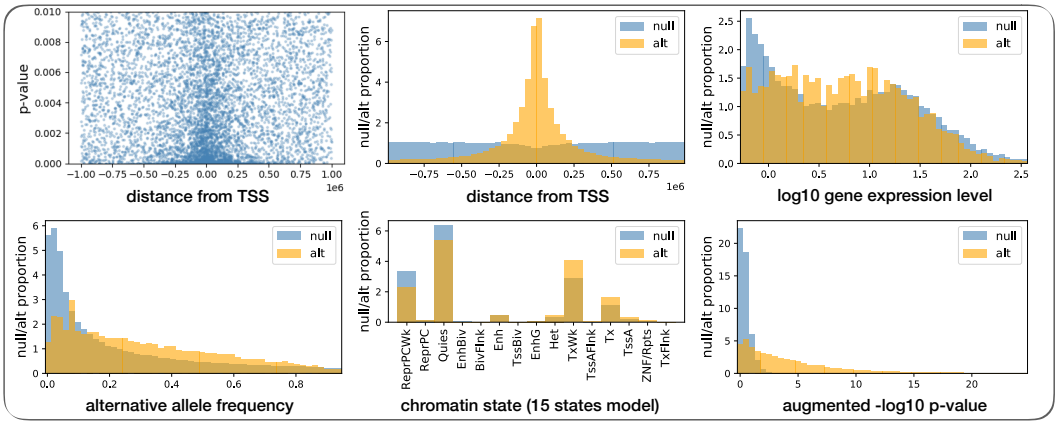
Supplemental Materials

1 Additional Results

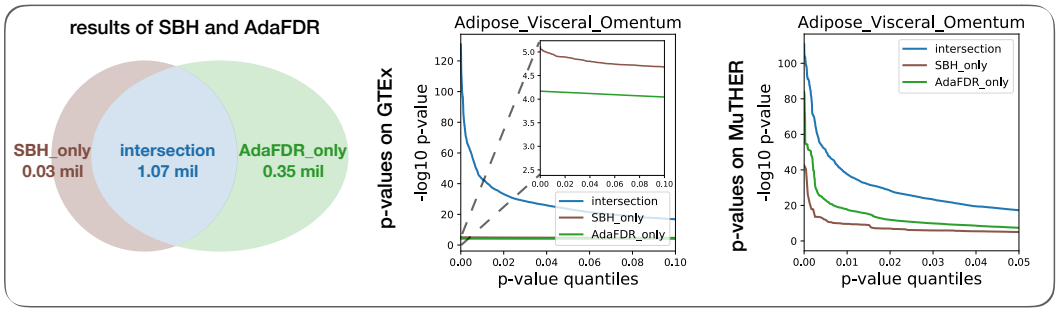
a number of selected eQTLs in the two colon tissues

unit: million	BH	SBH	IHW	AdaFDR	AdaFDR (aug)	AdaFDR (ctrl)
Colon_ Sigmoid	0.72	0.73 (+1.0%)	0.77 (+6.6%)	0.97 (+34.6%)	1.18 (+63.9%)	1.03 (+43.0%)
Colon_ Transverse	0.87	0.88 (+0.9%)	0.92 (+5.5%)	1.14 (+31.8%)	1.33 (+52.9%)	1.16 (+33.3%)

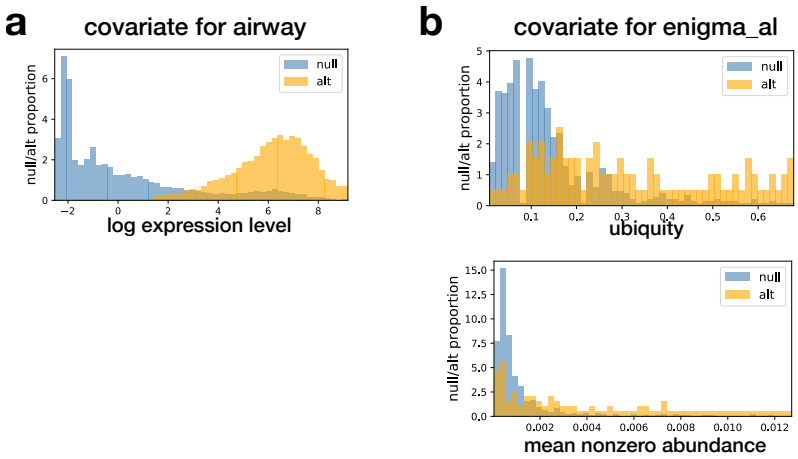
b relation between p-values and covariates for Colon_Sigmoid



c validation on MuTHER for Adipose_Visceral_Omentum

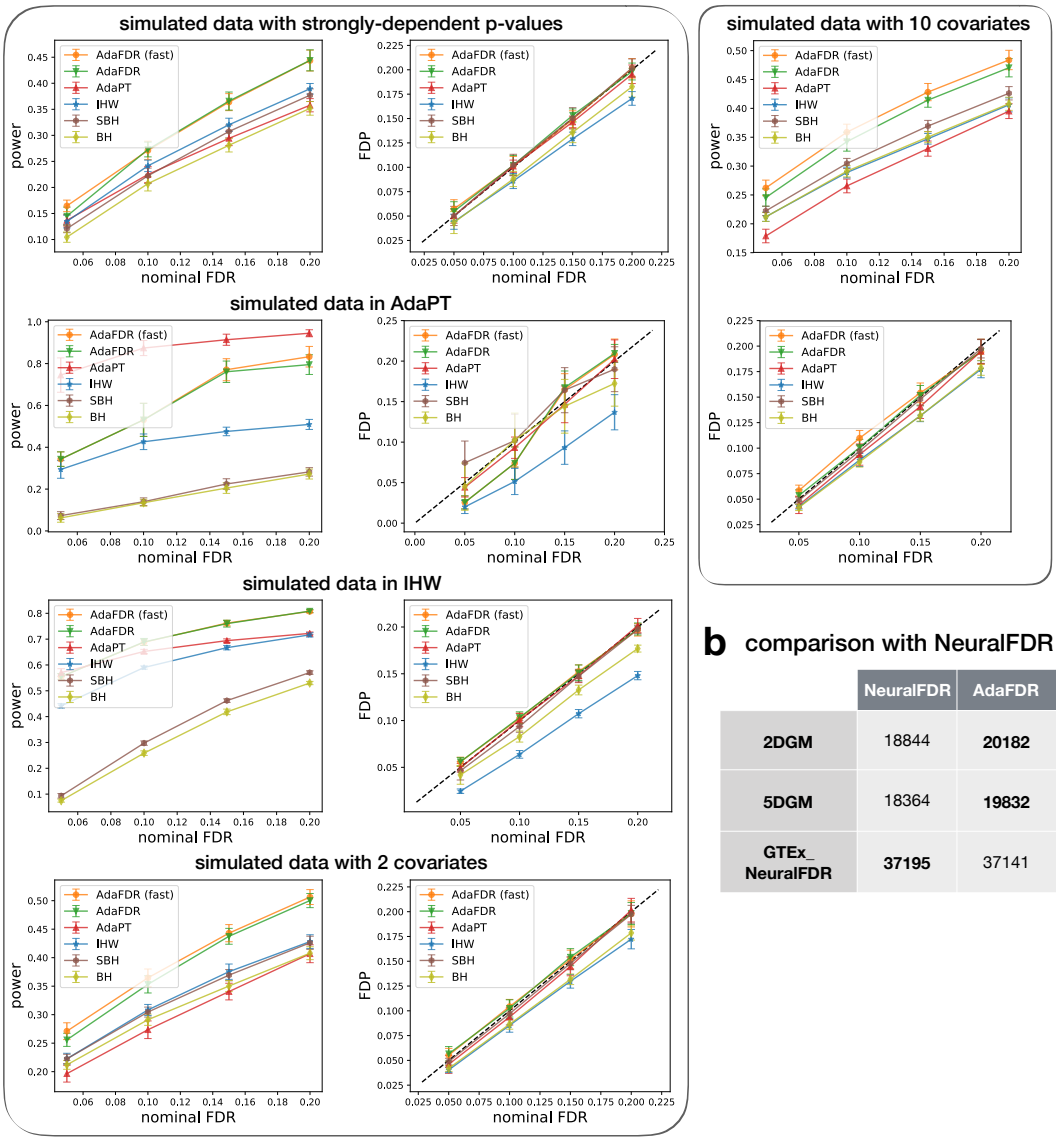


Supplementary Figure 1. Additional results on the GTEx data. (a) Results on the two colon tissues. (b) Feature visualization for Colon_Sigmoid (c) Validation on MuTHER for Adipose_Visceral_Omentum.



Supplementary Figure 2. (a) The covariate visualization for the RNA-Seq airway data. (b) The covariate visualization for the microbiome enigma_al data. Top: ubiquity; bottom: mean nonzero abundance.

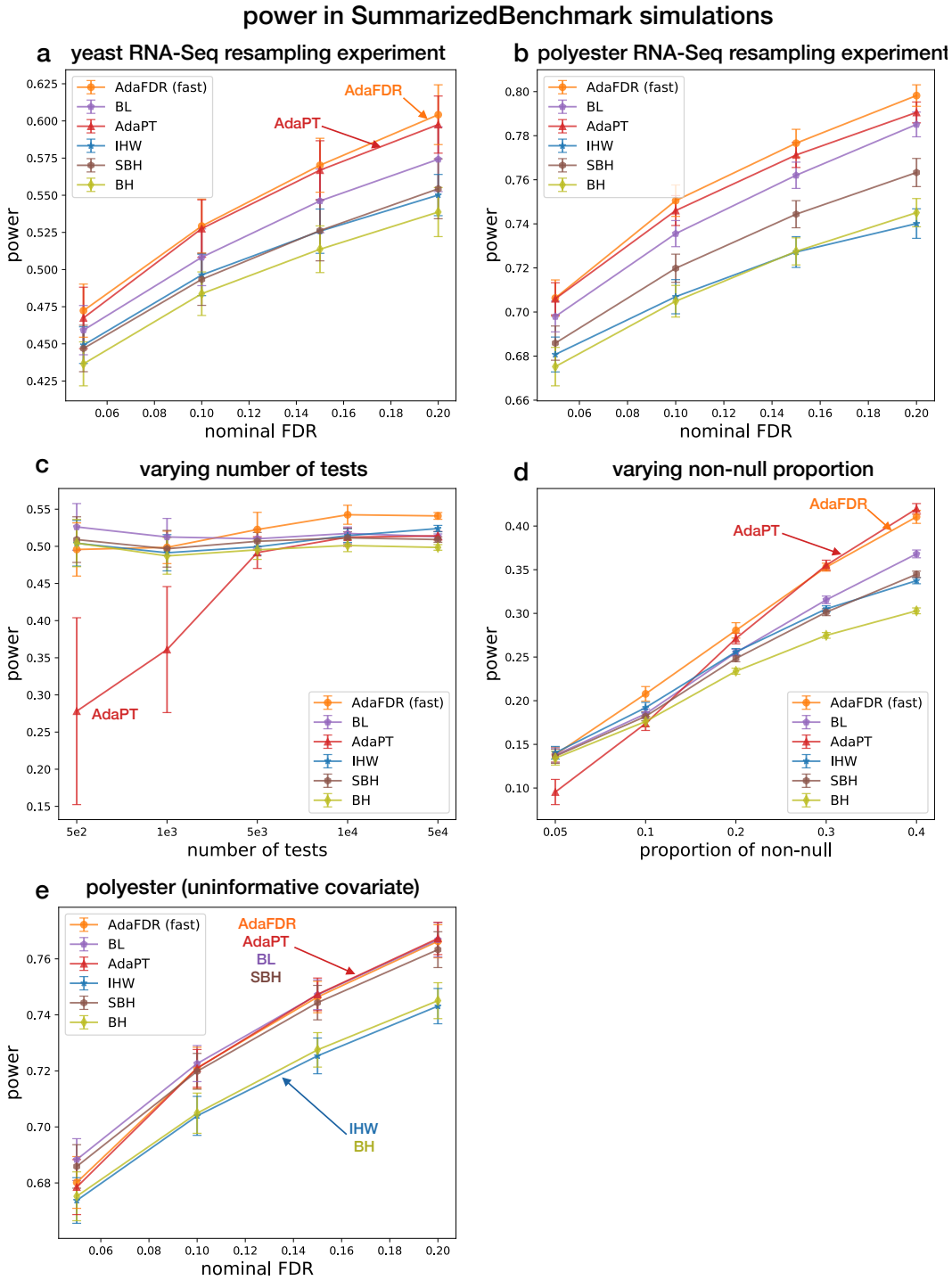
a more simulations for FDP and power



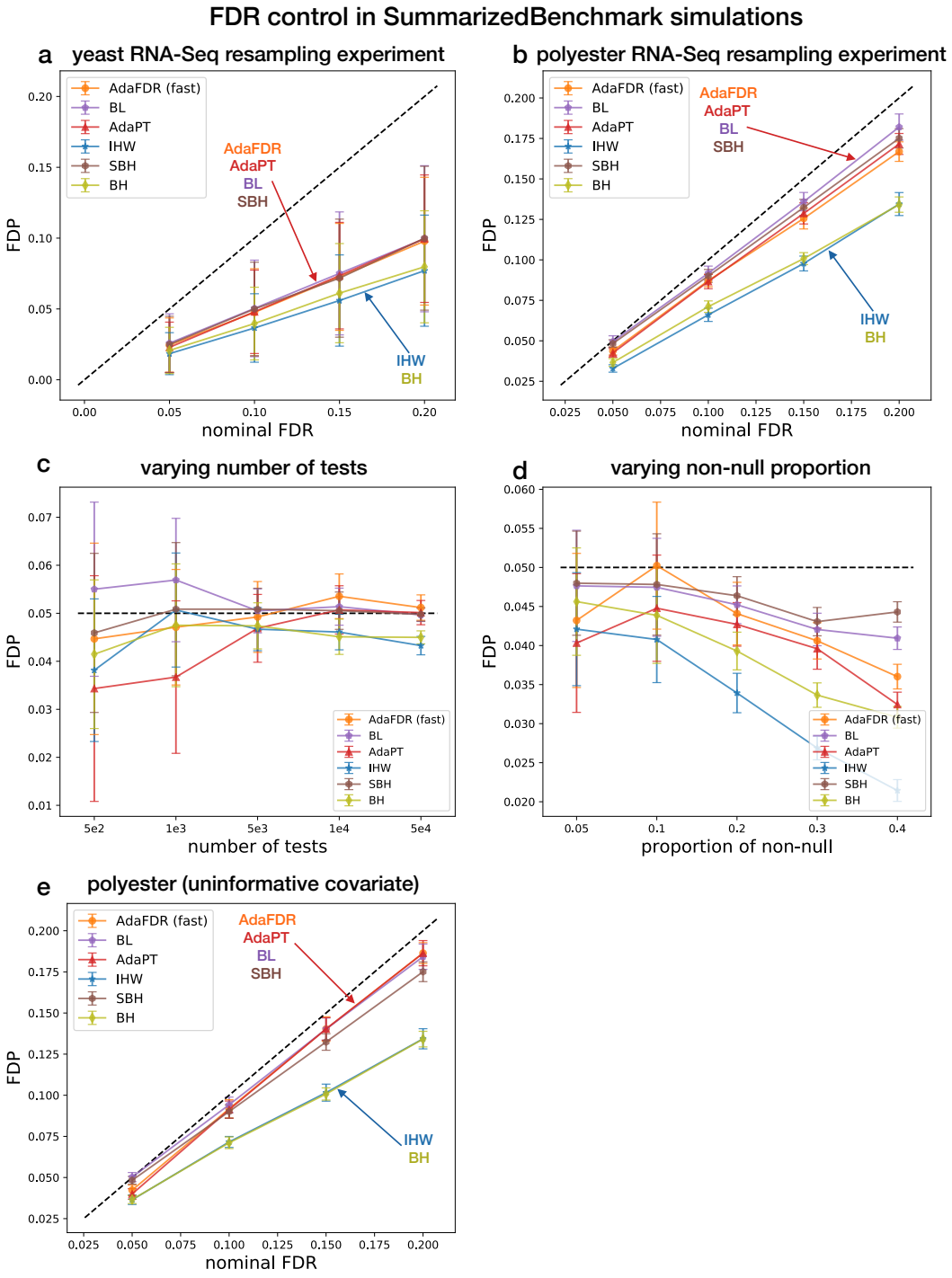
b comparison with NeuralFDR

	NeuralFDR	AdaFDR
2DGM	18844	**20182**
5DGM	18364	**19832**
GTEx_ NeuralFDR	**37195**	37141

Supplementary Figure 3. (a) Additional simulations for FDP and power. Descriptions of the data are in Supplementary SubSection 3.6. (b) Comparison between `NeuralFDR` and `AdaFDR`.



Supplementary Figure 4. Power in five SummarizedBenchmark simulations[18] with the corresponding FDP shown in Supplementary Figure 5. Panels a-d correspond to Figure 3 in[18] while panel e corresponds to the first row of Table S2 in[18]. Performance of an extra method BL[47] is provided. Performance of `AdaFDR` is very similar to `AdaFDR-fast` and is hence omitted to reduce clutter. Ten resamplings were done for RNA-Seq experiments (a,b,e) while twenty were done for others; 95% confidence intervals are provided. Panels a, b are two RNA-Seq spike-in resampling experiments with an informative covariate, panel c contains a simulated data with the number of tests varying from 500 to 50k, while panel d contains a simulated data with the non-null proportion of tests varying from 0.95 to 0.6. In all four experiments, `AdaFDR` and AdaPT have the highest power (with `AdaFDR` being slightly better). We note that AdaPT does not have such high power in the same experiments in[18]. This is probably because we used adapt_gam while adapt_glm is used in[18]; the former has a better performance but takes a longer time to run. Panel e uses the same set of p-values as panel b but with an uninformative covariate. We can see the performance of IHW reduces to BH while others reduce to SBH, a phenomenon also mentioned in[18]. `AdaFDR` maintains high power here indicating that it is not overfit to the uninformative covariate.



Supplementary Figure 5. FDR control in five SummarizedBenchmark simulations[18] with the corresponding power shown in Supplementary Figure 4. The detailed description of the data can also be found in Supplementary Figure 4. Performance of an extra method BL[47] is provided. Performance of `AdaFDR` is very similar to `AdaFDR-fast` and is hence omitted to reduce clutter. All methods control the FDR accurately, except in panel c, where BL slightly exceeds the FDR control when the number of tests is small.

2 Additional Algorithm Information

2.1 Feature preprocessing

We perform feature preprocessing to integrate both numerical covariates and categorical covariates. First for each categorical covariate, the categories are reordered based on the ratio of the alternative probability and the null probability, estimated on the training set using the same method as above. Then quantile normalization is performed for each covariate separately. Note that after this transformation, all covariates will have values between 0 and 1. Also, overfitting is not a concern since the entire proprecessing is done without seeing p-values from the testing set.

2.2 Remark on Theorem 1

Theorem 1 is similar to, but stronger than that for `NeuralFDR`. First, `NeuralFDR` requires the scale factor to be selected from a finite set of L numbers and has an extra multiplicative factor $\sqrt{\log L}$ in the error term ε. In contrast, `AdaFDR` selects the scale factor over all positive numbers and the $\sqrt{\log L}$ term is gone. This is done by using a stochastic process argument instead of the union bound. Second, `NeuralFDR` uses an empirical Bayes model where the tuples $(P_i, \mathbf{x}_i, h_i)$ are generated i.i.d. following some hierarchical model. `AdaFDR`, however, requires a less restrictive assumption made only on the conditional distribution of null p-values, whereas the covariates and alternative p-values can have arbitrary dependence.

2.3 Initialization via EM algorithm

Here we present the EM algorithm that is used to fit the mixture model (2) on a set of N points $\{\mathbf{x}_i\}_{i=1}^N$. Recall that due to quantile normalization, the value of $\mathbf{x}_i$ is within $[0,1]^d$. Therefore, each component in the mixture model is truncated to be within $[0,1]^d$, i.e. truncated GLM or truncated Gaussian. Since we need to use the samples each associated with a sample weight, let us consider the general case where each sample $\mathbf{x}_i$ receives a positive weight $v_i \in \mathbb{R}_+$.

For the sake of convenience, let us reparameterize the parameters to have the standard probability distribution

$$f_{\text{all}}(\mathbf{x}; \mathbf{w}, \mathbf{a}, \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K) = w_0 f_{\text{slope}}(\mathbf{x}; \mathbf{a}) + \sum_{k=1}^K w_k f_{\text{bump}}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k),\ \mathbf{x} \in [0,1]^d, \tag{1}$$

where $\mathbf{w} \in [0,1]^{K+1}$ with $\sum_{k=0}^K w_k = 1$ and

$$f_{\text{slope}}(\mathbf{x}; \mathbf{a}) = \exp \mathbf{a}^{\mathbf{T}}\mathbf{x} \prod_{j=1}^d \frac{a_j}{\exp(a_j) - 1},$$

$$f_{\text{bump}}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) = \prod_{j=1}^d \frac{1}{Z_{kj}\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x_j - \mu_{kj})^2}{2\sigma_{kj}^2}\right),$$

$$\text{for } Z_{kj} = \int_0^1 \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x_j - \mu_{kj})^2}{2\sigma_{kj}^2}\right) dx.$$

It is not hard to see that (1) is equivalent to the mixture threshold (2) up to a scale factor that can be specified by b in (2); knowing one, the parameters for the other can be computed without difficulty.

The EM algorithm can be described as follows. For the initialization, the responsibility $\mathbf{r}_i \in [0,1]^{K+1}, i \in [N]$ for each point $\mathbf{x}_i$ is initialized as

$$\mathbf{r}_i^{\text{init}} = [0.5, \frac{1}{2K}, \frac{1}{2K}, \cdots, \frac{1}{2K}],$$

where the first component corresponds to the slope component and the rest correspond to the K bump components. Then, the algorithm iterates between the E-step and the M-step as follows until convergence:

1. **Expection (E-step)**: For each point $\mathbf{x}_i$, update the responsibility

$$\mathbf{r}_i^{\text{new}} = \frac{1}{f_{\text{all}}(\mathbf{x}_i; \mathbf{w}^{\text{old}}, \mathbf{a}^{\text{old}}, \{\boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\sigma}_k^{\text{old}}\}_{k=1}^K)} [w_0^{\text{old}} f_{\text{slope}}(\mathbf{x}_i; \mathbf{a}^{\text{old}}),$$
$$w_1^{\text{old}} f_{\text{bump}}(\mathbf{x}_i; \boldsymbol{\mu}_1^{\text{old}}, \boldsymbol{\sigma}_1^{\text{old}}), w_2^{\text{old}} f_{\text{bump}}(\mathbf{x}_i; \boldsymbol{\mu}_2^{\text{old}}, \boldsymbol{\sigma}_2^{\text{old}}), \cdots, w_K^{\text{old}} f_{\text{bump}}(\mathbf{x}_i; \boldsymbol{\mu}_K^{\text{old}}, \boldsymbol{\sigma}_K^{\text{old}})].$$

2. **Maximization (M-step)**: Update the component weights $\mathbf{w}^{\text{new}}$ by

$$w_k^{\text{new}} = \frac{\sum_{i=1}^N v_i w_{ik}^{\text{old}}}{\sum_{k=0}^K \sum_{i=1}^N v_i w_{ik}^{\text{old}}},\ k = 0, 1, \cdots, K$$

Update the parameters for the slope component and each of the K bump component:

$$\mathbf{a}^{\text{new}} = \text{ML}_{\text{slope}}(\{\mathbf{x}_i, v_i r_{i0}^{\text{new}}\}_{i=1}^N)$$
$$\boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\sigma}_k^{\text{new}} = \text{ML}_{\text{bump}}(\{\mathbf{x}_i, v_i r_{ik}^{\text{new}}\}_{i=1}^N),\ k \in [K].$$

The ML estimates of slope and bump, i.e. $\text{ML}_{\text{slope}}(\{\mathbf{x}_i, v_i r_{i0}^{\text{new}}\}_{i=1}^N)$ and $\text{ML}_{\text{bump}}(\{\mathbf{x}_i, v_i r_{ik}^{\text{new}}\}_{i=1}^N)$, are described as follows.
ML estimate of the slope. The log likelihood function of a single observation $\mathbf{x}_i$ can be written as

$$l_i(\mathbf{a}) = \log f_{\text{slope}}(\mathbf{x}_i; \mathbf{a}) = \sum_{j=1}^{d} \log\left(\frac{a_j}{\exp(a_j)-1}\right) + \mathbf{a}^T \mathbf{x}_i. \tag{2}$$

Further the weighted average log likelihood function,

$$\bar{l}(\mathbf{a}) = \frac{\sum_{i=1}^N v_i r_{i0}^{\text{new}} l_i(\mathbf{a})}{\sum_{i=1}^N v_i r_{i0}^{\text{new}}} = \sum_{j=1}^{d} \log\left(\frac{a_j}{\exp(a_j)-1}\right) + \frac{\mathbf{a}^T}{\sum_{i=1}^n v_i r_{i0}^{\text{new}}} \sum_{i=1}^N v_i r_{i0}^{\text{new}} \mathbf{x}_i. \tag{3}$$

We add a regularization term $c\|\mathbf{a}\|_2^2$ to encourage small values of $c\|\mathbf{a}\|_2^2$, i.e.

$$\bar{l}(\mathbf{a}) = \sum_{j=1}^{d} \log\left(\frac{a_j}{\exp(a_j)-1}\right) + \frac{\mathbf{a}^T}{\sum_{i=1}^N v_i r_{i0}^{\text{new}}} \sum_{i=1}^N v_i r_{i0}^{\text{new}} \mathbf{x}_i - c\|\mathbf{a}\|_2^2. \tag{4}$$

We found that setting $c = 0.005$ gives a stable result. We solve the ML estimation problem by setting the derivative to be zero. Namely, for the j-th element a_j,

$$\frac{\partial \bar{l}}{\partial a_j} = \frac{1}{a_j} - \frac{e^{a_j}}{e^{a_j}-1} + \frac{1}{\sum_{i=1}^N v_i r_{i0}^{\text{new}}} \sum_{i=1}^N v_i r_{i0}^{\text{new}} x_{ij} - 2ca_j = 0. \tag{5}$$

Rearranging terms on both sides we have that the ML estimate $\hat{a}_j$ satisfies

$$\frac{e^{\hat{a}_j}}{e^{\hat{a}_j}-1} - \frac{1}{\hat{a}_j} + 2c\hat{a}_j = \frac{1}{\sum_{i=1}^N v_i r_{i0}^{\text{new}}} \sum_{i=1}^N v_i r_{i0}^{\text{new}} x_{ij}. \tag{6}$$

Since the left-hand-side term is monotonic in $\hat{a}_j$, the ML solution $\hat{a}_j$ can be computed via binary search.
ML estimate of the k-th bump. Since the density function can be factorized as a product of different dimensions, the ML estimation can be done for each dimension separately. Now consider observation $\mathbf{x}_i$. The log likelihood function corresponding to dimension j can be written as

$$l_{ij}(\mu_{kj}, \sigma_{kj}) = -\log Z_{kj} - \frac{1}{2}\log(2\pi) - \log\sigma_{kj} - \frac{1}{2\sigma_{kj}^2}(x_{ij} - \mu_{kj})^2. \tag{7}$$

Then the weighted average log likelihood function for dimension j can be written as

$$\bar{l}_j(\mu_{kj}, \sigma_{kj}) = \frac{\sum_{i=1}^N v_i r_{ik}^{\text{new}} l_{ij}(\mu_{kj}, \sigma_{kj})}{\sum_{i=1}^N v_i r_{ik}^{\text{new}}} \tag{8}$$

$$= -\log Z_{kj} - \frac{1}{2}\log(2\pi) - \log\sigma_{kj} - \frac{1}{2\sigma_{kj}^2 \sum_{i=1}^n v_i r_{ik}^{\text{new}}} \sum_{i=1}^N v_i r_{ik}^{\text{new}} (x_{ij} - \mu_{kj})^2. \tag{9}$$

Since $\bar{l}_j(\mu_{kj}, \sigma_{kj})$ is convex, we compute the ML estimation $\hat{\mu}_{kj}$ and $\hat{\sigma}_{kj}$ via gradient descent, where the derivatives are given as follows.

$$\frac{\partial \bar{l}_j}{\partial \mu_{kj}} = -\frac{1}{Z_{kj}}\frac{\partial Z_{kj}}{\partial \mu_{kj}} + \frac{1}{\sigma_{kj}^2 \sum_{i=1}^N v_i r_{ik}^{\text{new}}} \sum_{i=1}^N v_i r_{ik}^{\text{new}} (x_{ij} - \mu_{kj}) \tag{10}$$

$$\frac{\partial \bar{l}_j}{\partial \sigma_{kj}} = -\frac{1}{Z_{kj}}\frac{\partial Z_{kj}}{\partial \sigma_{kj}} - \frac{1}{\sigma_{kj}} + \frac{1}{\sigma_{kj}^3 \sum_{i=1}^N v_i r_{ik}^{\text{new}}} \sum_{i=1}^N v_i r_{ik}^{\text{new}} (x_{ij} - \mu_{kj})^2, \tag{11}$$

where the derivatives with respect to Z_{kj} are

$$\frac{\partial Z_{kj}}{\partial \mu_{kj}} = \frac{1}{\sigma_{kj}}[\phi(\beta_1) - \phi(\beta_2)], \quad \frac{\partial Z_{kj}}{\partial \sigma_{kj}} = \frac{1}{\sigma_{kj}}[\beta_1\phi(\beta_1) - \beta_2\phi(\beta_2)], \tag{12}$$

for $\beta_1 = \frac{-\mu_{kj}}{\sigma_{kj}}$, $\beta_2 = \frac{1-\mu_{kj}}{\sigma_{kj}}$ and $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$.

2.4 Implementation of other methods

1. AdaPT: adapt_gam is used with a 5-degree spline for each dimension. This choice is based on a discussion with the authors of AdaPT.

2. IHW: The covaraites are first clustered into 20 clusters using Kmeans clustering. Then IHW is run with the default setting and the cluster labels as the univariate covariate. This automatically incorporates the multi-dimensional case. For the univariate case, this does not change the result much as compared to directly running IHW. For example, for the airway data, directly running IHW gives 4873 discoveries while Kmeans+IHW gives 4862 discoveries.

3. BL: First the null distribution $\pi_0(\mathbf{x})$ is estimated using lm_pi0 with 5 degrees of freedom. Then BH is used with p-values weighted by $1/\pi_0(\mathbf{x}_i)$. This is the same as the usage in[18].

3 Data

3.1 eQTL study

GTEx For eQTL study, we used Genotype-Tissue Expression (GTEx) dataset[7]. This dataset aims at characterizing variation in gene expression levels across individuals and diverse tissues of the human body. We used the V7 release of GTEx analysis data (dbGaP Accession phs000424.v7.p2). The dataset contains 11688 samples, and in total there are 53 tissues from 714 donors (44 of them with sample size >70 are used in the GTEx paper). We filtered the tissues based on the following criteria. First, the tissue needs to have eQTL analysis, where the number of samples with genotype is greater than 70. Second, we set the number of samples threshold to be 100 in order to make the p-values more reliable. Third, we would like the tissue to have a corresponding roadmap[8] cell type, so that we can leverage the cell-specific chromatin state data from roadmap. After filtering, we were left with 17 cell types. The meta-information of the filtered GTEx dataset is listed in Table 1.

Table 1. Information for selected GTEx tissue types

Tissue name	Sample size	Roadmap cell type	Number of hypothesis
Adipose Subcutaneous	298	E063	1.72E+08
Adipose Visceral Omentum	185	E063	1.73E+08
Artery Aorta	197	E065	1.66E+08
Breast Mammary Tissue	183	E027	1.80E+08
Cells EBV-transformed lymphocytes	114	E116	1.60E+08
Colon Sigmoid	124	E106	1.70E+08
Colon Transverse	169	E075, E076	1.77E+08
Esophagus Gastroesophageal Junction	127	E079	1.67E+08
Esophagus Mucosa	241	E079	1.67E+08
Esophagus Muscularis	218	E079	1.66E+08
Heart Atrial Appendage	159	E104	1.61E+08
Heart Left Ventricle	190	E095	1.50E+08
Lung	278	E096	1.82E+08
Muscle Skeletal	361	E107, E108	1.47E+08
Pancreas	149	E098	1.59E+08
Stomach	170	E110, E111	1.69E+08
Whole Blood	338	E062	1.45E+08

In this filtered dataset, each hypothesis is a gene-variant pair. Nominal P values for each gene-variant pair were estimated using a two-tailed t-test. Each gene-variant is associated with 4 or 5 covariates listed below:

- **gene expression** We obtained the median gene expression from the gene in gene-variant pair and used as a feature.
- **alternative allele frequency** We mapped each SNP to the dbSNP database[51]. We took the alternative allele frequency as a feature. If there were multiple alternative alleles, we took the smallest one. For the SNPs we cannot find a mapping, this feature is imputed with mean alternative allele frequency.
- **TSS distance** The distance from the SNP to the transcription starting site is used as a feature. It is defined as $pos_{SNP} - pos_{TSS}$.
- **cell-specific chromatin state** We took the position of the SNP and mapped it to roadmap database. Each SNP falls into the 15-state chromatin model. This state is used as a categorical feature.
- **p-value from another tissue (optional)** Optionally, we used the P value from another tissue as a covariate. If we cannot find the same gene-variant pair in another tissue, we impute with the mean P value. This covariate is only used for "AdaFDR (aug)" and "AdaFDR (ctrl)" experiments.

MuTHER In the Multiple Tissus Human Expression Resource project[46], samples from 850 individuals were collected and 3 tissues, namely adipose, LCL, and skin, were studied. We used only the data for the adipose tissue, where a nominal p-value is provided for each SNP-gene pair.

3.2 RNA-Seq data

We used three RNA-Seq datasets to validate our algorithm. The first one `bottomly`[15] is an RNA-Seq dataset used to detect differential gene expression between mouse strains. We used the same data preprocessing pipeline as in IHW[12]. p-values were calculated using `DESeq2`, and the mean of normalized counts for each gene were chosen to be the covariate. The second dataset `airway`[14] is an RNA-Seq dataset used to identify the differentially expressed genes in airway smooth muscle cell lines in response to dexamethasone. The dataset is processed with the same pipeline as `bottomly`. The thrid dataset `Pasilla`[52] is an RNA-Seq dataset for detecting genes that are differentially expressed between the normal and Pasilla-knockdown conditions. This dataset is available in `Pasilla` package and it is analyzed in the vignette of `genefilter` package using independent filtering method. The p-values were generated using `DESeq` package and the logarithm of normalized count were used as the covariate. All the preprocessing steps can be reproduced using vignettes of R package `IHWPaper`[53].

3.3 Microbiome data

The two microbiome experiments are from the benchmark paper[18].

3.4 Proteomics data

The proteomics data is from the IHW paper[12].

3.5 fMRI data

The two fMRI data are from the fMRI paper[20].

3.6 Simulated data

Data 1. Simulated data with one covariate. The covariate $\mathbf{x}_i \sim \text{Unif}[0,1]$ and the probability of being an alternative hypothesis given the covariate $\mathbb{P}(h_i = 1|\mathbf{x}_i)$ is defined using the mixture model (1) as

$$\mathbb{P}(h_i = 1|\mathbf{x}_i) = 0.1 f_{\text{all}}(\mathbf{x}; \mathbf{w} = [0.5, 0.25, 0.25], a = 0.5, \mu_1 = 0.25, \mu_2 = 0.75, \sigma_1 = \sigma_2 = 0.05).$$

The null p-values are generated i.i.d. from Unif$[0,1]$ while the alternative p-values are generated i.i.d. from $P_i \sim \text{Beta}(\alpha = 0.3, \beta = 4)$. The number of hypotheses is 20000 and 10 datasets are generated with different random seeds.

Data 2. Simulated data with two covariates The covariate $\mathbf{x}_i \sim \text{Unif}[0,1]$ and the probability of being an alternative hypothesis given the covariate $\mathbb{P}(h_i = 1|\mathbf{x}_i)$ is defined using the mixture model (1) as

$$\begin{aligned}\mathbb{P}(h_i = 1|\mathbf{x}_i) =& 0.1 f_{\text{all}}(\mathbf{x}; \mathbf{w} = [0.5, 0.25, 0.25], \mathbf{a} = [0.5, 0.5], \\ & \boldsymbol{\mu}_1 = [0.25, 0.25], \boldsymbol{\mu}_2 = [0.75, 0.75], \boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = [0.1, 0.1]).\end{aligned}$$

The null p-values are generated i.i.d. from Unif$[0,1]$ while the alternative p-values are generated i.i.d. from $P_i \sim \text{Beta}(\alpha = 0.3, \beta = 4)$. The number of hypotheses is 20000 and 10 datasets are generated with different random seeds.

Data 3. Simulated data with ten covariates First a simulated data with two covariates is generated (data 2). Then, another 8 noisy dimensions are added to the covariates with each entry drawn i.i.d. from Unif$[0,1]$. The number of hypotheses is 20000 and 10 datasets are generated with different random seeds.

Data 4. Simulated data with weakly-dependent p-values The covariate $\mathbf{x}_i \sim \text{Unif}[0,1]$ and the probability of being an alternative hypothesis given the covariate $\mathbb{P}(h_i = 1|\mathbf{x}_i)$ is generated same as the simulated data with one covariate (data 1). The p-values are converted to z-scores via $p = 1 - \Phi(z)$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Every 10 consecutive null z-scores are generated from $\mathcal{N}(0, \Sigma)$, while every 10 consecutive alternative z-scores are generated from $\mathcal{N}(2, \Sigma)$, with the symmetric covariance matrix whose upper triangular part is specified as

$$\begin{aligned}&\Sigma_{ii} = 1, \\ &\Sigma_{ij} = 0.25, i < j \leq 4, \\ &\Sigma_{ij} = -0.25, j > 4.\end{aligned}$$

We note instead of 0.25, the value 0.4 is used in the original paper (Section 3.2,[4]). However such choice makes the covariance matrix not positive semi-definite. We decrease the value until the matrix becomes positive semi-definite. The number of hypotheses is 20000 and 10 datasets are generated with different random seeds.

Data 5. Simulated data with strongly-dependent p-values The setting is the same as the weakly dependent data (data 4) except the generation of z-scores. Here, every 5 consecutive null z-scores are generated from $\mathcal{N}(0, I)$, while every 5 consecutive alternative z-scores are generated from $\mathcal{N}(2, I)$. This perfect correlation means to model the linkage disequilibrium (LD) that

frequently occurs in SNPs. Due to the reduction of the inherent multiplicity, the number of hypotheses is increased to 50000. 10 datasets are generated with different random seeds.

Data 6. Simulated data used in AdaPT The same data for Figure 6a in[24] is used where the number of hypotheses is 2500. 10 datasets are generated with different random seeds.

Data 7. Simulated data used in IHW The data is generated according to Supplementary Section 4.2.2 in[12] where the number of hypotheses is 20000. While the original paper varies the effect size from 1 to 2.5 (the shift of z-scores for alternative p-values), here we only use a fixed effect size 2. 10 datasets are generated with different random seeds.

4 Proofs and Auxiliary Lemmas

4.1 Proof of Theorem 1

Proof. To avoid ambiguity, we make a few clarifications before the proof. First, the entire analysis is done while conditioning on the hypothesis splitting, all covariates $\{\mathbf{x}_i\}_{i\in[N]}$, the type of hypotheses $\{h_i\}_{i\in[N]}$, and the alternative p-values $\{P_i\}_{i\in\mathscr{H}_1}$, hence allowing arbitrary dependencies of them. Here we note that the reason for splitting the hypotheses at random is to attain good power. The randomness of the analysis comes from the null p-values, which are assumed to be i.i.d. uniformly distributed for convenience. A discussion on extending to the case where the null p-values, conditional on the covariates, are independently distributed and stochastically greater than the uniform distribution is provided at the end.

We also clarify a few notations. We use $t^*_{\mathscr{D}_1}$ to denote the threshold which is learned on fold 1 and will be applied on fold 2. γ_1^* denotes the scale factor of fold 1. For the testing-related quantities, we use subscript "1" to denote those evaluated on fold 1, including the number of discoveries $\mathrm{D}_1(\gamma_1^* t^*_{\mathscr{D}_2})$, the number of false discoveries $\mathrm{FD}_1(\gamma_1^* t^*_{\mathscr{D}_2})$, the mirror-estimated number of false discoveries $\widehat{\mathrm{FD}}_1(\gamma_1^* t^*_{\mathscr{D}_2})$ and the mirror-estimated false discovery proportion $\widehat{\mathrm{FDP}}_1(\gamma_1^* t^*_{\mathscr{D}_2})$. Note that here $t^*_{\mathscr{D}_2}$ is the threshold that is learned on fold 2 and then applied on fold 1. The term inside the bracket, $(\gamma_1^* t^*_{\mathscr{D}_2})$, may be omitted when there is no concern of being ambiguous. Quantities for fold 2 are defined in a similar fashion. Now we preceed to the proof.

Step 1: show that in order to prove the result, it suffices to show that

$$\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha) \leq \frac{\delta}{2}. \tag{13}$$

Indeed, if (13) it true, then by symmetry $\mathbb{P}(\mathrm{FDP}_1 \geq (1+\varepsilon)\alpha) \leq \frac{\delta}{2}$. Further by the union bound, with probability (w.p.) at least 1-δ,

$$\mathrm{FDP}_2 < (1+\varepsilon)\alpha, \text{ and } \mathrm{FDP}_1 < (1+\varepsilon)\alpha.$$

This further implies that w.p. at least 1-δ, the FDP on the whole dataset

$$\mathrm{FDP} = \frac{\mathrm{FD}_1+\mathrm{FD}_2}{\mathrm{D}_1+\mathrm{D}_2} \leq \left(\frac{\mathrm{FD}_1}{\mathrm{D}_1}\right) \vee \left(\frac{\mathrm{FD}_2}{\mathrm{D}_2}\right) = \mathrm{FDP}_1 \vee \mathrm{FDP}_2 < (1+\varepsilon)\alpha,$$

which gives the desired result. Hence in the rest of the proof, we denote effort to proving (13). Also, since we are only to deal with fold 2, we drop the subscript $\mathscr{D}_1$ for threshold learned on fold 1 to have $t^* \overset{def}{=} t^*_{\mathscr{D}_1}$.

Step 2: convert the probability $\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha)$ to some analyzable stochastic process.

Let $\mathscr{E}_0$ denote the set of random variables that we wish to condition on, including hypothesis splitting, all covariates $\{\mathbf{x}_i\}_{i\in[N]}$, the type of hypotheses $\{h_i\}_{i\in[N]}$, and the alternative p-values $\{P_i\}_{i\in\mathscr{H}_1}$. Let us consider the conditional version of (13):

$$\begin{aligned}\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha|\mathscr{E}_0) &= \mathbb{P}\left(\frac{\mathrm{FD}_2}{\mathrm{D}_2 \vee 1} \geq (1+\varepsilon)\alpha \middle| \mathscr{E}_0\right) \\ &= \mathbb{P}\left(\frac{\mathrm{FD}_2}{\mathrm{D}_2 \vee 1}\frac{1}{\alpha} - 1 \geq \varepsilon \middle| \mathscr{E}_0\right).\end{aligned}$$

Let $\eta \overset{def}{=} \left(\frac{\mathrm{FD}_2}{\mathrm{D}_2\vee 1}\frac{1}{\alpha} - 1\right)$. Recall that FD_2 and D_2 correspond to the best rescaled threshold on fold 2 $\gamma_2^* t^*$ and the best scale factor γ_2^* is selected from the set $\left\{\gamma : \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1} \leq \alpha, \mathrm{D}_2(\gamma t^*) \geq c_0 N\right\} \cup \{0\}$. Then η can be upper bounded by

$$\begin{aligned}\eta &= \frac{\mathrm{FD}_2(\gamma_2^* t^*)}{\mathrm{D}_2(\gamma_2^* t^*)\vee 1}\frac{1}{\alpha} - 1 \\ &\leq \sup_{\gamma \in \left\{\gamma:\ \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1} \leq \alpha,\ \mathrm{D}_2(\gamma t^*) \geq c_0 N\right\} \cup \{0\}} \left(\frac{\mathrm{FD}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1}\frac{1}{\alpha} - 1\right) \\ &\leq \sup_{\gamma \geq 0}\left(\frac{\mathrm{FD}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1}\frac{1}{\alpha} - 1\right)\mathbb{I}_{\left\{\gamma:\ \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1} \leq \alpha,\ \mathrm{D}_2(\gamma t^*) \geq c_0 N\right\}}.\end{aligned}$$

Furthermore, since the indicator function is one only when $\frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*)\vee 1} \leq \alpha$, which can also be written as $\frac{1}{\alpha} \leq \frac{\mathrm{D}_2(\gamma t^*)\vee 1}{\widehat{\mathrm{FD}}_2(\gamma t^*)}$ with the

convention that $\frac{x}{0} = \infty$ for any $x > 0$, we further have

$$\eta \leq \sup_{\gamma \geq 0} \left[\frac{\mathrm{FD}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*) \vee 1} \left(\frac{1}{\alpha} \wedge \frac{\mathrm{D}_2(\gamma t^*) \vee 1}{\widehat{\mathrm{FD}}_2(\gamma t^*)} \right) - 1 \right] \mathbb{I}_{\left\{ \gamma:\ \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*) \vee 1} \leq \alpha,\ \mathrm{D}_2(\gamma t^*) \geq c_0 N \right\}}$$
$$= \sup_{\gamma \geq 0} \left(\frac{\mathrm{FD}_2(\gamma t^*)}{(\alpha \mathrm{D}_2(\gamma t^*)) \vee \alpha \vee \widehat{\mathrm{FD}}_2(\gamma t^*)} - 1 \right) \mathbb{I}_{\left\{ \gamma:\ \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*) \vee 1} \leq \alpha,\ \mathrm{D}_2(\gamma t^*) \geq c_0 N \right\}}.$$

Again since indicator function is one only when $\mathrm{D}_2(\gamma t^*) \geq c_0 N$,

$$\eta \leq \sup_{\gamma \geq 0} \left(\frac{\mathrm{FD}_2(\gamma t^*)}{(\alpha c_0 N) \vee \widehat{\mathrm{FD}}_2(\gamma t^*)} - 1 \right) \mathbb{I}_{\left\{ \gamma:\ \frac{\widehat{\mathrm{FD}}_2(\gamma t^*)}{\mathrm{D}_2(\gamma t^*) \vee 1} \leq \alpha,\ \mathrm{D}_2(\gamma t^*) \geq c_0 N \right\}},$$
$$\leq 0 \vee \sup_{\gamma \geq 0} \left(\frac{\mathrm{FD}_2(\gamma t^*)}{(\alpha c_0 N) \vee \widehat{\mathrm{FD}}_2(\gamma t^*)} - 1 \right).$$

Furthermore with the notation $t_i^* \overset{def}{=} t^*(\mathbf{x}_i)$ where we recall that we have defined $t^* = t^*_{\mathscr{D}_1}$ before,

$$\eta \leq 0 \vee \sup_{\gamma \geq 0} \left(\frac{\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \leq \gamma t_i^*\}}}{(\alpha c_0 N) \vee \left(\sum_{i \in \mathscr{D}_2} \mathbb{I}_{\{P_i \geq 1 - \gamma t_i^*\}} \right)} - 1 \right)$$
$$\leq 0 \vee \sup_{\gamma \geq 0} \left(\frac{\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \leq \gamma t_i^*\}}}{(\alpha c_0 N) \vee \left(\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \geq 1 - \gamma t_i^*\}} \right)} - 1 \right).$$

Finally, we can complete the conversion by noting that

$$\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha | \mathscr{E}_0) = \mathbb{P}(\eta \geq \varepsilon | \mathscr{E}_0) \tag{14}$$
$$\leq \mathbb{P} \left[\sup_{\gamma \geq 0} \left(\frac{\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \leq \gamma t_i^*\}}}{(\alpha c_0 N) \vee \left(\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \geq 1 - \gamma t_i^*\}} \right)} - 1 \right) \geq \varepsilon \middle| \mathscr{E}_0 \right]. \tag{15}$$

Here, the first term in (15), i.e. $\frac{\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \leq \gamma t_i^*\}}}{(\alpha c_0 N) \vee \left(\sum_{i \in \mathscr{D}_2 \cap \mathscr{H}_0} \mathbb{I}_{\{P_i \geq 1 - \gamma t_i^*\}} \right)}$, can be understood as a stochastic process that as γ grows from 0 to infinity, new elements are added to the numerator and the denominator with equal probability. Hence this term should always be close to 1. We next proceed to prove the result following this intuition.

Step 3: Upper bound the probability of (15).

We note that the p-values involved in (15) are all null p-values from fold 2. Hence, they are i.i.d. uniformly distributed conditional on $\mathscr{E}_0$. Let $\mathscr{H}_{0,2} \overset{def}{=} \mathscr{D}_2 \cap \mathscr{H}_0$. For any $i \in \mathscr{H}_{0,2}, \gamma > 0$, define the random variables

$$B_{i,\gamma} = \mathbb{I}_{\{P_i \leq \gamma t_i^* \text{ or } P_i \geq 1 - \gamma t_i^*\}}, \quad R_i = \mathbb{I}_{\{P_i \leq 0.5\}} - \mathbb{I}_{\{P_i > 0.5\}}. \tag{16}$$

Since $\forall i \in \mathscr{H}_{0,2}$, $P_i | \mathscr{E}_0 \sim \mathrm{Unif}[0,1]$, we have $B_{i,\gamma} | \mathscr{E}_0 \sim \mathrm{Bern}(2\gamma t_i^*)$ and $R_i | \mathscr{E}_0$ are i.i.d. Rademacher random variables. In addition, it is easy to verify that $B_{i,\gamma}$ is independent of R_i and

$$\mathbb{I}_{\{P_i \leq \gamma t_i^*\}} = B_{i,\gamma} \mathbb{I}_{\{R_i = 1\}}, \quad \mathbb{I}_{\{P_i \geq 1 - \gamma t_i^*\}} = B_{i,\gamma} \mathbb{I}_{\{R_i = -1\}}.$$

Hence (15) can be written in terms of $B_{i,\gamma}$'s and R_i's as

$$\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha | \mathscr{E}_0) \leq \mathbb{P} \left[\sup_{\gamma \geq 0} \left(\frac{\sum_{i \in \mathscr{H}_{0,2}} B_{i,\gamma} \mathbb{I}_{\{R_i = 1\}}}{(\alpha c_0 N) \vee \sum_{i \in \mathscr{H}_{0,2}} B_{i,\gamma} \mathbb{I}_{\{R_i = -1\}}} - 1 \right) \geq \varepsilon \middle| \mathscr{E}_0 \right]$$
$$\leq \mathbb{P} \left[\sup_{\gamma \geq 0} \left(\frac{\sum_{i \in \mathscr{H}_{0,2}} B_{i,\gamma} R_i}{(\alpha c_0 N) \vee \sum_{i \in \mathscr{H}_{0,2}} B_{i,\gamma} \mathbb{I}_{\{R_i = -1\}}} \right) \geq \varepsilon \middle| \mathscr{E}_0 \right].$$

Furthermore, let $\gamma_0 = \frac{\alpha c_0 N}{\sum_{i\in\mathscr{H}_{0,2}} t_i^*}$. Divide the set of γ in the sup from $[0,\infty)$ into $[0,\gamma_0]$ and (γ_0,∞), and apply union bound to have

$$\begin{aligned}
&\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha|\mathscr{E}_0) \\
&\leq \mathbb{P}\left[\sup_{0\leq\gamma\leq\gamma_0}\left(\frac{\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}R_i}{(\alpha c_0 N)\vee\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}\mathbb{I}_{\{R_i=-1\}}}\right)\geq\varepsilon\middle|\mathscr{E}_0\right] \\
&\quad+\mathbb{P}\left[\sup_{\gamma>\gamma_0}\left(\frac{\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}R_i}{(\alpha c_0 N)\vee\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}\mathbb{I}_{\{R_i=-1\}}}\right)\geq\varepsilon\middle|\mathscr{E}_0\right] \\
&\leq \mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}R_i}{\alpha c_0 N}\geq\varepsilon\middle|\mathscr{E}_0\right)+\mathbb{P}\left(\sup_{\gamma>\gamma_0}\frac{\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}R_i}{\sum_{i\in\mathscr{H}_{0,2}} B_{i,\gamma}\mathbb{I}_{\{R_i=-1\}}}\geq\varepsilon\middle|\mathscr{E}_0\right).
\end{aligned}$$

Define the random set $\mathscr{B}_\gamma = \{i : i\in\mathscr{H}_{0,2}, B_{i,\gamma}=1\}$. We note that the sequence of sets $\{\mathscr{B}_\gamma\}_{\gamma\geq 0}$ is monotonic in the sense that as γ grows, more elements are incorporated into $\mathscr{B}_\gamma$. With this definition, the above inequality can be further written as

$$\mathbb{P}(\mathrm{FDP}_2 \geq (1+\varepsilon)\alpha|\mathscr{E}_0) \tag{17}$$

$$\leq \mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\alpha c_0 N}\geq\varepsilon\middle|\mathscr{E}_0\right)+\mathbb{P}\left(\sup_{\gamma>\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma}\mathbb{I}_{\{R_i=-1\}}}\geq\varepsilon\middle|\mathscr{E}_0\right). \tag{18}$$

Next we upper bound the two terms in (18) respectively. Here let us use $m \overset{def}{=} \alpha c_0 N$ for simplicity.
The first term in (18): For some $m_0 > 2m$ to be specified later, by the law of total probability,

$$\text{first term in (18)} = \mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}\geq\varepsilon\middle|\mathscr{E}_0\right) \tag{19}$$

$$\leq \mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}\geq\varepsilon, |\mathscr{B}_{\gamma_0}|\leq m_0\middle|\mathscr{E}_0\right)+\mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}\geq\varepsilon, |\mathscr{B}_{\gamma_0}|> m_0\middle|\mathscr{E}_0\right) \tag{20}$$

$$\leq \mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}\geq\varepsilon\middle||\mathscr{B}_{\gamma_0}|\leq m_0,\mathscr{E}_0\right)+\mathbb{P}\left(|\mathscr{B}_{\gamma_0}|> m_0|\mathscr{E}_0\right). \tag{21}$$

The two terms in (21) are upper bounded separately. Consider the first term. Recall that $\{\mathscr{B}_\gamma\}_{\gamma\geq 0}$ is a random sequence of monotonic sets; let $\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}$ denote any of its realization. Then since taking expectation over all possible $\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}$ s.t. $|\tilde{\mathscr{B}}_{\gamma_0}|\leq m_0$ is no greater than taking the sup of them,

$$\text{first term in (21)} \leq \sup_{\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}\ s.t.\ |\tilde{\mathscr{B}}_{\gamma_0}|\leq m_0}\mathbb{P}\left(\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}\geq\varepsilon\middle|\{\mathscr{B}_\gamma\}_{\gamma\geq 0}=\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0},\mathscr{E}_0\right).$$

Consider the term inside the probability, i.e. $\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}$, where due to conditioning $\{\mathscr{B}_\gamma\}_{\gamma\geq 0}=\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}$. Recall that the sequence $\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}$ is monotonic that as γ grows more elements are incorporated into the set but no element is removed from the set. Also up to the point $\gamma=\gamma_0$ there are altogether $|\tilde{\mathscr{B}}_{\gamma_0}|$ elements. Then the sup is equivalent to being evaluated over a sequence of $|\tilde{\mathscr{B}}_{\gamma_0}|+1$ monotonic sets, i.e. $\sup_{0\leq\gamma\leq\gamma_0}\frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m}$ is equal to $\sup_{0\leq k\leq|\tilde{\mathscr{B}}_{\gamma_0}|}\frac{\sum_{i\in[k]}\tilde{R}_i}{m}$ in distribution, where $\tilde{R}_1,\tilde{R}_2,\cdots$ is a sequence of i.i.d. Rademacher random variables independent of everything else. Therefore,

$$\begin{aligned}
\text{first term in (21)} &\leq \sup_{\{\tilde{\mathscr{B}}_\gamma\}_{\gamma\geq 0}\ s.t.\ |\tilde{\mathscr{B}}_{\gamma_0}|\leq m_0}\mathbb{P}\left(\max_{0\leq k\leq|\tilde{\mathscr{B}}_{\gamma_0}|}\frac{\sum_{i\in[k]}\tilde{R}_i}{m}\geq\varepsilon\right) \\
&= \mathbb{P}\left(\max_{1\leq k\leq m_0}\frac{\sum_{i\in[k]}\tilde{R}_i}{m}\geq\varepsilon\right)\leq 2e^{-\frac{m^2\varepsilon^2}{2m_0}},
\end{aligned}$$

where the last inequality is due to Lemma 1.

Now consider the second term in (21). Recall that $\mathbb{E}[|\mathscr{B}_{\gamma_0}|] = \sum_{i\in\mathscr{H}_{0,2}} 2\gamma_0 t_i = 2m$ by the definition of γ_0. By Lemma 2,

$$\text{second term in (21)} = \mathbb{P}\left[|\mathscr{B}_{\gamma_0}| > m_0 | \mathscr{E}_0\right] \le e^{-\frac{\frac{1}{2}(m_0-2m)^2}{2m+\frac{1}{3}(m_0-2m)}}. \tag{22}$$

By setting $m_0 = 3m$, we have

$$\text{first term in (18)} = \mathbb{P}\left(\sup_{0\le\gamma\le\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{m} \ge \varepsilon \middle| \mathscr{E}_0\right) \le 2e^{-\frac{m\varepsilon^2}{6}} + e^{-\frac{3m}{14}}. \tag{23}$$

The second term in (18): For some $m_1 \le 2m$ to be specified later, by the law of total probability,

$$\text{second term in (18)} = \mathbb{P}\left(\sup_{\gamma>\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma} \mathbb{I}_{\{R_i=-1\}}} \ge \varepsilon \middle| \mathscr{E}_0\right) \tag{24}$$

$$= \mathbb{P}\left(\sup_{\gamma>\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma} \mathbb{I}_{\{R_i=-1\}}} \ge \varepsilon, |\mathscr{B}_{\gamma_0}| \ge m_1 \middle| \mathscr{E}_0\right) + \mathbb{P}\left(\sup_{\gamma>\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma} \mathbb{I}_{\{R_i=-1\}}} \ge \varepsilon, |\mathscr{B}_{\gamma_0}| < m_1 \middle| \mathscr{E}_0\right) \tag{25}$$

$$\le \mathbb{P}\left(\sup_{\gamma>\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma} \mathbb{I}_{\{R_i=-1\}}} \ge \varepsilon \middle| |\mathscr{B}_{\gamma_0}| \ge m_1, \mathscr{E}_0\right) + \mathbb{P}\left(|\mathscr{B}_{\gamma_0}| < m_1 | \mathscr{E}_0\right). \tag{26}$$

Using the same argument for analyzing the first term in (21),

$$\text{first term in (26)} \le \mathbb{P}\left(\sup_{k\ge m_1} \frac{\sum_{i\in[k]} \tilde{R}_i}{\sum_{i\in[k]} \mathbb{I}_{\{\tilde{R}_i=-1\}}} \ge \varepsilon\right) \le \frac{2e^{-\frac{m_1\varepsilon^2}{4(\varepsilon+2)^2}}}{1-2e^{-\frac{m_1\varepsilon^2}{4(\varepsilon+2)^2}}},$$

where we recall that $\tilde{R}_1, \tilde{R}_2, \cdots$ is a sequence of i.i.d. Rademacher random variables indepedent of everything else, and the second inequality is due to Lemma 1.

Similar to (22), by Lemma 2,

$$\text{second term in (26)} = \mathbb{P}\left(|\mathscr{B}_{\gamma_0}| < m_1\right) \le e^{-\frac{\frac{1}{2}(2m-m_1)^2}{2m+\frac{1}{3}(2m-m_1)}}.$$

By setting $m_1 = m$, we have

$$\text{second term in (18)} = \mathbb{P}\left(\sup_{\gamma>\gamma_0} \frac{\sum_{i\in\mathscr{B}_\gamma} R_i}{\sum_{i\in\mathscr{B}_\gamma} \mathbb{I}_{\{R_i=-1\}}} \ge \varepsilon \middle| \mathscr{E}_0\right) \le \frac{2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}}{1-2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}} + e^{-\frac{3m}{14}}. \tag{27}$$

Combining (23) and (27) we have that for (18),

$$\mathbb{P}(\text{FDP}_2 \ge (1+\varepsilon)\alpha | \mathscr{E}_0) \le 2e^{-\frac{m\varepsilon^2}{6}} + \frac{2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}}{1-2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}} + 2e^{-\frac{3m}{14}}.$$

Furthermore,

$$\mathbb{P}(\text{FDP}_2 \ge (1+\varepsilon)\alpha) \le \sup_{\mathscr{E}_0} \mathbb{P}(\text{FDP}_2 \ge (1+\varepsilon)\alpha | \mathscr{E}_0) \tag{28}$$

$$\le 2e^{-\frac{m\varepsilon^2}{6}} + + \frac{2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}}{1-2e^{-\frac{m\varepsilon^2}{4(\varepsilon+2)^2}}} + 2e^{-\frac{3m}{14}}. \tag{29}$$

By equaling the term in the right-hand-side of (29) with $\frac{\delta}{2}$ we have $\varepsilon = \Theta(\sqrt{\frac{\log\frac{1}{\delta}}{m}})$. Recall that $m = \alpha c_0 N$ where c_0 is a constant, we have

$$\varepsilon = \Theta(\sqrt{\frac{\log\frac{1}{\delta}}{\alpha N}}),$$

which concludes the proof.

In order for the proof to hold, it is required that the mirror estimate $\widehat{\mathrm{FD}}_2(\gamma t^*)$ is stochastically no less than the true number of false discoveries $\mathrm{FD}_2(\gamma t^*)$ for any $\gamma \geq 0$. This is still true when the i.i.d. assumption for the null p-values is extended to the assumption that the null p-values, conditional on the covariates, are independently distributed and stochastically greater than the uniform distribution. Hence the result is directly extendable. □

4.2 Lemma 1 with proof

Lemma 1. *(Some properties of random walk) Let $R_1, R_2, \cdots$ be i.i.d. Rademacher random variables and let $S_k = \sum_{i=1}^k R_k$. Then for any integer $n > 1$ and for any real number $t > 0$,*

$$\mathbb{P}(\max_{1\leq k\leq n} S_k \geq t) \leq 2e^{-\frac{t^2}{2n}} \tag{30}$$

$$\mathbb{P}(\max_{k\geq n} \frac{1}{k} S_k \geq t) \leq \frac{2e^{-\frac{nt^2}{4}}}{1-2e^{-\frac{nt^2}{4}}} \tag{31}$$

$$\mathbb{P}(\max_{k\geq n} \frac{S_k}{\sum_{i=1}^k \mathbb{I}_{\{R_i=-1\}}} \geq t) \leq \frac{2e^{-\frac{nt^2}{4(t+2)^2}}}{1-2e^{-\frac{nt^2}{4(t+2)^2}}}, \tag{32}$$

where for the second and the third inequalities, we require t to be large enough for the probability to be positive.

Proof. (30) is proved via a standard reflection argument for random walk. First consider when t is an integer,

$$\begin{aligned}
&\mathbb{P}(\max_{1\leq k\leq n} S_k \geq t) = \mathbb{P}(\max_{1\leq k\leq n} S_k \geq t, S_n \geq t) + \mathbb{P}(\max_{1\leq k\leq n} S_k \geq t, S_n < t) \\
&= \mathbb{P}(S_n \geq t) + \mathbb{P}(\max_{1\leq k\leq n} S_k \geq t, S_n > t) = \mathbb{P}(S_n \geq t) + \mathbb{P}(S_n > t) \leq 2\mathbb{P}(S_n \geq t).
\end{aligned}$$

If t is not an integer,

$$\mathbb{P}(\max_{1\leq k\leq n} S_k \geq t) = \mathbb{P}(\max_{1\leq k\leq n} S_k \geq \lceil t \rceil) \leq 2\mathbb{P}(S_n \geq \lceil t \rceil) \leq 2\mathbb{P}(S_n \geq t).$$

Finally, using Hoeffding's inequality, one has

$$\mathbb{P}(\max_{1\leq k\leq n} S_k \geq t) \leq 2\mathbb{P}(S_n \geq t) \leq 2e^{-\frac{t^2}{2n}}.$$

(31) is proved via a technique called "peeling". Specifically,

$$\begin{aligned}
&\mathbb{P}(\max_{k\geq n} \frac{1}{k} S_k \geq t) \leq \mathbb{P}(\exists k \geq n, S_k \geq kt) \\
&\leq \sum_{j=0}^{\infty} \mathbb{P}(\exists k \in \{2^j n, 2^j n + 1, \cdots, 2^{j+1} n - 1\}, S_k \geq kt) \\
&\leq \sum_{j=0}^{\infty} \mathbb{P}(\exists k \in \{2^j n, 2^j n + 1, \cdots, 2^{j+1} n - 1\}, S_k \geq 2^j nt) \\
&\leq \sum_{j=0}^{\infty} \mathbb{P}(\max_{1\leq k\leq 2^{j+1} n} S_k \geq 2^j nt) \\
&\leq \sum_{j=0}^{\infty} 2\exp(-2^{j-2} nt^2) \overset{def}{=} \sum_{j=0}^{\infty} p_j,
\end{aligned}$$

where the last inequality is due to (30) that we have just proved. Note that for $j \geq 0$, $\frac{p_{j+1}}{p_j} = \exp(-2^{j-2} nt^2) \leq p_0$. Hence,

$$\mathbb{P}(\max_{k\geq n} \frac{1}{k} S_k \geq t) \leq \frac{p_0}{1-p_0} = \frac{2e^{-\frac{nt^2}{4}}}{1-2e^{-\frac{nt^2}{4}}}.$$

Finally, (32) is a direct consequence of (31):

$$\mathbb{P}(\max_{k\geq n}\frac{S_k}{\sum_{i=1}^{k}\mathbb{I}_{\{R_i=-1\}}}\geq t)=\mathbb{P}(\max_{k\geq n}\frac{2S_k}{k-S_k}\geq t)$$

$$=\mathbb{P}(\max_{k\geq n}\frac{1}{k}S_k\geq\frac{t}{t+2})\leq\frac{2e^{-\frac{nt^2}{4(t+2)^2}}}{1-2e^{-\frac{nt^2}{4(t+2)^2}}}.$$

□

4.3 Lemma 2 with proof

Lemma 2. *(Some properties of non-homogeneous Bernoulli sum) Let $B_i\sim Bern(p_i)$ be some independent Bernoulli random variables. Then*

$$\mathbb{P}(\sum_{i=1}^{n}B_i-\mathbb{E}[\sum_{i=1}^{n}B_i]\geq t)\leq e^{-\frac{\frac{1}{2}t^2}{\Sigma_{i=1}^{n}p_i+\frac{1}{3}t}} \tag{33}$$

$$\mathbb{P}(\sum_{i=1}^{n}B_i-\mathbb{E}[\sum_{i=1}^{n}B_i]\leq -t)\leq e^{-\frac{\frac{1}{2}t^2}{\Sigma_{i=1}^{n}p_i+\frac{1}{3}t}} \tag{34}$$

Proof. Define $X_i\overset{def}{=}B_i-p_i$. Then X_i's have zero means and are independent of each other. Also, note that $|X_i|\leq 1$ almost surely and $\sum_i\mathbb{E}[X_i^2]\leq\sum_i p_i$. Hence (33) and (34) can be obtained by applying Bernstein inequality on $\{X_i\}$ and $\{-X_i\}$ respectively. □