2020

Nonparametric False Discovery Rate Control for Identifying Simultaneous Signals

Sihai Dave Zhao

Yet Tian Nguyen

Electronic Journal of Statistics
Vol. 14 (2020) 110–142
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1663

Nonparametric false discovery rate control for identifying simultaneous signals

Sihai Dave Zhao

Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
e-mail: sdzhao@illinois.edu

and

Yet Tien Nguyen

Department of Mathematics & Statistics
Old Dominion University
Norfolk, VA 23529, USA
e-mail: ynguyen@odu.edu

Abstract: It is frequently of interest to identify simultaneous signals, defined as features that exhibit statistical significance across each of several independent experiments. For example, genes that are consistently differentially expressed across experiments in different animal species can reveal evolutionarily conserved biological mechanisms. However, in some problems the test statistics corresponding to these features can have complicated or unknown null distributions. This paper proposes a novel nonparametric false discovery rate control procedure that can identify simultaneous signals even without knowing these null distributions. The method is shown, theoretically and in simulations, to asymptotically control the false discovery rate. It was also used to identify genes that were both differentially expressed and proximal to differentially accessible chromatin in the brains of mice exposed to a conspecific intruder. The proposed method is available in the R package github.com/sdzhao/ssa.

Keywords and phrases: False discovery rate, replicability, multiple testing, simultaneous signals.

Received January 2019.

1. Introduction

Multiple hypothesis testing is now a staple of scientific research, and summary statistics, such as test statistics and p-values, from previously conducted experiments are now readily publicly available. Jointly analyzing summary statistics from different independent experiments can provide scientific insights that cannot be achieved from a single experiment alone. One important type of joint analysis is to identify features that are non-null in each of several experiments,

which will be referred to as simultaneous signals. To be precise, given n features in D experiments, define I_{id} to be an unobserved non-random signal indicator that equals 0 when the null hypothesis is true for feature i in experiment d, and equals 1 when the alternative hypothesis is true. The set of simultaneous signals is defined to be

$$\mathcal{S} = \{i \in \{1, \ldots, n\} : I_{i1} = \ldots = I_{iD} = 1\}, \tag{1}$$

so that the set of non-simultaneous signals is $\mathcal{S}^c = \{i : i \notin \mathcal{S}\}$.

The problem of identifying simultaneous signals arises in many different contexts. For example, it is of interest to identify genetic variants that are associated with multiple related conditions, such as psychiatric disorders [1, 27, 28], to uncover potentially shared disease mechanisms. Similarly, it is useful to identify regions in the genome that are simultaneously associated both with a disease outcome and with a gene expression, as these locations may contain important causal mutations [21, 48]. As another example, identifying findings that replicate across independent experiments is a crucial component of reproducible research [8, 23, 24]. Finally, comparative genomics research aims to find genes associated with similar phenotypes across different animal species, in hopes of finding evolutionarily conserved genomic programs [33, 39, 47].

This paper studies the problem of identifying simultaneous signals under false discovery rate control. Specifically, let T_{id} be a univariate test statistic corresponding to the ith feature in the dth experiment, such that a hypothesis test based on T_{id} can be performed to infer the true value of the signal indicator I_{id}. Let $\delta : \mathbb{R}^D \to \{0, 1\}$ be a simultaneous signal discovery procedure where $\delta(T_{i1}, \ldots, T_{iD}) = 1$ declares $i \in \mathcal{S}$ and $\delta(T_{i1}, \ldots, T_{iD}) = 0$ declares $i \in \mathcal{S}^c$. False discovery rate control methods aim to maximize the number of discovered simultaneous signals while maintaining the false discovery rate

$$\text{FDR}(\delta) = E\left[\frac{\sum_{i \in \mathcal{S}^c} \delta(T_{i1}, \ldots, T_{iD})}{\max\{1, \sum_{i=1}^{n} \delta(T_{i1}, \ldots, T_{iD})\}}\right] \tag{2}$$

to be at most α, for some prespecified $\alpha < 1$.

There has been a great deal of recent work on methods to control the false discovery rate when identifying simultaneous signals. An *ad hoc* approach is to use a standard procedure, like that of Benjamini and Hochberg [7], to discover significant features separately in each of the D experiments and then to identify discoveries common to all experiments. Bogomolov and Heller [9] developed a modified version of this idea and proved that their procedure maintains false discovery rate control. Another common strategy is to summarize the D statistics for each feature i into a single scalar statistic, for example by taking the maximum of their corresponding p-values [30]. This reduces the problem to a single sequence of multiple tests, but it is unclear how to choose the best summary function. A more principled approach treats the T_{id} as a single sequence of multivariate test statistics $(T_{i1}, \ldots, T_{iD})$. In this framework, it has been shown that the local false discovery rate [18] is the optimal scalar summary of the multivariate test statistics [11, 13, 14, 24]. This can be difficult to calculate in practice, so Chung

et al. [13] assumed a parametric model and used the EM algorithm to estimate unknown parameters, Chi [11] proposed a Taylor expansion approximation, Du and Zhang [14] used a single-index model approximation, and Heller and Yekutieli [24] employed an empirical Bayes approach. Xiang, Zhao and Cai [51] recently introduced a new framework for the joint analysis of multiple tests from independent studies, of which simultaneous signal identification is a special case.

All of these methods assume that the null distributions of the test statistics T_{id} are known to some extent. Many assume that p-values are available, meaning that the nulls must be known exactly. Others estimate parameters of the null distributions, which still requires knowing the parametric families to which the nulls belongs [40]. However, in many important problems in genomics, information about the null distributions of the T_{id} is not readily available, for at least three common reasons. First, small sample sizes can make it difficult to obtain the exact null distribution of standard test statistics [53]. Second, complex test statistics can have intractable null distributions. For example, the null distribution of the SKAT statistic [50], which tests the significance of a set of genetic variants, does not have a convenient closed form and in practice is computationally approximated. Finally, complex data types can give rise to null distributions that are difficult to model or characterize. For example, data from ChIP-seq experiments [29] are used to identify regions of the genome where transcription factors are found to bind, but the number, size, and locations of these regions are not predetermined. This makes accurate quantification of the statistical significance of the identified regions very difficult [12].

To date, relatively little work has considered false discovery rate control when null distributions are not completely known. Some results are available given test statistics from a single experiment. Knockoff filters [2, 3, 4, 10] assume only that the null distributions are identical and symmetric, and p-filters [5, 32] assume only that the test statistics can be converted to random variables between 0 and 1 that are stochastically larger than a uniformly distributed random variable. Resampling-based procedures [15, 36, 49, 52] do not require known null distributions, but can only be used if the raw data are available. This may not be true for some applications, such as in genetics, where it is common that only test statistics are easily accessible.

In contrast, results are lacking when there are two or more experiments of interest. Nonparametric methods for detecting the presence of simultaneous signals have been proposed [55, 56], but there do not seem to exist methods for identifying them when null distributions are unknown. Section 5 describes a simultaneous signal identification problem, encountered in a study of mouse behavioral genomics [38], where the null distributions in one of the experiments was unknown. No existing false discovery rate control method can be applied to this problem.

This paper proposes a novel nonparametric method for controlling the false discovery rate for identifying simultaneous signals when the test statistics have unknown null distributions. A tradeoff of its robustness is that it can be very conservative, especially if the proportion of simultaneous signals is high and the studies $D \leq 3$; see Sections 2.2 and 2.5 for further discussion. Section 2 describes

the proposed procedure and shows that it can asymptotically control the false discovery rate at the nominal level. Section 3 discusses an alternative procedure that has more power but requires much more restrictive conditions. Section 4 illustrates the performance of the proposed method in simulations and Section 5 applies it in the mouse behavioral genomics problem. Section 6 concludes with a discussion, and proofs of all technical results can be found in the Appendix.

2. Proposed procedure

2.1. Model

The observed data consist of D sequences of n univariate test statistics T_{id}, corresponding to features $i = 1, \ldots, n$ from experiments $d = 1, \ldots, D$. For each T_{id}, let F^0_{id} and F^1_{id} denote the cumulative distribution functions under the null and alternative hypotheses, respectively, and S^0_{id} and S^1_{id} denote the corresponding survival functions. Therefore

$$T_{id} \sim F^I_{id} = 1 - S^I_{id} \text{ when } I_{id} = I \tag{3}$$

for $I = 0, 1$, where the I_{id} are unobserved non-random indicators of whether the null hypothesis is actually true or false, as introduced in Section 1. Within each sequence d, it is assumed that the T_{id} are mutually independent, and the D sequences are also assumed to be mutually independent. In other words, T_{id} and $T_{i'd'}$ are independent if $i \neq i'$ or $d \neq d'$. Finally, the test statistics are assumed to be one-tailed, with larger values of T_{id} giving more evidence against the null. This is formalized in Assumption 1.

Assumption 1. *For all t, $S^0_{id}(t) < S^1_{id}(t)$.*

This paper adopts a fixed-effects model where the I_{id} are non-random quantities. A popular alternative in the multiple testing literature [24, 46, 51] is the random-effects framework, where the $(I_{i1}, \ldots, I_{iD})$ are modeled as independent and identically distributed random $D \times 1$ Bernoulli vectors, whose components I_{id} and $I_{id'}$ can be dependent for $d \neq d'$. The T_{id} are then assumed to be conditionally independent given $(I_{i1}, \ldots, I_{iD})$ such that

$$(T_{i1}, \ldots, T_{iD}) \mid (I_{i1}, \ldots, I_{iD}) \sim \prod_{d=1}^{D} F^{I_{id}}_{id}(t_d). \tag{4}$$

These fixed- and random-effects models are closely related [20, 44], and the latter can be useful for interpreting certain aspects of the proposed procedure; see Section 2.2.

2.2. Two sequences of test statistics

The proposed method is first introduced assuming that only two sequences of test statistics T_{id} are observed, $i = 1, \ldots, n$ and $d = 1, 2$. Section 2.5 describes a potential extension when there are more than two sequences of interest.

The overall strategy follows the framework of Storey, Taylor and Siegmund [45] for false discovery rate control in a single sequence of test statistics. The proposed procedure declares a feature i to be a simultaneous signal if

$$(T_{i1}, T_{i2}) \in [t, \infty) \times [t, \infty) \tag{5}$$

for an appropriately chosen threshold t. The goal is to choose a threshold t that discovers the most simultaneous signals while maintaining an acceptable false discovery rate.

This requires estimating the false discovery rate that would be attained by a particular threshold t. To motivate this estimator, suppose for now that $S^0_{id} = S^0_d$ and $S^1_{id} = S^1_d$ for all features i. This condition is much stronger than necessary and will be weakened in Assumption 2. Then under model (3), the expected proportion of false positives would equal

$$n^{-1} \sum_{i \in \mathcal{S}^c} S_1^{I_{i1}}(t) S_2^{I_{i2}}(t),$$

where $\mathcal{S}$ is the set of simultaneous signals defined in (1). The following result shows that this expected proportion can be upper-bounded by the product of marginal survival functions.

Proposition 1. *For $d = 1, 2$, define the marginal signal proportion*

$$\pi_d = \frac{|\{i : I_{id} = 1\}|}{n}$$

and $S_d(t) = (1 - \pi_d) S^0_d(t) + \pi_d S^1_d(t)$, where $S^0_d(t)$ and $S^1_d(t)$ are any survival functions that satisfy $S^0_d(t) < S^1_d(t)$. Then under Assumption 1,

$$n^{-1} \sum_{i \in \mathcal{S}^c} S_1^{I_{i1}}(t_1) S_2^{I_{i2}}(t_2) \leq S_1(t_1) S_2(t_2)$$

for any t_1 and t_2, with $\mathcal{S}$ defined in (1).

Proposition 1 motivates the following conservative estimator for the false discovery rate that would be attained by the rejection region $[t, \infty) \times [t, \infty)$:

$$\widehat{\mathrm{FDR}}_\rho(t) = \frac{\hat{S}_1(t) \hat{S}_2(t) + \rho}{\max\{n^{-1}, \hat{G}(t, t)\}}, \tag{6}$$

where $\hat{S}_d(t) = n^{-1} \sum_{i=1}^n I(T_{id} > t)$ are empirical marginal survival functions, $\hat{G}(t, t) = n^{-1} \sum_{i=1}^n I(T_{i1} > t, T_{i2} > t)$ is the total proportion of rejected features, and ρ is a positive constant that regularizes the asymptotic properties of the proposed procedure. An alternative to (6) would be to define $\widehat{\mathrm{FDR}}_\rho(t) = 0$ if $\hat{G}(t, t) = 0$, but (6) is more convenient for proving asymptotic false discovery rate control.

This leads to the proposed nonparametric discovery procedure

$$\begin{aligned} \hat{\delta}_\rho(T_{i1}, T_{i2}) &= I(T_{i1} \geq \hat{t}_\rho, T_{i2} \geq \hat{t}_\rho), \\ \hat{t}_\rho &= \inf\left\{t \in [0, \infty) : \widehat{\mathrm{FDR}}_\rho(t) \leq \alpha\right\}, \end{aligned} \tag{7}$$

for some desired false discovery rate $\alpha < 1$. Features with $\hat{\delta}_\rho(T_{i1}, T_{i2}) = 1$ are declared to be simultaneous signals. The threshold $\hat{t}_\rho$ maximizes the number of rejected features while maintaining a conservative estimate of the true false discovery rate to be below α.

This estimator does not require any knowledge of S^0_{id} or S^1_{id} beyond the stochastic ordering of Assumption 1. The tradeoff is that the proposed $\widehat{\text{FDR}}_\rho(t)$ can give a very conservative estimate of the true false discovery rate. The proof of Proposition 1 shows that

$$n^{-1} \sum_{i \in \mathcal{S}^c} S_1^{I_{i1}}(t_1) S_2^{I_{i2}}(t_2) = S_1(t_1) S_2(t_2) + \pi_{(1,1)} C_1(t_1, t_2) + \pi_1 \pi_2 C_2(t_1, t_2),$$

where $\pi_{(1,1)}$ is the proportion of simultaneous signals and $C_1(t_1, t_2)$ and $C_2(t_1, t_2)$ are positive quantities. The bound in Proposition 1 can therefore be very loose if the proportions of marginal or simultaneous signals are high. As a result, the actual false discovery rate attained by the proposed method can be much lower than the target level α. On the other hand, there are many settings where these signal proportions are expected to be low, such as in genomics studies where only small proportion of the genomic features are expected to be associated with the outcome of interest.

Remark 1. The function $\hat{G}(t, t)$ in the denominator of $\widehat{\text{FDR}}_\rho(t)$ (6) does not always converge to $S_1(t)S_2(t)$, even though T_{i1} and T_{i2} are independent under the fixed-effects model (3). This is because the (T_{i1}, T_{i2}) are not identically distributed, due to signal indicators (I_{i1}, I_{i2}) are that are different for different features i. This observation motivates an interesting alternative interpretation of the proposed procedure, which is more easily described using the random-effects model (4) from Section 2.1. In this framework, the (I_{i1}, I_{i2}) are modeled as identically distributed random Bernoulli tuples and T_{i1} and T_{i2} are independent conditional on (I_{i1}, I_{i2}). Therefore the (T_{i1}, T_{i2}) are marginally identically distributed, but dependence between I_{i1} and I_{i2} will make T_{i1} and T_{i2} marginally dependent. Since $\hat{G}(t, t)$ estimates the marginal bivariate distribution function of (T_{i1}, T_{i2}), $\widehat{\text{FDR}}_\rho(t)$ can be thought of measuring the departure of $\hat{G}(t, t)$ from independence. In this sense, procedure (7) appears closely related to testing for independence between the T_{i1} and the T_{i2}.

Remark 2. The rejection regions in (5) are rectangular, but many other non-rectangular shapes can be used. In fact, Heller and Yekutieli [24] showed that the optimal rejection region is a level curve of the local false discovery rate. On the other hand, rectangular regions are simple to implement and interpret, and they allow the expected proportion of false positives to be upper-bounded using only marginal survival functions, via Proposition 1. This is crucial to the nonparametric nature of the proposed procedure, and it is not clear whether there exist non-rectangular rejection regions that have this property.

2.3. Rank transformation

The rejection region (5) uses the same threshold for both T_{i1} and T_{i2}. This may not be appropriate when the null distributions S^0_{i1} and S^0_{i2} are not comparable,

because the test statistics from the two sequences will be on different scales. A simple solution is to transform the T_{id} within each sequence to their corresponding ranks, which makes the sequences more comparable. This is equivalent to replacing $\widehat{\text{FDR}}_\rho(t)$ with

$$\frac{n^{-1}\sum_{i=1}^n I\{\hat{F}_1(T_{i1}) \geq t\} n^{-1}\sum_{i=1}^n I\{\hat{F}_2(T_{i2}) \geq t\} + \rho}{\max[n^{-1}, n^{-1}\sum_{i=1}^n I\{\hat{F}_1(T_{i1}) \geq t, \hat{F}_2(T_{i2}) \geq t\}]},$$

where $\hat{F}_d = 1 - \hat{S}_d$. The rank transformation procedure is therefore equivalent to considering rejection regions $[\hat{F}_1^{-1}(t), \infty) \times [\hat{F}_2^{-1}(t), \infty)$ instead of region (5).

This procedure can be less powerful than a procedure that knows the true null distributions S_{id}^0 of the T_{id}. When the nulls are known, the correct way to place the test statistics on the same scale would be to convert the T_{id} to p-values, which would replace $\hat{F}_1$ and $\hat{F}_2$ above with $1-S_{i1}^0$ and $1-S_{i2}^0$. This would correspond to considering to rejection regions $[(S_{i1}^0)^{-1}(1-t), \infty) \times [(S_{i2}^0)^{-1}(1-t), \infty)$. However, these regions may not coincide with those considered by rank transformation. This can happen, for example, if $S_{i1}^0 = S_{i2}^0$ but $\hat{F}_1 \neq \hat{F}_2$, and can result in the rank transformation procedure having lower power. This is illustrated in simulations in Section 4.1.

An alternative to the rank transformation approach is to consider rejection regions $[t_1, \infty) \times [t_2, \infty)$ [11, 14]. This allows different thresholds for the different sequences to be learned from the data, without needing knowledge of the null distributions. However, the resulting procedure seems to require very restrictive conditions in order to guarantee false discovery rate control. This is discussed in detail in Section 3.

2.4. *Theoretical properties*

To motivate the proposed procedure, it was temporarily assumed that $S_{id}^0 = S_d^0$ and $S_{id}^1 = S_d^1$ for all features i. However, in general the test statistics for different features may have different null and alternative distributions. Even so, the proposed method still has good properties under the following assumption about the S_{id}^0 and S_{id}^1. Let $\mathbf{I}_i = (I_{i1}, \ldots, I_{iD})$ be the vector of true signal indicators corresponding to feature i in each sequence of test statistics.

Assumption 2. *The following hold for every sequence* $d = 1, \ldots, D$.

(a) For $I \in \{0, 1\}$, *there exist continuous functions* S_d^I *such that*

$$\lim_{n\to\infty} \frac{1}{|\{i : I_{id} = I\}|} \sum_{i:I_{id}=I} S_{id}^I(t) = S_d^I(t)$$

uniformly in t.

(b) For ever vector $\mathbf{I} \in \{0, 1\}^D$ *such that* $\mathbf{I} \neq (1, \ldots, 1)$,

$$\lim_{n\to\infty} \frac{1}{|\{i : \mathbf{I}_i = \mathbf{I}\}|} \sum_{i:\mathbf{I}_i=\mathbf{I}} \prod_{d=1}^D S_{id}^{I_d}(t_d) = \prod_{d=1}^D S_d^{I_d}(t_d)$$

uniformly in $(t_1, \ldots, t_D)$*, with* S_d^0 *and* S_d^1 *defined in item (a) above.*

(c) *For* $\mathbf{I} = (1, \ldots, 1)$*, there exists a continuous function* G^1 *such that uniformly in* $(t_1, \ldots, t_D)$*,*

$$\lim_{n\to\infty} \frac{1}{|\{i : \mathbf{I}_i = \mathbf{I}\}|} \sum_{i:\mathbf{I}_i=\mathbf{I}} \prod_{d=1}^{D} S_{id}^1(t_d) = G^1(t_1, \ldots, t_D).$$

(d) *For every* $\mathbf{I} \in \{0,1\}^D$*, there exist proportions* $\pi_{\mathbf{I}}$ *such that*

$$\lim_{n\to\infty} \frac{|\{i : \mathbf{I}_i = \mathbf{I}\}|}{n} \to \pi_{\mathbf{I}}.$$

Assumption 2 endows the fixed-effects model (3) with certain useful properties of the random-effects model (4), and is similar to assumptions introduced by Genovese and Wasserman [20] and others [44, 45]. Assumption 2(a) posits the existence of limiting survival functions S_d^0 and S_d^1, which can be thought of as the marginal null and alternative distributions of T_{id} under the random-effects model. Assumption 2(b) recovers the random-effects assumption that the components of $(T_{i1}, \ldots, T_{id})$ are independent conditional on $\mathbf{I}_i$. In fact it is slightly weaker, requiring that the T_{id} are conditionally independent only when $\mathbf{I}_i \neq (1, \ldots, 1)$, as the joint distribution G^1 defined in Assumption 2(c) need not equal the product of marginal survival functions. Finally, the $\pi_{\mathbf{I}}$ defined in Assumption 2(d) can be viewed as the probability that $\mathbf{I}_i = \mathbf{I}$ under the random-effects model.

The main result of this paper is that the proposed procedure can achieve asymptotic false discovery rate control.

Theorem 1. *Under Assumptions 1 and 2, the proposed procedure* (7) *with* $\rho > 0$ *satisfies*

$$\limsup_{n\to\infty} \text{FDR}(\hat{\delta}_\rho) \leq \alpha,$$

where FDR *is the true false discovery rate defined in* (2)*.*

Finite-sample rather than asymptotic false discovery rate control would be ideal, and the proof of Theorem 1 suggests that this might be possible if the null and alternative distributions did not vary across features, and if the marginal survival functions S_d in $\widehat{\text{FDR}}_\rho(t)$ (6) were known rather than estimated. The condition that $\rho > 0$ is necessary for technical reasons, but the simulations in Section 4 indicates that using $\rho = 0$ still gives good results.

Procedure (7) can provably control the false discovery rate for simultaneous signals without any knowledge of the null or alternative distributions, aside from the stochastic ordering condition of Assumption 1. Because of this, it pays a price in terms of power to detect simultaneous signals. A major reason is that the bound in Proposition 1 is not tight, which causes the selected threshold $\hat{t}_\rho$ (7) to be larger than necessary. In particular, the proof of Proposition 1 indicates that the bound is most tight when each sequence of test statistics has very few signals and when there are very few simultaneous signals.

2.5. Two or more sequences of test statistics

In some problems, the goal may be to discover features that are simultaneously significant across $D > 2$ sequences of test statistics. The proposed method can be extended to this setting by consider rejection regions of the form $[t, \infty)^D$ for a threshold t. Under model (3) and Assumption 2, the expected number of false positives that would be discovered by this region can be upper-bounded using the following generalization of Proposition 1.

Proposition 2. *For $D \geq 2$,*

$$n^{-1} \sum_{i \in \mathcal{S}^c} \prod_{d=1}^{D} S_d^{I_{id}}(t_d) \leq \sum_{d,d' \in 1,\ldots,D, d \neq d'} S_d(t_d) S_{d'}(t_{d'})$$

for any $t_1, \ldots, t_D$, with S_d defined in Proposition 1 and $\mathcal{S}$ defined in (1).

Following the reasoning in Section 2.2, Proposition 2 therefore motivates the following discovery procedure for any number $D \geq 2$ of sequences:

$$\begin{aligned} &\hat{\delta}_\rho(T_{i1}, \ldots, T_{iD}) = I(T_{i1} \geq \hat{t}_\rho, \ldots, T_{iD} \geq \hat{t}_\rho), \\ &\hat{t}_\rho = \inf \left[t \in [0, \infty) : \frac{\sum_{d,d' \in 1,\ldots,D, d \neq d'} \hat{S}_d(t) \hat{S}_{d'}(t) + \rho}{\max\{n^{-1}, n^{-1} \sum_{i=1}^{n} I(T_{i1} > t, \ldots, T_{iD} > t)\}} \leq \alpha \right], \end{aligned} \tag{8}$$

and features with $\hat{\delta}_\rho(T_{i1}, \ldots, T_{iD}) = 1$ are declared as simultaneous signals across all D sequences. It is straightforward to extend the proof of Theorem 1 to this generalized procedure. This reduces to procedure (7) when $D = 2$.

Procedure (8) can suffer from very low power because the bound in Proposition 2 becomes exceedingly conservative for larger D. When $D > 2$, this bound is derived from repeated applications of the basic bound for $D = 2$. This is likely not an optimal approach, and further work is necessary to design a more powerful nonparametric discovery procedure for more than two sequences. Nevertheless, simulations in Section 4 suggest that the method is still serviceable when $D \leq 3$, which covers many practical problems. Furthermore, few other methods are available when the null and alternative distributions are unknown.

3. Alternative procedure

For $D = 2$, Remark 1 of Section 2.2 pointed out that a more natural alternative to rejection region (5) is the rectangle $[t_1, \infty) \times [t_2, \infty)$, which allows a different threshold for each sequence of test statistics. Applying Proposition 1 to this rejection region suggests the conservative false discovery rate estimator

$$\widetilde{\mathrm{FDR}}_\rho(t_1, t_2) = \frac{\hat{S}_1(t_1) \hat{S}_2(t_2) + \rho}{\max\{n^{-1}, \hat{G}(t_1, t_2)\}}, \tag{9}$$

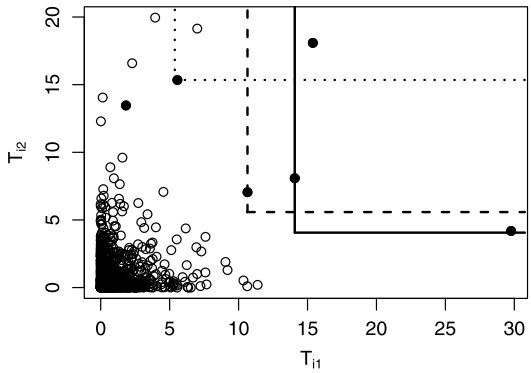

FIG 1. *The alternative false discovery rate control procedure* (10) *does not produce a unique rejection region. The solid, dashed, and dotted lines demarcate three distinct regions that maximize the number of rejections while satisfying* $\widetilde{\text{FDR}}_\rho(t_1, t_2) \leq 0.05$. *Filled circles denote true simultaneous signals.*

with $\hat{S}_d$ and ρ as defined in (6) and $\hat{G}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} I(T_{i1} > t_1, T_{i2} > t_2)$. This can be shown to be an asymptotically uniformly conservative estimate of the false discovery rate incurred by the rejection region.

Theorem 2. *For any discovery procedure of the form* $\delta(T_{i1}, T_{i2}) = I(T_{i1} \geq t_1, T_{i2} \geq t_2)$, *under Assumptions 1 and 2,*

$$\lim_{n\to\infty} \inf_{t_1 \leq \eta_1, t_2 \leq \eta_2} \left\{ \widetilde{\text{FDR}}_\rho(t_1, t_2) - \text{FDR}(\delta) \right\} \geq 0$$

almost surely, for fixed $\eta_1, \eta_2 < \infty$.

This leads to the following alternative to the proposed procedure (7):

$$\begin{aligned} &\tilde{\delta}_\rho(T_{i1}, T_{i2}) = I(T_{i1} \geq \hat{t}_{\rho 1}, T_{i2} \geq \hat{t}_{\rho 2}), \\ &(\hat{t}_{\rho,1}, \hat{t}_{\rho 2}) = \underset{(t_1,t_2)\in\Pi}{\arg\max}\, \hat{G}(t_1, t_2) \text{ subject to } \widetilde{\text{FDR}}_\rho(t_1, t_2) \leq \alpha, \end{aligned} \tag{10}$$

where the set $\Pi = \{(\infty, \infty)\} \cup \{(T_{i1}, T_{i'2}) : 1 \leq i, i' \leq n\}$ is the union of the point (∞, ∞) along with the Cartesian product of the two sequences of observed test statistics. The $\hat{t}_{\rho 1}$ and $\hat{t}_{\rho 2}$ are chosen to maximize $\hat{G}(t_1, t_2)$, which is equivalent to maximizing the number of rejected features, subject to controlling the estimated false discovery rate bound.

The $\hat{t}_{\rho 1}$ and $\hat{t}_{\rho 2}$ defined in (10) are not unique, in that there can exist multiple distinct rejection regions that maximize $\hat{G}(t_1, t_2)$. Figure 1 illustrates an example where $n = 1{,}000$, the marginal signal proportions were $\pi_1 = 0.019$ and $\pi_2 = 0.012$, there were six simultaneous signals, the null $T_{id} \sim \chi^2_1$, and the non-null $T_{id} \sim \chi^2_1(9)$. Each of the rectangular rejection regions at the $\alpha = 0.05$ level rejects a different set of three features. Nevertheless, under certain conditions it can be shown that any $(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})$ satisfying (10) can achieve asymptotic false discovery rate control.

Theorem 3. *Under Assumptions 1 and 2, if there exist* $t_1', t_2' < \infty$ *such that* $\lim_{n\to\infty} \widetilde{\text{FDR}}_\rho(t_1, t_2) < \alpha$, *the alternative procedure* (10) *satisfies*

$$\limsup_{n\to\infty} \text{FDR}(\tilde{\delta}_\rho) \leq \alpha.$$

However, the requirement in Theorem 3 that there exist t_1' and t_2' such that $\lim_{n\to\infty} \widetilde{\text{FDR}}_\rho(t_1', t_2') \leq \alpha$ turns out to be very stringent. It can be shown following the proof of Proposition 1 that such t_1' and t_2' must satisfy

$$\frac{S_1(t_1')S_2(t_2') + \rho}{S_1(t_1')S_2(t_2') + (\pi_{(1,1)} - \pi_1\pi_2)\{S_1^1(t_1') - S_1^0(t_1')\}\{S_2^1(t_2') - S_2^0(t_2')\}} \leq \alpha,$$

where $\pi_{(1,1)}$ is the proportion of simultaneous signals. But this can only happen when $\pi_{(1,1)} > \pi_1\pi_2$, where π_d is the proportion of signals in study d. If it does not hold, the alternative procedure may not be able to maintain the false discovery rate at the nominal level. This is in contrast to Theorem 1 for the proposed procedure $\hat{\delta}_\rho$ in (7), which guarantees asymptotic false discovery rate control without needing this condition.

This is important, for example, when applying $\tilde{\delta}_\rho$ (10) to sequences where no simultaneous signals actually exist. Then $\pi_{(1,1)} = 0$, and $\tilde{\delta}_\rho$ may incorrectly identify one or two features as simultaneous signals because it cannot maintain the nominal false discovery rate. This alternative procedure is thus not pursued in the remainder of this paper. Furthermore, the rank transformation described in Section 2.3 obviates the need for a different threshold for each sequence of test statistics.

4. Simulations

4.1. Effect of rank transformation

This section explores the effectiveness of the rank transformation described in Section 2.3 for $D = 2$ sequences for a variety of null distributions of different scales. The proposed discovery procedure $\hat{\delta}_\rho$ in (7), with $\rho = 0$, was applied in three ways at a nominal $\alpha = 0.05$ false discovery rate. First, $\hat{\delta}_\rho$ was applied to directly to T_{id} to demonstrate the consequences of ignoring the different scales of the test statistics. Next, p-values P_{id} were calculated based on the T_{id}, and the proposed method was applied without rank transformation to $-\log_{10} P_{id}$, a p-value transformation common in the genomics literature. This represents an oracle version of the proposed method that uses information about the true null distributions to place the test statistics on comparable scales. Finally, $\hat{\delta}_\rho$ was applied to rank-transformed T_{id} to attempt to recover the oracle performance.

Figure 2 reports the false discovery rates and average numbers of discovered simultaneous signals over 200 replications of the following simulation settings. Each setting had $n = 10{,}000$ features, and the two sequences were generated with either equal or unequal numbers of signals in each sequence.

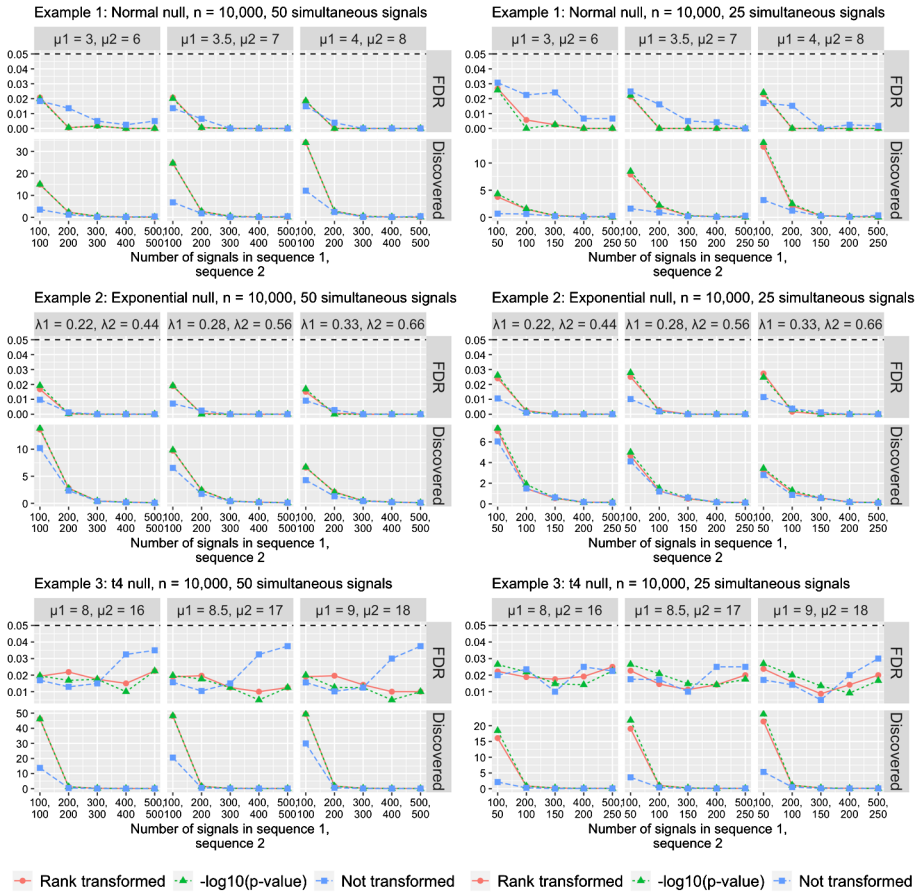

FIG 2. *False discovery rates and average numbers of discovered simultaneous signals of proposed method applied to rank-transformed* T_{id}, $-\log_{10} P_{id}$, *and untransformed* T_{id}, *where* P_{id} *are p-values calculated based on* T_{id} *using their true null distributions. Horizontal dashed lines mark the nominal* $\alpha = 0.05$ *false discovery rate level. The rank-transformed procedure can have less power than the procedure applied to the* $-\log_{10} P_{id}$.

Example 1 (normal null). For sequences $d = 1, 2$, $T_{id} = Z_{id}^2$ where the Z_{id} were independently drawn from $\mathrm{N}(\mu_{id}, d^2)$. If $I_{id} = 0$, $\mu_{id} = 0$, and if $I_{id} = 1$, $\mu_{id} \sim N(\mu_d, 1)$ and was fixed across all replications, for various μ_d. Because the maximum of n independent $\mathrm{N}(0, d^2)$ goes like $d(2\log n)^{1/2}$, μ_d was at most $d(2\log n)^{1/2}$. This was to ensure that the simultaneous signals could not be identified by simply choosing the features with $T_{id} > d(2\log n)^{1/2}$ in both sequences.

Example 2 (exponential null). Simulations followed Example 1 except that the T_{id} were independently generated from $\mathrm{Exp}(\lambda_{id})$ where $\lambda_{id} = 2d$ for $I_{id} = 0$ and $\lambda_{id} \sim \max\{0.1, N(\lambda_d, 0.01)\}$ and fixed across replications for $I_{id} = 1$, for various λ_d. Because the maximum of n independent $\mathrm{Exp}(2d)$ random variable

goes like $(\log n)/(2d)$, λ_d was at least $2d/\log n$, so that the expected value of T_{id} for $I_{id} = 1$ was at most $(\log n)/(2d)$.

Example 3 (t_4 null). Simulations followed Example 1 except that $T_{id} = Y_{id}^2$ where the Y_{id} were independently generated from $d(\mu_{id} + t_4)$ with t_4 denoting a random variable drawn from a t-distribution with four degrees of freedom. If $I_{id} = 0$, $\mu_{id} = 0$, and if $I_{id} = 1$, and $\mu_{id} \sim N(\mu_d, 1)$ and was fixed across all replications, where μ_d was at most $\log n$.

When the two sequences had the same number of signals, the rank-transformed procedure performed essentially exactly as well as the oracle procedure using the p-values calculated with knowledge of the true null distributions. However, when the two sequences had different numbers of signals, the rank-transformed method had lower power, particularly for t_4-distributed T_{id}. As discussed in Section 2.3, this is because the rejection regions considered by the rank-transformed procedure no longer coincide with those considered by the oracle procedure.

4.2. Unknown null distributions

The main motivation of this paper was to develop a method to control the false discovery rate for discovering simultaneous signals when the null and/or alternative distributions of the T_{id} are unknown. These simulations explore this for $D = 2$ sequences. In this section, each sequence of test statistics was generated with the same number of signals, following Bogomolov and Heller [9]. See Appendix A for additional simulation settings.

For each d and i, ten correlated z-scores $(Z_{id1}, \ldots, Z_{id10})$ were generated from $\mathrm{N}(\mu_{id}, \Sigma_{id})$, where $\mu_{id} = (0, \ldots, 0)$ for $I_{id} = 0$ and $\mu_{id} = (\mu_{id1}, \ldots, \mu_{id10})$ with $\mu_{idj} \sim \mathrm{N}(\mu_d, 1)$ and fixed across replications for $I_{id} = 1$, for the same values of μ_d used in Example 1 of Section 4.1. Each Σ_{id} was equal to the empirical correlation matrix of a different set of 10 genes selected from a gene expression study of multiple myeloma, obtained from Shi et al. [43]. These z-scores were converted to correlated p-values $(P_{id1}, \ldots, P_{id10})$, and finally $T_{id} = -2\sum_{j=1}^{10} \log P_{idj}$. The vectors of z-scores were generated independently across both d and i, so that the T_{id} were also independent.

This setting models applications where group testing is applied to multiple groups of correlated genomic features. The groups, indexed by i, are independent, but the features within the groups are not. The null distribution of each T_{id} is complicated and in practice would not be known, as it depends on the unknown correlations between the features.

The proposed procedure was implemented with $\rho = 0$ and rank-transformed T_{id}. To demonstrate the difficulty of this problem, three existing methods described in Section 1 were also implemented:

1. The method of Chung et al. [13] uses p-values P_{id} calculated from the T_{id}.
2. The empirical Bayes method of Heller and Yekutieli [24] uses an estimate of the null distribution of the T_{id}. It first calculates z-scores from the T_{id}

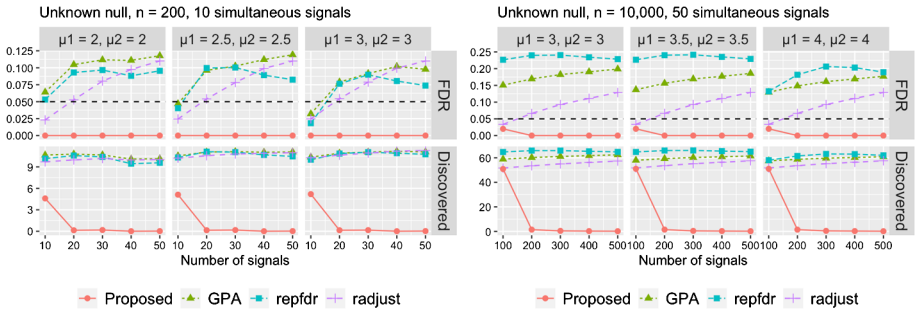

FIG 3. *False discovery rates and average numbers of discovered simultaneous signals when T_{id} has an unknown null distribution. Horizontal dashed lines mark the nominal $\alpha = 0.05$ false discovery rate level. Only the proposed method was capable of maintaining the nominal α. Proposed: proposed approach (7); GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

using their known theoretical nulls, then assumes that the z-scores are normally distributed but with unknown mean and variance, and finally estimates the unknown parameters using the method of Efron [16]. See Efron [18] for a discussion of why using this so-called "empirical null" can be more appropriate than using the theoretical null in multiple testing problems.

3. The method of Bogomolov and Heller [9] is based on first selecting promising features from each sequence based on the P_{id}.

These existing approaches all require known null distributions, so they were implemented assuming that the T_{id} followed χ^2_{20}, which would only be true if Σ_{id} equaled the identity matrix.

Figure 3 reports the false discovery rates and average numbers of discovered simultaneous signals over 200 replications. The proposed method always maintained the false discovery rate at the nominal level, something that none of the other methods could achieve. It appears to be the only existing simultaneous signal discovery procedure that can provably control the false discovery rate in this setting when the T_{id} have complex or unknown null distributions. The tradeoff is that the proposed method is conservative, in that the achieved false discovery rate can be much lower than the desired level α. This was discussed in Section 2.2. Additional simulation results in Appendix A, where there were no true simultaneous signals and where the two sequences had different numbers of signals, lead to similar conclusions.

4.3. Known null distributions

In many standard simultaneous signal detection problems, the null distributions of the T_{id} are known. A number of methods already exist to address these cases, such as those described above in Section 4.2. It is interesting to compare them

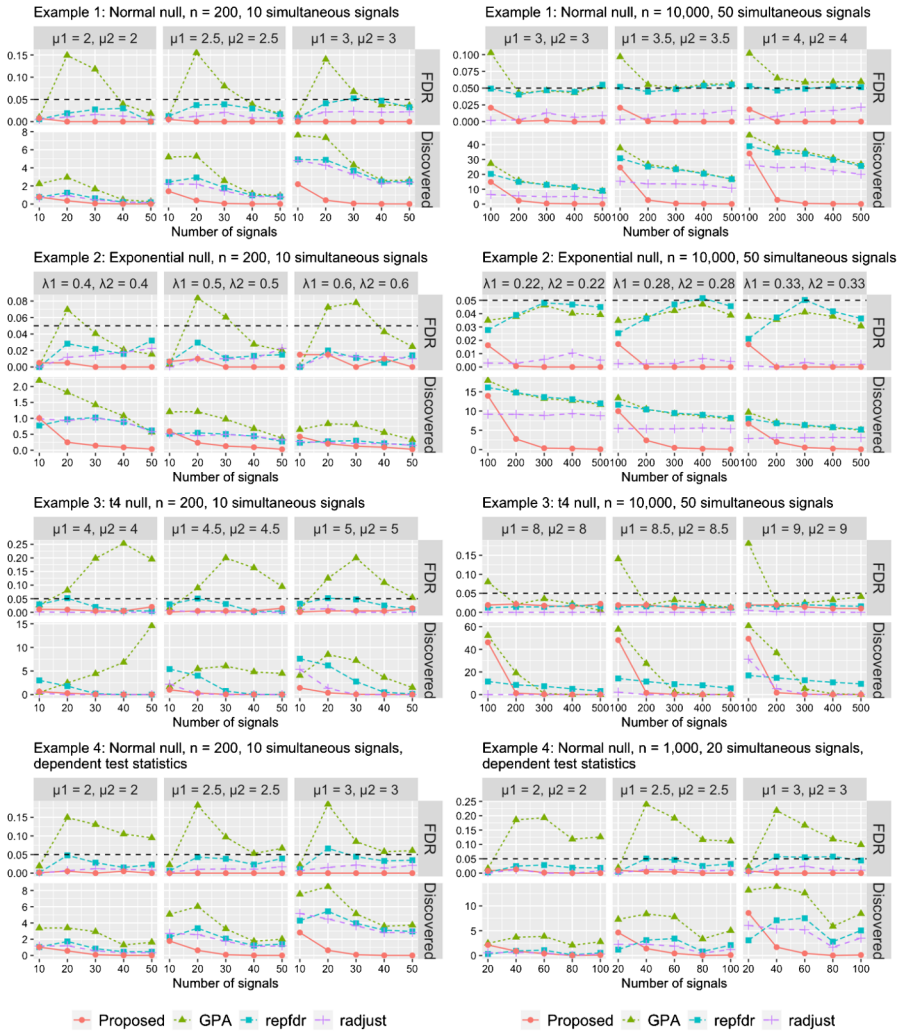

FIG 4. *False discovery rates and average numbers of discovered simultaneous signals when null distributions are known. Horizontal dashed lines mark the nominal* $\alpha = 0.05$ *false discovery rate level. Proposed: proposed approach* (7)*; GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

in these standard settings to the performance of the proposed method. The simulations in this section explore this for $D = 2$ sequences with a non-zero number of simultaneous signals. As in Section 4.2, each sequence had the same number of signals.

Figure 4 reports the false discovery rates and average numbers of discovered simultaneous signals over 200 replications of the following simulation settings. The first three examples are similar to those in Section 4.1 except that in this section, the null distributions of the T_{id} were set to be equal for both sequences.

Example 1 (normal null). This follows Example 1 of Section 4.1, except that $T_{id} = Z_{id}^2$ where the Z_{id} were independently drawn from $\mathrm{N}(\mu_{id}, 1^2)$.

Example 2 (exponential null). This follows Example 2 of Section 4.1 except that $T_{id} \sim \mathrm{Exp}(\lambda_{id})$ where $\lambda_{id} = 2$ for $I_{id} = 0$.

Example 3 (t_4 null). This follows Example 3 of Section 4.1 except that $T_{id} = Y_{id}^2$ where the Y_{id} were independently generated from $\mu_{id} + t_4$.

Example 4 (dependent test statistics). For sequences $d = 1, 2$, $T_{id} = Z_{id}^2$ where the Z_{id} were generated as in Example 1 of this section, except that all Z_{id} within a sequence d were correlated. Their correlation was set equal to the empirical correlation matrix of n genes chosen from a gene expression study of multiple myeloma [43]. This violates the assumption in model (3) of independence between test statistics in a sequence.

The procedure of Chung et al. [13] discovered the largest number of simultaneous signals, but could not maintain the nominal false discovery rate even though the null distributions of the T_{id} were known. The remaining methods were able to maintain the nominal false discovery rate throughout, even when test statistics were dependent. The methods of Heller and Yekutieli [24] and Bogomolov and Heller [9] performed similarly, with the former having somewhat higher power. The proposed method, in many cases, had comparable power when the marginal signal proportions were very low, though as previously mentioned it had very low power otherwise. It was able to control the false discovery rate even when test statistics were dependent in Example 4, though this likely was due in part to the procedure's conservativeness. As before, additional simulations in Appendix A lead to similar conclusions.

4.4. More than two sequences of test statistics

The generalized discovery procedure (8) in Section 2.5 was applied to $D = 3$ sequences with $n = 10{,}000$ features. In sequences $d = 1, 2$, $T_{id} = Z_{id}^2$ where the Z_{id} were independently generated from $\mathrm{N}(\mu_{id}, 1)$, with $\mu_{id} = 0$ when $I_{id} = 0$ and otherwise drawn from $\mathrm{N}(5, 1)$ and fixed across replications. In the third sequence, T_{i3} was generated from a complicated distribution meant to model data from ChIP-seq experiments [29], such as the example studied in Section 5. The proposed procedure is the only one applicable to this setting due to the unknown null distribution of the T_{i3}.

First, $\lambda_{i1} = \lambda_{i2}$ were drawn from $\mathrm{N}(100, 5)$ when $I_{i3} = 0$ and then fixed across replications. These model population average ChIP-seq peak heights at genomic location i under experimental and control conditions, respectively, that are equal under the null hypothesis. When $I_{i3} = 1$, λ_{i1} and λ_{i2} were independently drawn from $\mathrm{Exp}(0.001)$, modeling differences in average peak heights between the experimental conditions under the alternative hypothesis. Next, O_{il} for $l = 1, 2$ were generated from $\mathrm{Poisson}(\lambda_{il})$, modeling observed ChIP-seq peak counts. Finally, $T_{i3} = |\log(O_{i1}/O_{i2})|$, and will tend to be larger when $I_{i3} = 1$ because $\lambda_{i1} \neq \lambda_{i2}$.

TABLE 1
False discovery rates and average numbers of discovered simultaneous signals for $D = 3$ sequences of test statistics, using the proposed approach (8) with a nominal 0.05 false discovery rate level.

	$\mu_1 = 2, \mu_2 = 2$		$\mu_1 = 2.5, \mu_2 = 2.5$		$\mu_1 = 3, \mu_2 = 3$	
Marginal signals	50	100	50	100	50	100
FDR	0.00	0.00	0.00	0.00	0.00	0.00
Discoveries	4.39	0.08	12.25	0.08	20.56	0.08

Table 1 reports the results over 200 replications and shows that the proposed procedure maintained the nominal false discovery rate while still being able to detect a significant proportion of the true simultaneous signals. That the attained false discovery rates are much lower than the nominal 0.05 indicates that the procedure is conservative, as discussed in Section 2.5.

5. Data analysis

The field of sociogenomics studies molecular correlates of social behavior [34]. Saul et al. [38] studied the transcriptomic response to social challenge in mice that were exposed to intruder mice introduced to their cages. At 30, 60, and 120 minutes after intruder removal, they collected RNA-seq data from the amygdala, frontal cortex, and hypothalamus in order to determine which genes were differentially expressed between mice exposed to the intruder and mice exposed to a nonsocial control condition. They also collected ChIP-seq H3K27ac data at 30 and 120 minutes, to identify regions of chromatin that were differentially accessible between experimental and control mice. These data are available from the Gene Expression Omnibus under accession number GSE80345.

This section analyzes these data to find mouse genes that are both differentially expressed and next to differentially accessible regions of chromatin. Integrating these pieces of evidence can identify genes whose expression changes may be directly caused by differential binding of transcription factors to nearby regions of DNA [38]. This analysis can be cast as a simultaneous signal detection problem. Each mouse gene constitutes a genomic feature i, which can be associated with both a differential expression test statistic T_{i1} and a test statistic T_{i2} for the differential accessibility of a neighboring region of chromatin. The goal is to identify genes whose T_{i1} and T_{i2} are simultaneously non-null.

Following Saul et al. [38], the $T_{i1} = Z_{i1}^2$ where Z_{i1} were standard z-scores obtained using the edgeR software package [35]. Defining T_{i2} was more involved. Methods exist for calculating differential accessibility test statistics for genomic regions using ChIP-seq data [22, 42, 54], but these first identify regions of interest from the same data that the test statistics come from. This makes accurate p-values difficult to calculate [12]. This analysis takes a simple approach and by defining $T_{i2} = |\log(O_{i1}/O_{i2})|$, where O_{i1} and O_{i2} were the observed number of H3K27ac reads, in the experimental and control sample respectively, within 100 kb up- and down-stream of the ith gene.

The null distribution of T_{i2} is highly nontrivial, and the proposed method (7) is the only existing false discovery rate control procedures that can be used

TABLE 2
Mouse genes found to be both differentially expressed and next to differentially accessible chromatin at a nominal false discovery rate of 0.1.

Amygdala		Frontal cortex		Hypothalamus	
30 min	120 min	30 min	120 min	30 min	120 min
	Klk6	Nts		Ai606473	Lhx9
				Foxg1	
				Gpr88	
				Meis2	
				Penk	
				Slc5a7	

without knowledge of this null. Table 2 presents the genes identified at a nominal false discovery rate of 0.1. It indicates that the hypothalamus is the most transcriptionally responsive to social challenge, particularly at 30 minutes. A number of these genes have been previously implicated in mouse behavior. For example, mice without Gpr88 and Penk have been shown to exhibit low anxiety and resistance to mild stress [25, 26], and Foxg1 was highlighted in Saul et al. [38] as providing evidence for the role of neuropeptide signaling and neuron differentiation. These findings raise novel mechanistic hypotheses about the molecular response to social challenge.

6. Discussion

Most of this paper has assumed that the test statistics are independent across features. In the single-sequence false discovery rate control problem with dependent test statistics, an important step is to estimate parameters of the null distribution of the test statistics rather than using the theoretical null [17, 19, 41]. The nonparametric false discovery rate bound (6) already uses empirical distribution estimates, so the proposed procedure may also be able to control the false discovery rate under dependence. More work is required to fully characterize the behavior of the proposed method with dependent test statistics.

In some cases the two sequences of p-values are not of equal importance, as in replicability analysis [8, 9, 23, 24], which distinguishes between a primary versus a follow-up study. The proposed method makes no such distinction, but could be potentially be modified. Suppose for two sequences that sequence 2 were of greater interest. Then the rejection region could be defined as $[t, \infty) \times [ct, \infty)$ for some fixed constant $0 < c < 1$. This may allow weaker signals to be captured from the more important study.

A major outstanding issue is the suboptimal power of the proposed method. In some problems, with large numbers of test statistic sequences and/or sequences with moderate or high numbers of signals, the key bounds on the expected proportion of false positives in Propositions 1 and 2 are very loose. This is the tradeoff the method's robustness to unknown test statistic null distributions. Developing nonparametric detection methods with good power in these settings is an important direction for future work.

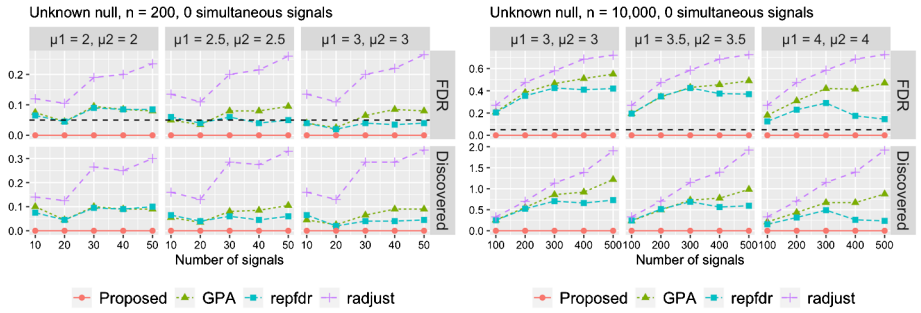

FIG 5. *False discovery rates and average numbers of discovered simultaneous signals when T_{id} has an unknown null distribution and there are no simultaneous signals. Horizontal dashed lines mark the nominal $\alpha = 0.05$ false discovery rate level. GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

Acknowledgments

The authors would like to thank the Associate Editor and the two anonymous referees for their exceedingly careful and useful reviews, which have substantially improved this manuscript. The authors would also like to thank Dr. Michael Saul for providing the data from the mouse sociogenomic experiment. The work of Dave Zhao was funded in part by NSF grant DMS-1613005.

Appendix A: Additional simulation results

Figures 5 and 6 report simulation results with unknown and known null distributions, respectively, when the number of true simultaneous signals was zero. The simulation settings are described in Sections 4.2 and 4.3. The proposed method is always able to maintain the false discovery rate at or below the nominal $\alpha = 0.05$ level, and is the only one capable of achieving this even when the null distributions are unknown, as in Figure 3.

Figures 7 and 8 report simulation results with unknown and known null distributions, respectively, when the two sequences have different numbers of signals. The simulation settings also follow those in Sections 4.2 and 4.3. The results show similar trends compared to results from settings with equal marginal signal proportions, reported in Figures 3 and 4 in the main text.

Appendix B: Proof of Proposition 1

By definition,

$$\begin{aligned} &S_1(t_1)S_2(t_2) \\ =&(1-\pi_1)(1-\pi_2)S_1^0(t_1)S_2^0(t_2) + (1-\pi_1)\pi_2 S_1^0(t_1)S_2^1(t_2) + \\ &\pi_1(1-\pi_2)S_1^1(t_1)S_2^0(t_2) + \pi_1\pi_2 S_1^1(t_1)S_2^1(t_2) \end{aligned}$$

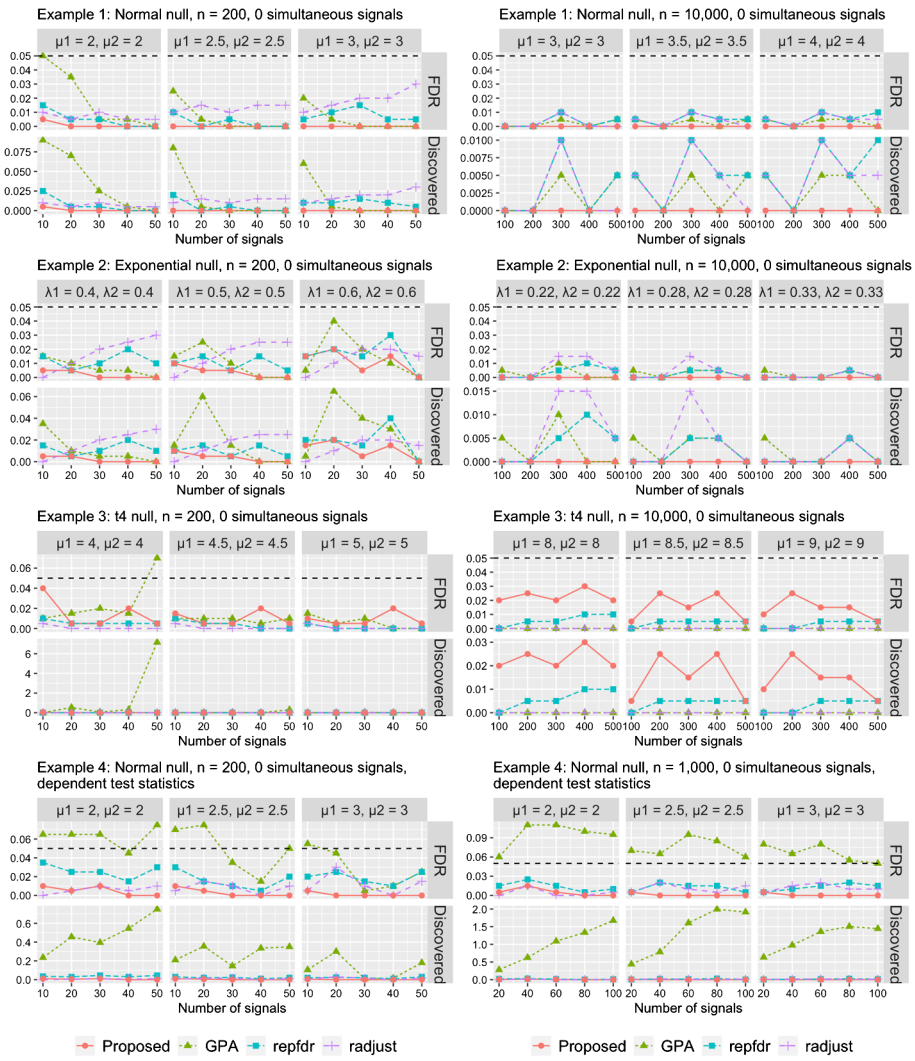

FIG 6. *False discovery rates and average numbers of discovered simultaneous signals when null distributions are known and there are no simultaneous signals. Horizontal dashed lines mark the nominal* $\alpha = 0.05$ *false discovery rate level. Proposed: proposed approach* (7)*; GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

$$
\begin{aligned}
&= (1-\pi_1-\pi_2)S_1^0(t_1)S_2^0(t_2) + \\
&\quad \pi_1\pi_2\{S_1^0(t_1)S_2^0(t_2) + S_1^1(t_1)S_2^1(t_2) - S_1^0(t_1)S_2^1(t_2) - S_1^1(t_1)S_2^0(t_2)\} + \\
&\quad \pi_2 S_1^0(t_1)S_2^1(t_2) + \pi_1 S_1^1(t_1)S_2^0(t_2).
\end{aligned}
$$

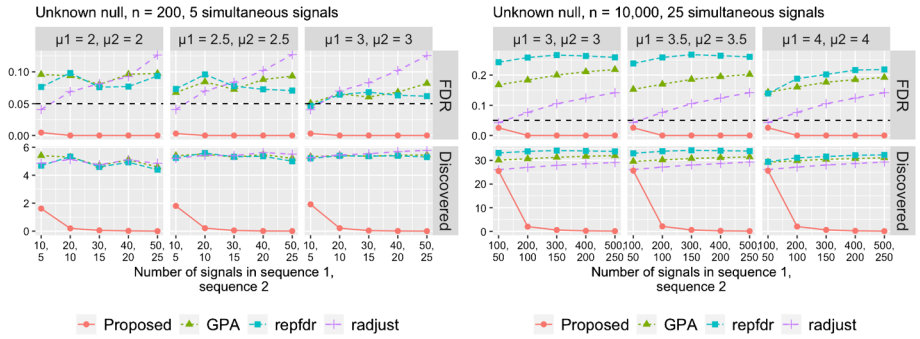

FIG 7. *False discovery rates and average numbers of discovered simultaneous signals when T_{id} has an unknown null distribution and the two sequences have different numbers of signals. Horizontal dashed lines mark the nominal $\alpha = 0.05$ false discovery rate level. GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

Define the proportion $\pi_{\mathbf{I}} = |\{i : \mathbf{I}_i = \mathbf{I}\}|/n$. Then $\pi_1 = \pi_{(1,0)} + \pi_{(1,1)}$ and $\pi_2 = \pi_{(0,1)} + \pi_{(1,1)}$, so the above expression becomes

$$\begin{aligned}
&S_1(t_1)S_2(t_2)\\
&= (\pi_{(0,0)} - \pi_{(1,1)})S_1^0(t_1)S_2^0(t_2) + (\pi_{(0,1)} + \pi_{(1,1)})S_1^0(t_1)S_2^1(t_2) +\\
&\quad (\pi_{(1,0)} + \pi_{(1,1)})S_1^1(t_1)S_2^0(t_2) + \pi_1\pi_2\{S_1^1(t_1) - S_1^0(t_1)\}\{S_2^1(t_2) - S_2^0(t_2)\}\\
&= \pi_{(0,0)}S_1^0(t_1)S_2^0(t_2) + \pi_{(0,1)}S_1^0(t_1)S_2^1(t_2) + \pi_{(1,0)}S_1^1(t_1)S_2^0(t_2) +\\
&\quad \pi_{(1,1)}[S_1^0(t_1)S_2^1(t_2) + \{S_1^1(t_1) - S_1^0(t_1)\}S_2^0(t_2)] +\\
&\quad \pi_1\pi_2\{S_1^1(t_1) - S_1^0(t_1)\}\{S_2^1(t_2) - S_2^0(t_2)\}.
\end{aligned}$$

Since $S_1^1(t_1) > S_1^0(t_1)$ by the stochastic ordering in Assumption 1, the last two lines of the previous display are always positive. The result follows because

$$\begin{aligned}
&\pi_{(0,0)}S_1^0(t_1)S_2^0(t_2) + \pi_{(0,1)}S_1^0(t_1)S_2^1(t_2) + \pi_{(1,0)}S_1^1(t_1)S_2^0(t_2)\\
&= n^{-1}\sum_{i\in\mathcal{S}^c} S_1^{I_{i1}}(t_1)S_2^{I_{i2}}(t_2).
\end{aligned}$$

Appendix C: Proof of Theorem 1

Define $\mathcal{R}(t) = \{i : I(T_{i1} \geq t, T_{i2} \geq t)\}$ to be the set of features rejected at threshold t. Then

$$\mathrm{FDR}(\hat{\delta}_\rho) = \sum_{i\in\mathcal{S}^c\cap\mathcal{R}(\hat{t}_\rho)} E\frac{I(T_{i1} \geq \hat{t}_\rho, T_{i2} \geq \hat{t}_\rho)}{\max\{1, \sum_{i=1}^n I(T_{i1} \geq \hat{t}_\rho, T_{i2} \geq \hat{t}_\rho)\}},$$

with $\mathcal{S}$ defined as in (1). Since $\hat{t}_\rho$ satisfies $\widehat{\mathrm{FDR}}_\rho(\hat{t}_\rho) \leq \alpha$,

$$\frac{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho}{\max\{n^{-1}, n^{-1}\sum_{i=1}^n I(T_{i1} \geq \hat{t}_\rho, T_{i2} \geq \hat{t}_\rho)\}} \leq \alpha$$

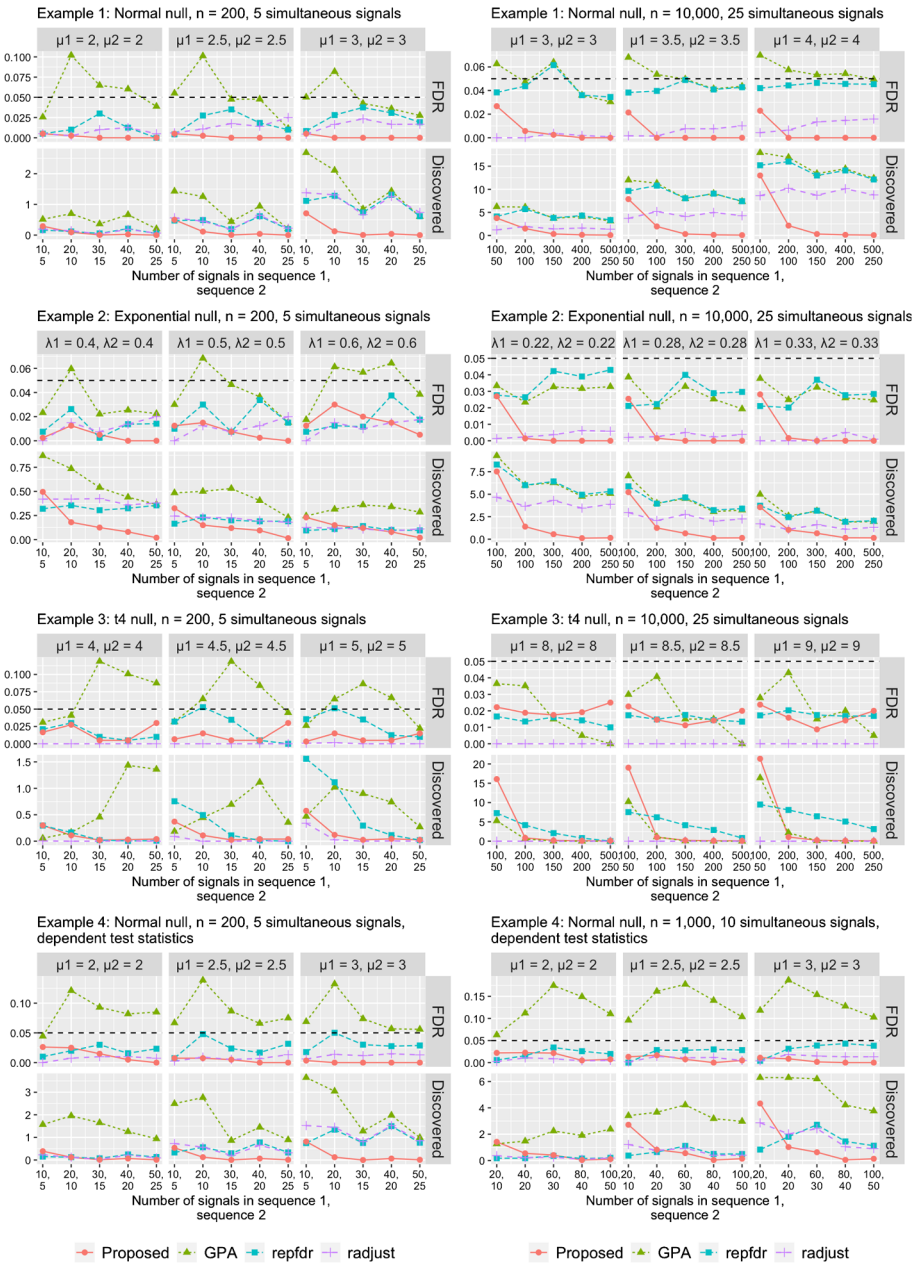

FIG 8. *False discovery rates and average numbers of discovered simultaneous signals when null distributions are known and the two sequences have different numbers of signals. Horizontal dashed lines mark the nominal* $\alpha = 0.05$ *false discovery rate level. Proposed: proposed approach* (7)*; GPA: method of Chung et al. [13]; repfdr: method of Heller and Yekutieli [24]; radjust: method of Bogomolov and Heller [9].*

by the definition of $\widehat{\text{FDR}}_\rho(\hat{t}_\rho)$ in (6). Therefore,

$$\text{FDR}(\hat{\delta}_\rho) \le \frac{\alpha}{n} \sum_{i \in \mathcal{S}^c \cap \mathcal{R}(\hat{t}_\rho)} E \frac{I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho}.$$

Analogous to the proposed procedure (7), define the constrained optimization problem

$$\hat{t}_\rho^{(-i)} = \inf \left[t \in [0, \infty) : \frac{\prod_{d=1}^2 n^{-1}\{\sum_{j \ne i} I(T_{jd} \ge t) + 1\} + \rho}{\max[n^{-1}, n^{-1}\{\sum_{j \ne i} I(T_{j1} \ge t, T_{j2} \ge t) + 1\}]} \le \alpha \right]. \tag{11}$$

This type of leave-one-out construction of $\hat{t}_\rho^{(-i)}$ has also been used in proofs of false discovery rate control in a single sequence of test statistics [6, 32, 37].

For any feature $i \in \mathcal{R}(\hat{t}_\rho)$, $I(T_{id} \ge \hat{t}_\rho) = 1$ and $I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho) = 1$, so

$$\begin{aligned} &\frac{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho}{\max\{n^{-1}, n^{-1}\sum_{i=1}^n I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho)\}} \\ &= \frac{\prod_{d=1}^2 n^{-1}\{\sum_{j \ne i} I(T_{jd} \ge \hat{t}_\rho) + 1\} + \rho}{\max[n^{-1}, n^{-1}\{\sum_{j \ne i} I(T_{j1} \ge \hat{t}_\rho, T_{j2} \ge \hat{t}_\rho) + 1\}]}. \end{aligned}$$

This means that $\hat{t}_\rho$ is feasible for problem (11), so $\hat{t}_\rho^{(-i)} \le \hat{t}_\rho$. Next, this in turn implies that since $i \in \mathcal{R}(\hat{t}_\rho)$, $1 = I(T_{id} \ge \hat{t}_\rho) \le I(T_{id} \ge \hat{t}_\rho^{-(i)}) \le 1$ and $1 = I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho) \le I(T_{i1} \ge \hat{t}_\rho^{-(i)}, T_{i2} \ge \hat{t}_\rho^{-(i)}) \le 1$, hence, $I(T_{id} \ge \hat{t}_\rho) = I(T_{id} \ge \hat{t}_\rho^{-(i)}) = 1$ and $I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho) = I(T_{i1} \ge \hat{t}_\rho^{-(i)}, T_{i2} \ge \hat{t}_\rho^{-(i)}) = 1$, so

$$\begin{aligned} &\frac{\hat{S}_1(\hat{t}_\rho^{(-i)})\hat{S}_2(\hat{t}_\rho^{(-i)}) + \rho}{\max\{n^{-1}, n^{-1}\sum_{i=1}^n I(T_{i1} \ge \hat{t}_\rho^{(-i)}, T_{i2} \ge \hat{t}_\rho^{(-i)})\}} \\ &= \frac{\prod_{d=1}^2 n^{-1}\{\sum_{j \ne i} I(T_{jd} \ge \hat{t}_\rho^{(-i)}) + 1\} + \rho}{\max[n^{-1}, n^{-1}\{\sum_{j \ne i} I(T_{j1} \ge \hat{t}_\rho^{(-i)}, T_{j2} \ge \hat{t}_\rho^{(-i)}) + 1\}]}, \end{aligned}$$

which is at most α by construction of $\hat{t}_\rho^{(-i)}$. Thus $\hat{t}_\rho^{(-i)}$ is feasible for problem (7) and $\hat{t}_\rho \le \hat{t}_\rho^{(-i)}$.

The previous results imply that $\hat{t}_\rho^{(-i)} = \hat{t}_\rho$ for $i \in \mathcal{R}(\hat{t}_\rho)$. Then

$$\begin{aligned} \text{FDR}(\hat{\delta}_\rho) &\le \frac{\alpha}{n} \sum_{i \in \mathcal{S}^c \cap \mathcal{R}(\hat{t}_\rho)} E \frac{I(T_{i1} \ge \hat{t}_\rho, T_{i2} \ge \hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \\ &\le \frac{\alpha}{n} \sum_{i \in \mathcal{S}^c \cap \mathcal{R}(\hat{t}_\rho)} E \frac{I(T_{i1} \ge \hat{t}_\rho^{(-i)}, T_{i2} \ge \hat{t}_\rho^{(-i)})}{\hat{S}_1(\hat{t}_\rho^{(-i)})\hat{S}_2(\hat{t}_\rho^{(-i)}) + \rho} \\ &\le \frac{\alpha}{n} \sum_{i \in \mathcal{S}^c \cap \mathcal{R}(\hat{t}_\rho)} E \frac{I(T_{i1} \ge \hat{t}_\rho^{(-i)}, T_{i2} \ge \hat{t}_\rho^{(-i)})}{\prod_{d=1}^2 \{n^{-1}\sum_{j \ne i} I(T_{jd} \ge \hat{t}_\rho^{(-i)}) + 1\} + \rho}, \end{aligned}$$

where the third line follows because it was shown above that $i \in \mathcal{R}(\hat{t}_\rho)$ implies $i \in \mathcal{R}(\hat{t}_\rho^{(-i)})$. Since neither $\hat{t}_\rho^{(-i)}$ nor the denominator of the final expression depends on (T_{i1}, T_{i2}), and because the T_{id} are independent across sequences d, for every $i \in \mathcal{R}(\hat{t}_\rho)$

$$
\begin{aligned}
&E\frac{I(T_{i1} \geq \hat{t}_\rho^{(-i)}, T_{i2} \geq \hat{t}_\rho^{(-i)})}{\prod_{d=1}^{2}\{n^{-1}\sum_{j\neq i} I(T_{jd} \geq \hat{t}_\rho^{(-i)}) + 1\} + \rho} \\
&= E\frac{S_{i1}^{I_{i1}}(\hat{t}_\rho^{(-i)}) S_{i2}^{I_{i2}}(\hat{t}_\rho^{(-i)})}{\prod_{d=1}^{2}\{n^{-1}\sum_{j\neq i} I(T_{jd} \geq \hat{t}_\rho^{(-i)}) + 1\} + \rho} = E\frac{S_{i1}^{I_{i1}}(\hat{t}_\rho^{(-i)}) S_{i2}^{I_{i2}}(\hat{t}_\rho^{(-i)})}{\hat{S}_1(\hat{t}_\rho^{(-i)})\hat{S}_2(\hat{t}_\rho^{(-i)}) + \rho}.
\end{aligned}
$$

Then again because $\hat{t}_\rho = \hat{t}_\rho^{(-i)}$ on $\mathcal{R}(\hat{t}_\rho)$,

$$
\text{FDR}(\hat{\delta}_\rho) \leq \frac{\alpha}{n} \sum_{i \in \mathcal{S}^c \cap \mathcal{R}(\hat{t}_\rho)} E\frac{S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \leq \alpha E\frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho}.
$$

It remains to show that

$$
\limsup_{n\to\infty} E\frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \leq 1.
$$

By the Fatou-Lebesgue theorem, it suffices to show that

$$
\limsup_{n\to\infty} \frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \leq 1
$$

almost surely. Define $\mathcal{N} = \{(0,0),(0,1),(1,0)\}$. Then the left-hand expression can be rewritten as

$$
\begin{aligned}
&\limsup_{n\to\infty} \frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \\
&\leq 1 + \limsup_{n\to\infty} \frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho) - \sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(\hat{t}_\rho) S_2^{I_2}(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} + \qquad (12) \\
&\limsup_{n\to\infty} \frac{\sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(\hat{t}_\rho) S_2^{I_2}(\hat{t}_\rho) - S_1(\hat{t}_\rho)S_2(\hat{t}_\rho) - \rho}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} + \qquad (13) \\
&\limsup_{n\to\infty} \frac{S_1(\hat{t}_\rho)S_2(\hat{t}_\rho) - \hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho}, \qquad (14)
\end{aligned}
$$

with $\pi_{\mathbf{I}}$ and S_d defined as in Assumption 2.

First, the second term of (12) obeys

$$
\limsup_{n\to\infty} \left| \frac{n^{-1}\sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(\hat{t}_\rho) S_{i2}^{I_{i2}}(\hat{t}_\rho) - \sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(\hat{t}_\rho) S_2^{I_2}(\hat{t}_\rho)}{\hat{S}_d(\hat{t}_\rho)\hat{S}_d(\hat{t}_\rho) + \rho} \right|
$$

$$\leq \frac{1}{\rho} \lim_{n\to\infty} \sup_{t\in[0,\infty)} \left| n^{-1} \sum_{i\in\mathcal{S}^c} S_{i1}^{I_{i1}}(t) S_{i2}^{I_{i2}}(t) - \sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(t) S_2^{I_2}(t) \right| = 0,$$

almost surely, by Assumption 2. Next, the numerator of (13) satisfies

$$\sup_{t\in[0,\infty)} \left\{ \sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(t) S_2^{I_2}(t) - S_1(t)S_2(t) - \rho \right\} < 0$$

by Proposition 1, and because $\rho > 0$,

$$\limsup_{n\to\infty} \frac{\sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(\hat{t}_\rho) S_2^{I_2}(\hat{t}_\rho) - S_1(\hat{t}_\rho)S_2(\hat{t}_\rho) - \rho}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \leq 0$$

almost surely. It remains to show that (14) goes to zero. Since

$$\limsup_{n\to\infty} \left| \frac{S_1(\hat{t}_\rho)S_2(\hat{t}_\rho) - \hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho)}{\hat{S}_1(\hat{t}_\rho)\hat{S}_2(\hat{t}_\rho) + \rho} \right| \leq \frac{1}{\rho} \lim_{n\to\infty} \sup_{t\in[0,\infty)} \left| S_1(t)S_2(t) - \hat{S}_1(t)\hat{S}_2(t) \right|$$

and the $\hat{S}_d(t)$ are averages of independent but non-identically distributed terms that satisfy the conditions of Theorem 8.3 of Pollard [31],

$$\lim_{n\to\infty} \sup_{t\in[0,\infty)} \left| \hat{S}_d(t) - \frac{1}{n} \sum_{i=1}^{n} S_{id}^{I_{id}}(t) \right| = 0$$

almost surely for all d. By Assumption 2,

$$\lim_{n\to\infty} \sup_{t\in[0,\infty)} \left| \frac{1}{n} \sum_{i=1}^{n} S_{id}^{I_{id}}(t) - S_d(t) \right| = 0,$$

where $S_d(t) = (1-\pi_d)S_d^0(t) + \pi_d S_d^1(t)$ is defined in Proposition 1. Therefore

$$\begin{aligned} \lim_{n\to\infty} \sup_{t\in[0,\infty)} |S_1(t)S_2(t) - \hat{S}_1(t)\hat{S}_2(t)| &\leq \lim_{n\to\infty} \sup_{t\in[0,\infty)} |S_1(t)S_2(t) - \hat{S}_1(t)S_2(t)| + \\ &\quad \lim_{n\to\infty} \sup_{t\in[0,\infty)} |\hat{S}_1(t)S_2(t) - \hat{S}_1(t)\hat{S}_2(t)| \\ &\leq \lim_{n\to\infty} \sum_{d=1}^{2} \sup_{t\in[0,\infty)} |S_d(t) - \hat{S}_d(t)| = 0 \end{aligned}$$

almost surely. This concludes the proof.

Appendix D: Proof of Theorem 2

Define

$$\begin{aligned} V_{ab}(t_1,t_2) &= \sum_{i:I_{i1}=a, I_{i2}=b} I(T_{i2} \geq t_1, T_{i2} \geq t_2), \quad a,b = 0,1, \\ R(t_1,t_2) &= \sum_i I(T_{i1} \geq t_1, T_{i2} \geq t_2). \end{aligned} \tag{15}$$

Then the true false discovery rate attained by a discovery rule of the form $\delta(T_{i1}, T_{i2}) = I(T_{i1} \geq t_1, T_{i2} \geq t_2)$ can be written as

$$\text{FDR}(\delta) = E\left[\frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}}\right].$$

It will first be shown that for any $\eta_1, \eta_2 < \infty$,

$$\lim_{n\to\infty} \inf_{t_1\leq\eta_1, t_2\leq\eta_2} \left[\widetilde{\text{FDR}}_\rho(t_1,t_2) - \frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}}\right] \geq 0 \tag{16}$$

almost surely, for $\widetilde{\text{FDR}}_\rho(t_1,t_2)$ defined in (9). Next it will be shown that

$$\sup_{t_1\leq\eta_1, t_2\leq\eta_2} \left|\frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}} - \text{FDR}(\delta)\right| \to 0, \tag{17}$$

almost surely, which will complete the proof.

To show (16), it suffices to show

$$\lim_{n\to\infty} \inf_{t_1\leq\eta_1, t_2\leq\eta_2} [\hat{S}_1(t_1)\hat{S}_2(t_2) - n^{-1}\{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)\}] \geq 0$$

almost surely. Arguments from the proof of Theorem 1 can be used to show that

$$\lim_{n\to\infty} \sup_{t_1,t_2\in[0,\infty)} |\hat{S}_1(t_1)\hat{S}_2(t_2) - S_1(t_1)S_2(t_2)| = 0,$$

$$\lim_{n\to\infty} \sup_{t_1,t_2\in[0,\infty)} |G^0(t_1,t_2) - n^{-1}\{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)\}| = 0$$

almost surely, where $G^0(t_1,t_2) = \sum_{\mathbf{I}\in\mathcal{N}} \pi_{\mathbf{I}} S_1^{I_1}(t_1) S_2^{I_2}(t_2)$ with $\mathcal{N}$ defined as $\{(0,0),(0,1),(1,0)\}$. Combining these with Proposition 1 proves (16).

To prove (17), define

$$G(t_1,t_2) = G^0(t_1,t_2) + \pi_{(1,1)} G^1(t_1,t_2) \tag{18}$$

for $G^1(t_1,t_2)$ from in Assumption 2. Then

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{t_1\leq\eta_1, t_2\leq\eta_2} \left|\frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)}\right| \\
&\leq \lim_{n\to\infty} \frac{n}{\max\{1, R(\eta_1,\eta_2)\}} \times \\
&\quad \sup_{t_1\leq\eta_1, t_2\leq\eta_2} \left|n^{-1}\{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)\} - G^0(t_1,t_2)\right| + \\
&\quad \lim_{n\to\infty} \frac{n}{\max\{1, R(\eta_1,\eta_2)\}} \frac{1}{G(\eta_1,\eta_2)} \times \\
&\quad \sup_{t_1\leq\eta_1, t_2\leq\eta_2} \left|G(t_1,t_2) - n^{-1}\max\{1, R(t_1,t_2)\}\right|.
\end{aligned}$$

Arguments from the proof of Theorem 1 can be used to show that both terms on the right-hand side equal zero almost surely. Next, the dominated convergence theorem implies that

$$\begin{aligned}
0 &= E \lim_{n\to\infty} \times \\
&\quad \sup_{t_1\le\eta_1, t_2\le\eta_2} \left| \frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right| \\
&= \lim_{n\to\infty} E \times \\
&\quad \sup_{t_1\le\eta_1, t_2\le\eta_2} \left| \frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}} - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right| \\
&\ge \lim_{n\to\infty} \sup_{t_1\le\eta_1, t_2\le\eta_2} \left| \mathrm{FDR}(\delta) - \frac{G^0(t_1,t_2)}{G(t_1,t_2)} \right|.
\end{aligned}$$

Combining these results proves (17).

Appendix E: Proof of Theorem 3

The theorem is trivially true when $(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) = (\infty, \infty)$. Otherwise, suppose there exist fixed $\eta_1, \eta_2 < \infty$ such that $\hat{t}_{\rho 1} \le \eta_1$ and $\hat{t}_{\rho 2} \le \eta_2$ with probability 1. Then by (16) from the proof of Theorem 2,

$$\begin{aligned}
&\liminf_{n\to\infty} \left[\widetilde{\mathrm{FDR}}_\rho(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) - \frac{V_{(0,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(1,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(0,1)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})}{\max\{1, R(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})\}} \right] \\
&\ge \lim_{n\to\infty} \inf_{t_1\le\eta_1, t_2\le\eta_2} \left[\widetilde{\mathrm{FDR}}_\rho(t_1,t_2) - \right. \\
&\qquad \left. \frac{n^{-1}}{n^{-1}} \frac{V_{(0,0)}(t_1,t_2) + V_{(1,0)}(t_1,t_2) + V_{(0,1)}(t_1,t_2)}{\max\{1, R(t_1,t_2)\}} \right] \ge 0
\end{aligned}$$

almost surely. This implies that

$$\limsup_{n\to\infty} \frac{V_{(0,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(1,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(0,1)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})}{\max\{1, R(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})\}} \le \alpha$$

almost surely. Then by the Fatou-Lebesgue theorem,

$$\limsup_{p\to\infty} \mathrm{FDR}(\tilde{\delta}) \le E \limsup_{p\to\infty} \frac{V_{(0,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(1,0)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) + V_{(0,1)}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})}{\max\{1, R(\hat{t}_{\rho 1}, \hat{t}_{\rho 2})\}} \le \alpha$$

for the discovery procedure $\tilde{\delta}$ (10).

It remains to construct η_1 and η_2. The pointwise limit of $\widetilde{\mathrm{FDR}}_\rho(t_1,t_2)$ (9) is

$$\mathrm{FDR}_\rho(t_1,t_2) = \{S_1(t_1)S_2(t_2) + \rho\}/G(t_1,t_2),$$

for $G(t_1, t_2)$ defined in (18). By assumption, there exists some $\epsilon > 0$ such that $\text{FDR}_\rho(t_1', t_2') = \alpha - \epsilon$. Kolmogorov's strong law of large numbers and Slutsky's theorem show that for n sufficiently large,

$$|\widetilde{\text{FDR}}_\rho(t_1', t_2') - \text{FDR}_\rho(t_1', t_2')| \leq \epsilon/2$$

with probability 1. This implies that $\widetilde{\text{FDR}}_\rho(t_1', t_2') \leq \alpha - \epsilon/2$, so (t_1', t_2') is a feasible solution of the optimization problem (10). Then $\hat{G}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) \geq \hat{G}(t_1', t_2')$. Using arguments from the proof of Theorem 1, it can be shown that $\hat{G}(t_1, t_2)$ converges almost surely to $G(t_1, t_2)$ uniformly in (t_1, t_2). Therefore for any $\eta > 0$, there exists a sufficiently large n such that

$$G(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) \geq \hat{G}(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) - \eta/4 \geq \hat{G}(t_1', t_2') - \eta/4 \geq G(t_1', t_2') - \eta/2$$

with probability 1. Choose $\eta = G(t_1', t_2')$, which must be positive because t_1' and t_2' are both finite by assumption. This shows that $G(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) \geq \eta/2 > 0$ with probability 1. Now define η_1 such that $S_1^{-1}(\eta/2)$ and $\eta_2 = S_2^{-1}(\eta/2)$. Then

$$G(\hat{t}_{\rho 1}, \hat{t}_{\rho 2}) \geq \eta/2 = S_1(\eta_1) = G(\eta_1, 1) \geq G(\eta_1, \hat{t}_{\rho 2}),$$

which implies that $\hat{t}_{\rho 1} \leq \eta_1$ with probability 1. By similar reasoning, $\hat{t}_{\rho 2} \leq \eta_2$ with probability 1. Finally, since $\eta > 0$, η_1 and η_2 are both finite as well.

Appendix F: Proof of Proposition 2

We prove the Proposition 2 for $D = 3$. Similar arguments can be applied for cases $D \geq 4$. First, the expression

$$n^{-1} \sum_{i \in \mathcal{S}^c} S_1^{I_{i1}}(t_1) S_2^{I_{i2}}(t_2) S_3^{I_{i3}}(t_3)$$

equals

$$\begin{aligned}
&\pi_{(0,0,0)} S_1^0(t_1) S_2^0(t_2) S_3^0(t_3) + \pi_{(0,1,0)} S_1^0(t_1) S_2^1(t_2) S_3^0(t_3) + \\
&\pi_{(1,0,1)} S_1^1(t_1) S_2^0(t_2) S_3^1(t_3) + \pi_{(1,0,0)} S_1^1(t_1) S_2^0(t_2) S_3^0(t_3) + \\
&\pi_{(0,0,1)} S_1^0(t_1) S_2^0(t_2) S_3^1(t_3) + \pi_{(0,1,0)} S_1^0(t_1) S_2^1(t_2) S_3^0(t_3) + \\
&\pi_{(1,1,0)} S_1^1(t_1) S_2^1(t_2) S_3^0(t_3),
\end{aligned}$$

where $\pi_{\mathbf{I}} = |\{i : \mathbf{I}_i = \mathbf{I}\}|/n$. By the stochastic ordering in Assumption 1, $S_d^0(t) < S_d^1(t)$ for $d = 1, 2, 3$, so the previous expression is at most

$$\begin{aligned}
&\{\pi_{(0,0,0)} S_1^0(t_1) S_2^0(t_2) + \pi_{(0,1,0)} S_1^0(t_1) S_2^1(t_2) + \pi_{(1,0,1)} S_1^1(t_1) S_2^0(t_2)\} S_3^1(t_3) + \\
&\{\pi_{(1,0,0)} S_2^0(t_2) S_3^0(t_3) + \pi_{(0,0,1)} S_2^0(t_2) S_3^1(t_3) + \pi_{(0,1,0)} S_2^1(t_2) S_3^0(t_3)\} S_1^1(t_1) + \\
&\pi_{(1,1,0)} S_1^1(t_1) S_2^1(t_2) S_3^0(t_3) \\
\leq &\{\pi_{(0,0,0)} S_1^0(t_1) S_2^0(t_2) + \pi_{(0,1,0)} S_1^0(t_1) S_2^1(t_2) + \pi_{(1,0,1)} S_1^1(t_1) S_2^0(t_2)\} + \\
&\{\pi_{(1,0,0)} S_2^0(t_2) S_3^0(t_3) + \pi_{(0,0,1)} S_2^0(t_2) S_3^1(t_3) + \pi_{(0,1,0)} S_2^1(t_2) S_3^0(t_3)\} + \\
&\pi_{(1,1,0)} S_1^1(t_1) S_2^1(t_2) S_3^0(t_3)
\end{aligned}$$

$$\{\pi_{(1,0,0)}S_2^0(t_2)S_3^0(t_3) + \pi_{(0,0,1)}S_2^0(t_2)S_3^1(t_3) + \pi_{(0,1,0)}S_2^1(t_2)S_3^0(t_3)\} +$$
$$\{\pi_{(0,1,0)}S_1^0(t_1)S_3^0(t_3) + \pi_{(0,1,1)}S_1^0(t_1)S_3^1(t_3) + \pi_{(1,1,0)}S_1^1(t_1)S_3^0(t_3)\}.$$

Now define $\pi_{(I_1,I_2,\cdot)} = \pi_{(I_1,I_2,0)} + \pi_{(I_1,I_2,1)}$ for $I_1, I_2 \in \{0,1\}$, and define $\pi_{(I_1,\cdot,I_2)}$ and $\pi_{(\cdot,I_1,I_2)}$ similarly. Then the previous expression is upper-bounded by

$$\{\pi_{(0,0,\cdot)}S_1^0(t_1)S_2^0(t_2) + \pi_{(0,1,\cdot)}S_1^0(t_1)S_2^1(t_2) + \pi_{(1,0,\cdot)}S_1^1(t_1)S_2^0(t_2)\} +$$
$$\{\pi_{(\cdot,0,0)}S_2^0(t_2)S_3^0(t_3) + \pi_{(\cdot,0,1)}S_2^0(t_2)S_3^1(t_3) + \pi_{(\cdot,1,0)}S_2^1(t_2)S_3^0(t_3)\} +$$
$$\{\pi_{(0,\cdot,0)}S_1^0(t_1)S_3^0(t_3) + \pi_{(0,\cdot,1)}S_1^0(t_1)S_3^1(t_3) + \pi_{(1,\cdot,0)}S_1^1(t_1)S_3^0(t_3)\}.$$

Applying Proposition 1 to each of these terms gives the desired result.

References

[1] Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., Sklar, P., Disorder, T. P. G. C. P. B., Groups, S. W., Roddey, J. C., Chen, C.-H., McEvoy, L., Desikan, R. S., Djurovic, S. and Dale, A. M. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics* **9** e1003455.

[2] Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics* **11** 1983–2001. MR3651021

[3] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085. MR3375876

[4] Barber, R. F. and Candes, E. J. (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint* arXiv:1602.03574. MR3988764

[5] Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1247–1268. MR3689317

[6] Benditkis, J., Heesen, P. and Janssen, A. (2018). The false discovery rate (FDR) of multiple tests in a class room lecture. *Statistics & Probability Letters* **134** 29–35. MR3758578

[7] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 289–300. MR1325392

[8] Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association* **108** 1480–1492. MR3174723

[9] Bogomolov, M. and Heller, R. (2018). Assessing replicability of findings across two studies of multiple features. *Biometrika* **105** 505–516. MR3842881

[10] Candes, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 551–577. MR3798878

[11] Chi, Z. (2008). False discovery rate control with multivariate p-values. *Electronic Journal of Statistics* **2** 368–411. MR2411440

[12] Chitpin, J. G., Awdeh, A. and Perkins, T. J. (2018). RECAP reveals the true statistical significance of ChIP-seq peak calls. *bioRxiv* 260687.

[13] Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics* **10** e1004787.

[14] Du, L. and Zhang, C. (2014). Single-index modulated multiple testing. *The Annals of Statistics* **42** 30–79. MR3226157

[15] Dudoit, S., van der Laan, M. J. and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology* **3** 1–69. MR2101462

[16] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99** 96–104. MR2054289

[17] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102** 93–103. MR2293302

[18] Efron, B. (2010a). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* **1**. Cambridge University Press. MR2724758

[19] Efron, B. (2010b). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* **105** 1042–1055. MR2752597

[20] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 499–517. MR1924303

[21] Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B. and Roussos, P. (2018). A Bayesian Framework for Multiple Trait Colocalization from Summary Association Statistics. *Bioinformatics* **1** 8.

[22] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38** 576–589.

[23] Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences* **111** 16262–16267.

[24] Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics* **8** 481–498. MR3191999

[25] Meirsman, A. C., Le Merrer, J., Pellissier, L. P., Diaz, J., Clesse, D., Kieffer, B. L. and Becker, J. A. (2016). Mice lacking GPR88 show motor deficit, improved spatial learning, and low anxiety reversed by delta opioid antagonist. *Biological Psychiatry* **79** 917–927.

[26] Melo, I., Drews, E., Zimmer, A. and Bilkei-Gorzo, A. (2014). Enkephalin knockout male mice are resistant to chronic mild stress. *Genes, Brain and Behavior* **13** 550–558.

[27] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013a). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45** 984–994.

[28] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013b). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381** 1371.

[29] Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10** 669.

[30] Phillips, D. and Ghosh, D. (2014). Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics. *The Annals of Applied Statistics* **8** 801–823. MR3262535

[31] Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics* i–86. JSTOR. MR1089429

[32] Ramdas, A., Barber, R. F., Wainwright, M. J. and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*. To appear. MR3988773

[33] Rittschof, C. C., Bukhari, S. A., Sloofman, L. G., Troy, J. M., Caetano-Anollés, D., Cash-Ahmed, A., Kent, M., Lu, X., Sanogo, Y. O., Weisner, P. A., Zhang, H., Bell, A. M., Ma, J., Sinha, S., Robinson, G. E. and Stubbs, L. (2014). Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proceedings of the National Academy of Sciences* **111** 17929–17934.

[34] Robinson, G. E., Grozinger, C. M. and Whitfield, C. W. (2005). Sociogenomics: social life in molecular terms. *Nature Reviews Genetics* **6** 257.

[35] Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

[36] Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics* **35** 1378–1408. MR2351090

[37] Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A (2008–)* 135–168. MR2551809

[38] Saul, M., Seward, C. H., Troy, J. M., Zhang, H., Sloofman, L. G.,

Lu, X., Weisner, P. A., Caetano-Anolles, D., Sun, H., Zhao, S. D., Chandrasekaran, S., Sinha, S. and Stubbs, L. (2017). Transcriptional regulatory dynamics drive coordinated metabolic and neural response to social challenge in mice. *Genome Research* gr–214221.

[39] Saul, M. C., Blatti, C., Yang, W., Bukhari, S. A., Shpigler, H. Y., Troy, J. M., Seward, C. H., Sloofman, L., Chandrasekaran, S., Bell, A. M., Stubbs, L., Robinson, G. E., Zhao, S. D. and Sinha, S. (2018). Cross-species systems analysis of evolutionary toolkits of neurogenomic response to social challenge. *Genes, Brain and Behavior* e12502.

[40] Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics* **2** 1332–1359. MR2655662

[41] Schwartzman, A. (2012). Comment: FDP vs FDR and the effect of conditioning. *Journal of the American Statistical Association* **107** 1039–1041. MR3010890

[42] Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J. and Nestler, E. J. (2013). diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PloS One* **8** e65598.

[43] Shi, L. et al. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28** 827.

[44] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 479–498. MR1924302

[45] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 187–205. MR2035766

[46] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102** 901–912. MR2411657

[47] Thompson, D., Regev, A. and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annual Review of Cell and Developmental Biology* **31** 399–428.

[48] Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H., McNeill, M., Smyth, D. J., Niblett, D., Cambien, F., Consortium, C., Tiret, L., Todd, J. A., Clayton, D. G. and Blankenberg, S. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21** 2815–2824.

[49] Westfall, P. H., Young, S. S. et al. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* **279**. John Wiley & Sons.

[50] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.

[51] Xiang, D., Zhao, S. D. and Cai, T. T. (2019). Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. to appear. MR3997098

[52] Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82** 171–196. MR1736442

[53] Yu, D., Huber, W. and Vitek, O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29** 1275–1282.

[54] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9** R137.

[55] Zhao, S. D., Cai, T. T. and Li, H. (2017). Optimal detection of weak positive latent dependence between two sequences of multiple tests. *Journal of Multivariate Analysis* **160** 169–184. MR3688697

[56] Zhao, S. D., Cai, T. T., Cappola, T. P., Margulies, K. B. and Li, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *Journal of the American Statistical Association* **112** 1032–1046. MR3735358