# Adaptive Filtering Multiple Testing Procedures for Partial Conjunction Hypotheses

Jingshu Wang, Chiara Sabatti, Art B. Owen *
Department of Statistics, Stanford University

October, 2016

**Abstract**

The partial conjunction (PC) alternative hypothesis $H_1^{r/n}$ stipulates that at least $r$ of $n$ related basic hypotheses are non-null, making it a useful measure of replicability. Motivated by genomic problems we consider a setting with a large number $M$ of partial conjunction null hypotheses to test, based on an $n \times M$ matrix of $p$-values. When $r > 1$ the hypothesis $H_0^{r/n}$ is composite. Validity versus the case with $r-1$ alternative hypotheses holding can lead to very conservative tests. We develop a filtering approach for $H_0^{r/n}$ based on the $M$ $p$-values for $H_0^{(r-1)/n}$. This filtering approach has greater power than straightforward PC testing. We prove that it can be used to control the familywise error rate, the per family error rate, and the false discovery rate among $M$ PC tests. In simulations we find that our filtering approach properly controls the FDR while achieving good power.

*Keywords:* meta-analysis, replicability, screening, BH procedure

*This work was supported by the US National Science Foundation under grant DMS-1521145

# 1 Introduction

Replication has been referred to as "the cornerstone of science" (Moonesinghe et al., 2007). However in modern science, there is a replication crisis where many of the scientific findings, especially results of experiments in medicine, psychology and biology, lack replicability (Moonesinghe et al., 2007; Baker, 2016). From the data analysis point of view, new statistical methods and models are needed to encourage and assess replicable findings. Testing for partial conjunction hypotheses serves this purpose. When the results of $n$ studies are combined, a partial conjunction (PC) hypothesis tests whether at least $r$ out of $n$ hypotheses are false where $r$ is some chosen number. For $r \geq 2$, rejecting the partial conjunction null explicitly guarantees that the finding is replicable in at least some of the studies. Besides testing for replication in replicated experiments, PC hypotheses can also be used in testing for repeated effects across different conditions (Sun and Wei, 2011; Heller et al., 2007).

When combining results of $n$ studies, a typical approach is meta-analysis which aggregates information across studies to increase statistical power. However, it is possible that the null is then rejected largely on the basis of just one extremely significant component hypothesis test. Such a rejection may be undesirable as it could arise from some irreproducible property (such as technical or model bias) of the setting in which that one component test was made. In contrast, researchers in functional magnetic resonance imaging (fMRI) have adopted conjunction (logical 'and') testing (Price and Friston, 1997) in which an hypothesis must be rejected in all $n$ settings where it is tested. Conjunction tests lose power for large $n$ as they are based on the largest of $n$ p-values. The PC hypothesis is a compromise between the above approaches, achieving power gain and replicability at the same time. It was first used in Friston et al. (2005) and then studied by Benjamini and Heller (2008). The extremes $r = 1$ and $r = n$ correspond to the usual meta-analysis tests and conjunction testing respectively.

In modern data analyis, for example genetic data analysis of high-throughput experiments, it is common that thousands of hypotheses are tested at the same time in one study. Thus a typical application of the PC hypothesis involves a multiple hypotheses testing problem, where many of PC hypotheses are tested together. The data is typically given as a p-value matrix. A classical multiple testing procedure will first calculate a valid p-value for each PC hypothesis and apply standard procedures on these p-values to achieve simultaneous error control. However, these classical procedures are well-known for their low power (Heller and Yekutieli, 2014; Sun et al., 2015) in finding non-null PC hypotheses and discourage the applications of PC hypotheses.

In this paper, we propose new procedures for PC hypotheses that are much more powerful than classical methods but still achieve simultaneous error control. These procedures are based on an adaptive filtering idea. The key reason that the classical procedures are conservative is that a PC null hypothesis with $r \geq 2$ is composite. A valid p-value requires type-I error control under the

worst case of a null PC hypothesis, while the simultaneous error for multiple testing is controlled at the true configuration of the hypotheses. As a consequence, under some configurations of the true null PC hypothesis, a valid p-value would be too large. The adaptive filtering idea is to find filtering regions, such that conditional on these regions, the PC $p$-value is still valid while its efficiency is not sensitive to the actual configuration of the null hypothesis. We prove that our procedures can be designed to control several simultaneous error rates under independence of all test statistics. Our simulations show that they also control these error rates under local and weak dependency structure of the PC hypotheses.

To our knowledge, this is the first frequentist multiple testing procedure that largely increases the power of classical procedures for PC hypotheses with any given $n$ and $r$. Two related methods are the procedures in Bogomolov and Heller (2015) and the empirical Bayes approach in Heller and Yekutieli (2014). However, Bogomolov and Heller (2015) is designed for the case of $n = r = 2$ and the empirical Bayes approach cannot work well when $n$ is large in our simulations.

The structure of the paper is as follows. Section 2 sets up the problem and lists related concepts and notation. Section 3 reviews the classical multiple testing procedures on PC hypotheses and their problems. Section 4 defines our adaptive filtering procedures. Section 5 proves the validity of our procedures under independence and an important monotonicity property. Section 6 discusses extension of the procedures to more general scenarios and compares them to previous methods. Section 7 describes simulations and finally, Section 8 provides a real data example.

# 2 Preliminaries

Here we introduce our notation for partial conjunction null hypotheses and test statistics. Then we review some needed facts about simultaneous testing, and introduce a few definitions.

## 2.1 Partial conjunction hypotheses and tests

We want to simultaneously test for the PC null hypotheses:

$$H_{0j}^{r/n} : \text{fewer than } r \text{ out of } n \text{ hypotheses are nonnull}$$

for $j = 1, 2, \ldots, M$. Each PC null hypothesis $H_{0j}^{r/n}$ involves $n$ individual null hypotheses $H_{0ij} : \theta_{ij} \in \mathbf{\Theta}_{0ij}$ for $i = 1, 2, \cdots, n$. The corresponding alternative hypothesis is denoted as $H_{1ij} : \theta_{ij} \in \mathbf{\Theta}_{1ij}$. Let $(\omega_{ij})_{n\times M}$ be an underlying hypothesis indicator matrix with $\omega_{ij} = 0$ if $H_{0ij}$ is true and $\omega_{ij} = 1$ if $H_{1ij}$ is true.

We group together hypotheses from $n$ studies into the vector $\boldsymbol{\omega}_{\cdot j} = (\omega_{1j}, \ldots, \omega_{nj}) \in \{0,1\}^n$. Then the partial conjunction null and alternative hypotheses are

$$H_{0j}^{r/n} : \|\boldsymbol{\omega}_{\cdot j}\|_0 \leq r-1, \quad \text{and} \quad H_{1j}^{r/n} : \|\boldsymbol{\omega}_{\cdot j}\|_0 \geq r$$

respectively, where $\|\cdot\|_0$ is the $L_0$ norm counting the number of nonzero entries of a vector. We further use $v_j = 1$ to indicate that $H_{1j}^{r/n}$ holds and $v_j = 0$ otherwise. The p-value matrix corresponding to the individual hypotheses $(H_{0ij})_{n\times M}$ is $(p_{ij})_{n\times M}$. It is the observed value of a corresponding random matrix $(P_{ij})_{n\times M}$.

A multiple testing procedure is represented as a vector of decision functions $(\varphi_1, \varphi_2, \cdots, \varphi_M)$ where $\varphi_j = 1$ if we reject $H_{0j}^{r/n}$ and $\varphi_j = 0$ otherwise. Note that each decision function $\varphi_j = \varphi_j\big((p_{ij})_{n\times M}\big)$ depends on the whole p-value matrix $(p_{ij})_{n\times M}$ because of multiple testing adjustment.

For each $j = 1, 2, \cdots, M$, let the p-value vector be $p_{\cdot j} = (p_{1j}, \cdots, p_{nj})$ and its sorted p-values be $p_{(1)j} \leq p_{(2)j} \leq \cdots \leq p_{(n)j}$. Our tests $\varphi_j$ will depend most strongly on $p_{\cdot j}$ but elements of $p_{\cdot j'}$ for $j \neq j'$ also play a role.

## 2.2 Simultaneous error control in multiple hypotheses testing

If $\varphi_j = 1$ while $v_j = 0$ then a false discovery, or type-I error, has occured. The number of false discoveries is $V = \sum_{j=1}^M \varphi_j 1_{v_j=0}$. The number of total discoveries is $R = \sum_{j=1}^M \varphi_j$. There are many measures to control the type-I error (Dudoit and Van Der Laan, 2007) in multiple testing problems, among which the most commonly used ones are familywise error rate (FWER), the per family error rate (PFER) and the false discovery rate (FDR). The FWER is the $\mathbb{P}(V \geq 1)$, the PFER is $\mathbb{E}(V)$ and the FDR is $\mathbb{E}(V/\max(R,1))$.

Let the $M$ hypotheses have p-values $P_1, P_2, \cdots, P_M$. A classical procedure controlling FWER at level $\alpha$ is the Bonferroni correction which rejects $H_{0j}$ if $P_j \leq \alpha/M$. The Bonferroni correction also controls the PFER at level $\alpha$ (Tukey, 1953).

The most widely used procedure controlling the FDR is the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). One representation of the BH procedure is to reject $H_{0j}$ if $P_j \leq \gamma_0$ where $\gamma_0$ is a data-dependent threshold defined as

$$\gamma_0 = \max\Big\{\gamma \in \big\{\alpha, \cdots, \frac{2}{M}\cdot\alpha, \frac{\alpha}{M}\big\} : M\gamma \leq \alpha \sum_{j=1}^M 1_{P_j \leq \gamma}\Big\},$$

where we take $\gamma_0 = 0$ by convention, if it is the maximum of the empty set.

There is a major difference between simultaneous error control for multiple testing and the type-I error control for single hypothesis testing. The type-I error for a single hypothesis must

be controlled even when the worst case of null hypothesis is true. The simultaneous error in multiple testing, whatever measurement we use, must be controlled at the true configuration of the hypotheses.

## 2.3 Complete and partial monotonicity

Wang and Owen (2015) show the importance of a monotonicity constraint for a single PC test. Here, monotonicity means that the combined p-value is a non-decreasing function of individual p-values that go into it. Without the monotonicity constraint, any monotone PC P-value is inadmissible and the counter-examples are very counter intuitive. We think that it is a reasonable requirement for a practically useful test.

For multiple PC hypotheses testing, we will also need a monotonicity constraint. We define both "complete monotonicity" and "partial monotonicity".

**Definition 1** (Complete monotonicity)**.** A multiple testing procedure has complete monotonicity if each decision function $\varphi_j(p_{\cdot 1}, \cdots, p_{\cdot M})$ is a non-increasing function in all the elements of $(p_{ij})_{n\times M}$ for $j = 1, 2, \cdots, M$.

**Definition 2** (Partial monotonicity)**.** A multiple testing procedure has partial monotonicity if for all $j \in 1{:}M$, its decision function $\varphi_j(p_{\cdot 1}, \cdots, p_{\cdot M})$ is non-increasing in all elements of $(p_{1j}, p_{2j}, \ldots, p_{nj})$.

Complete monotonicity guarantees that if any $p$-value $p_{ij}$ is reduced while all others are held fixed, then we should reject any PC null hypotheses that we originally rejected and maybe more. Partial monotonicity only requires the test of hypothesis $j$ to be monotone in the $p$-values for that same hypothesis. It allows a reduction in $p_{ij'}$ for $j' \neq j$ to reverse a rejection of $H_{0j}^{r/n}$.

Most multiple comparison procedures for $n = 1$ (for example, the BH procedure) satisfy complete monotonicity. However, complete monotonicity is a stringent assumption and prevents us from finding efficient procedures for partial conjunction nulls. This problem will be discussed in more detail in Section 5.2. We think that partial monotonicity is the appropriate requirement for a reasonable multiple testing procedure of partial conjunction hypotheses.

## 2.4 Some notation for sets and subsets

We finish this section with some notation for index sets. We use $1{:}n$ to denote $\{1, 2, \cdots, n\}$. The index set $u \subset 1{:}n$ has cardinality $|u|$ and complement $-u = 1{:}n \setminus u$. We also define null hypotheses $H_{0uj}^{1/|u|} = \cap_{i\in u} H_{0ij}$ which is the global null hypothesis of $\{H_{0ij} : i \in u\}$. If $u = \{i_1, i_2, \cdots, i_{|u|}\}$, then we use $\boldsymbol{p}_{uj}$ to denote the vector $\left(p_{i_1 j}, p_{i_2 j}, \cdots, p_{i_{|u|} j}\right)$. Finally, our convention is that the maximum of a null set $\varnothing$ is 0, as we used already in defining $\gamma_0$ of the BH procedure.

# 3 Classical procedures on PC hypotheses

Wang and Owen (2015) discussed the superiority of a class of PC p-values, called GBHPC p-values, for a single PC null hypothesis. It is a generalization of the BHPC p-values discussed in Benjamini and Heller (2008) for PC hypotheses.

**Definition 3** (GBHPC P-value)**.** For each $u \subset 1{:}n$ with $|u| = n - r + 1$ let $g_u$ be a function from $[0,1]^{n-r+1}$ to $[0,1]$ such that $g_u$ is non-decreasing and is a valid meta-analysis p-value for $H_{0u}^{1/u}$. Then

$$P_{r/n} = f^\star(P_1, \cdots, P_n) = \max_{\substack{u \subset 1:n \\ |u| = n-r+1}} g_u(\boldsymbol{P}_u) \tag{1}$$

is a generalized BHPC (GBHPC) p-value.

Here, a valid meta-analysis P-value $g_u(\boldsymbol{P}_u)$ means a valid combined P-value for the corresponding global null hypothesis $H_{0u}^{1/|u|}$. The original BHPC statistics are of the GBHPC form but use the same function $g(\cdot)$ for all subsets $u$ with $|u| = n - r + 1$. From the non-decreasing property they take the form $P_{r/n} = g(P_{(r)}, P_{(r+1)}, \ldots, P_{(n)})$. The extension from BHPC to GBHPC is mostly technical, and the BHPC form is most natural for applications. Wang and Owen (2015) show that under monotonicity the BHPC tests are admissible, and conversely, any admissible monotone PC test is of the GBHPC form, though not necessarily the BHPC form. Next we list a few examples of BHPC p-values from Benjamini and Heller (2008).

**Example 1.**

Simes' method:

$$P_{r/n}^S = \min_{i=r,\cdots,n} \left\{ \frac{n-r+1}{j-r+1} P_{(i)} \right\}$$

which is valid for positively dependent P-values.

**Example 2.** Fisher's method:

$$P_{r/n}^F = \mathbb{P}\left( \chi^2_{(2(n-r+1))} \geq -2 \sum_{i=r}^{n} \log P_{(i)} \right),$$

which is valid for independent P-values.

**Example 3.** Bonferroni's method:

$$P_{r/n}^B = (n-r+1) P_{(r)},$$

which is valid under any dependence structure of P-values.

Let $P_{r/n,j}$ be a GBHPC P-value for the $j$th PC null hypothesis. We could simply apply standard multiple testing procedures to the GBHPC P-values $\{P_{r/n,j} : j = 1, 2, \cdots, M\}$. For instance, to control FWER and PFER at level $\alpha$, one can use the Bonferroni correction, $\varphi_j = 1_{P_{r/n,j} \le \alpha/M}$. This guarantees $\mathbb{E}(V) \le \alpha$ under any dependence structure and configuration of the individual hypotheses. Similarly, to control the FDR one can apply BH to $\{P_{r/n,j} : j = 1, 2, \cdots, M\}$.

The above procedures would generally be conservative, especially when in many application problems one would expect that the true nonnull individual hypotheses are sparse. To quantify the conservativeness, we consider the following example for Bonferroni correction when $r = n$, then discuss the more general case, before proposing our alternative in Section 4.

## 3.1 The case $r = n$

If each partial conjunction null hypothesis is $H_{0j}^{n/n}$ $(r = n)$, then the only possible GBHPC P-value is $P_{n/n,j} = P_{(n)j}$. With a Bonferroni correction, $H_{0j}^{n/n}$ is rejected if $p_{(n)j} \le \alpha/M$ when FWER (or PFER) is controlled at level $\alpha$.

Define sets $\mathcal{I}_k \subset 1{:}M$ such that $j \in \mathcal{I}_k$ means that precisely $k$ of the alternative hypotheses $H_{1ij}$ hold for the given $j$. For $k = 1, \ldots, n-1$ we have

$$\mathcal{I}_k = \{j \in 1{:}M : H_{0j}^{(k+1)/n} \text{ is true but } H_{0j}^{k/n} \text{ is false}\},$$

while $\mathcal{I}_0 = \{j \in 1{:}M : H_{0j}^{1/n} \text{ is true}\}$ and $\mathcal{I}_n = \{j \in 1{:}M : H_{0j}^{n/n} \text{ is false}\}$. Assume that for each $j$, the $n$ individual P-values $P_{1j}, \cdots, P_{nj}$ are independent.

Under the above setup, an upper bound for the PFER is

$$\begin{aligned}
\mathbb{E}(V) &= \sum_{j=1}^{M} \mathbb{E}(1_{P_{(n)j} \le \alpha/M} \cdot 1_{v_j = 0}) \\
&\le \sum_{j=1}^{M} \mathbb{P}(P_{ij} \le \alpha/M \text{ for all } i \text{ with } \omega_{ij} = 0) \\
&\le \sum_{k=0}^{n-1} |\mathcal{I}_k| \cdot \frac{\alpha^{n-k}}{M^{n-k}}.
\end{aligned}$$

This estimate of $\mathbb{E}(V)$ can be quite close to the true value if all the tests of non-null hypotheses $H_{1ij}$ have high power.

We are interested in the setting where $M$ is large. Let $\delta_k = |\mathcal{I}_k|/M$ and suppose that $(\delta_0, \delta_1, \ldots, \delta_n)$ approaches some limit as $M \to \infty$. The bound for PFER is

$$\alpha\Big(\delta_{n-1} + \delta_{n-2}\frac{\alpha}{M} + \delta_{n-3}\Big(\frac{\alpha}{M}\Big)^2 + \cdots + \delta_0\Big(\frac{\alpha}{M}\Big)^n\Big)$$

and so it is clear that in this limit the bound on $\mathbb{E}(V)$ is dominated by $\delta_{n-1}\alpha$ so long as $\delta_{n-1}$ is bounded away from 0. If instead $\delta_{n-1} = 0$, then $\mathbb{E}(V) = O(M^{-1})$. More generally if $\delta_{n-1}$ is small then $\mathbb{E}(V)$ will be much less than its nominal level $\alpha$ for large $M$. In genetics problems we might have $M$ in the thousands (microarrays) or millions (SNPs) with $\delta_0 \approx 1$ and $\delta_{n-1} \approx 0$. Under these circumstances the procedure would be very conservative, in fact much more conservative than Bonferroni usually is.

### 3.2 More general $r$ and $n$

For general $2 \leq r \leq n$, the magnitude of $\mathbb{E}(V)$ for Bonferroni correction of GBHPC p-values will depend mainly on $\delta_{r-1}$ in the large $M$ setting. The partial conjunction null $H_0^{r/n}$ is a composite null hypothesis and we can design more efficient procedures if we know the fractions $\delta_k$. Unfortunately, $\delta_k$ are unknown. If we can estimate $\delta_k$ from the given p-value matrix, it is possible that there exists a more efficient procedure for simultaneous testing of PC null hypotheses. This is what motivates the Bayesian methods (Heller and Yekutieli, 2014; Flutre et al., 2013). From a frequentist perspective, we do not estimate $\delta_k$ but design a filtering region such that conditional on the filtering region the rejection probability under the PC null hypothesis is not sensitive to the actual configuration of the null.

## 4 Adaptive filtering procedures

In the previous section we saw that for the partial conjunction null $H_0^{r/n}$, the most influential null configuration arises when $H_0^{r/n}$ is true but $H_0^{(r-1)/n}$ is false. This finding motivates the definition of the filtering p-values we use in our adaptive filtering procedures.

### 4.1 The adaptive filtering heuristics

The key reason that original procedures on partial conjunction hypotheses are conservative is that the partial conjunction null is composite and the inequality $\mathbb{P}(P_{r/n} \leq \gamma) \leq \gamma$ can be very loose. In the extreme setting with $r = n$ (in Section 3.1) $\mathbb{P}(P_{(n)j} \leq \gamma)$ can be as low as $\gamma^n$ under the complete null but as high as $\gamma$ if only one individual hypothesis is null. We will define filtering regions $\mathcal{A}_\gamma \subset [0,1]^n$ such that $\mathbb{P}(P_{r/n,j} \leq \gamma \mid P_{\cdot j} \in \mathcal{A}_\gamma) \leq \gamma$ is still valid under any cases in the partial conjunction null space but can be much tighter than $\mathbb{P}(P_{r/n,j} \leq \gamma) \leq \gamma$ for some configurations.

Consider a simple example of $n = r = 2$ and assume that for each $j$, $P_{1j}$ and $P_{2j}$ are independent. Each partial conjunction null hypothesis consists of three possible scenarios: $\boldsymbol{\omega}_{\cdot j} \in$

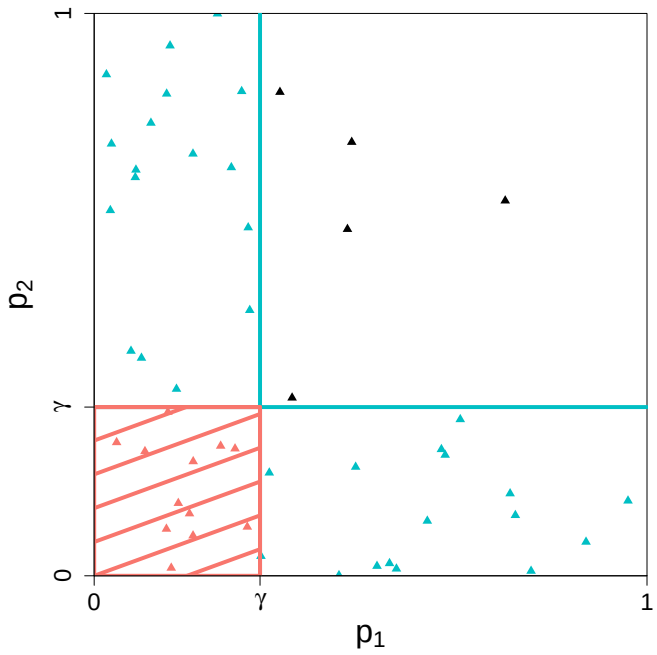


Figure 1: Illustration of estimating $V$ for $n = r = 2$.

$\{(0,0),(0,1),(1,0)\}$. For any given threshold $\gamma \in [0,1]$, define the fitering region $\mathcal{A}_\gamma = \{(p_1,p_2) : \min(p_1,p_2) \le \gamma\}$. As illustrated in Figure 1, the blue "L" shaped region is the filtering region and the red square is the rejection region $\mathcal{R}_\gamma = \{(p_1,p_2) : \max(p_1,p_2) \le \gamma\}$. It's easy to check geometrically that $\mathbb{P}(P_{\cdot j} \in \mathcal{R}_\gamma \mid P_{\cdot j} \in \mathcal{A}_\gamma) \le \gamma$ if $H_{0j}^{2/2}$ is true regardless of what the true null case is, since at least one of $P_{1j}$ and $P_{2j}$ is stochastially larger than uniform.

In multiple hypothesis testing, the above inequality provides an alternative way to estimate $V$, the number of false positives. Instead of estimating $V$ as $\hat{V}_B = \gamma \cdot M$ which is used in the classical Bonferroni correction and BH procedure, $V$ can be estimated as

$$\hat{V} = \gamma \cdot \Big( \sum_{j=1}^{M} 1_{\min(P_{1j},P_{2j}) \le \gamma} \Big).$$

Obviously, $\hat{V} \le \hat{V}_B$ and the gap can be large if most of the hypotheses are complete nulls. It is only for simple null hypotheses that $\hat{V}_B$ provides a good estimate of $V$.

The regions $\mathcal{A}_\gamma$ act as filters as one need to consider a partial conjunction hypothesis as candidate hypothesis only when $P_{\cdot j} \in \mathcal{A}_\gamma$. The filtering regions $\mathcal{A}_\gamma$ are "dynamic" as they depend on

the threshold $\gamma$.

The threshold $\gamma$ is determined adaptively. For controlling FWER and PFER at level $\alpha$ (which is to guarantee $\mathbb{E}(V) \leq \alpha$), one can select $\gamma$ such that

$$\gamma \cdot \Big( \sum_{j=1}^{M} 1_{\min(P_{1j}, P_{2j}) \leq \gamma} \Big) \approx \alpha. \tag{2}$$

For controlling FDR at level $\alpha$, one estimate the false discovery proportion $V/R$ as $\hat{V}/R$ and select $\gamma$ such that

$$\frac{\gamma \cdot \Big( \sum_{j=1}^{M} 1_{\min(P_{1j}, P_{2j}) \leq \gamma} \Big)}{\sum_{j=1}^{M} 1_{\max(P_{1j}, P_{2j}) \leq \gamma}} \approx \alpha. \tag{3}$$

## 4.2 Definition of the procedures

Here we define adaptive filtering procedures that reject a null hypothesis $H_{0j}^{r/n}$ if the corresponding partial conjunction P-value based on Bonferroni's method is below some threshold: $P_{r/n}^{B} \leq \gamma$. The threshold $\gamma$ is determined adaptively from the p-value matrix. The choice of $\gamma$ also depends on the filtering regions which are defined as:

$$\mathcal{A}_\gamma := \big\{ (p_1, p_2, \cdots, p_n) : (n - r + 1) p_{(r-1)} \leq \gamma \big\}.$$

We propose an adaptive filtering Bonferroni test (two versions) and an adaptive filtering Benjamini-Hochberg test. They all depend on the filtering $P$-values

$$F_j := (n - r + 1) P_{(r-1)j} \tag{4}$$

and selection $P$-values

$$S_j := P_{r/n,j}^{B} = (n - r + 1) P_{(r)j}. \tag{5}$$

**Definition 4** (Adaptive filtering Bonferroni)**.** For a control level $\alpha$, and with $F_j$ and $S_j$ given by (4) and (5) respectively, reject $H_{0j}^{r/n}$ if $S_j \leq \gamma_0^{\text{Bon}}$ where

$$\gamma_0^{\text{Bon}} = \max \Big\{ \gamma \in \big\{ \alpha, \frac{\alpha}{2}, \cdots, \frac{\alpha}{M} \big\} : \gamma \cdot \sum_{j=1}^{M} 1_{F_j \leq \gamma} \leq \alpha \Big\}.$$

**Definition 5** (Adaptive filtering Benjamini-Hochberg)**.** For a control level $\alpha$, and with $F_j$ and $S_j$ given by (4) and (5) respectively, reject $H_{0j}^{r/n}$ if $S_j \leq \gamma_0^{\mathrm{BH}}$ where

$$\gamma_0^{\mathrm{BH}} = \max\Big\{\gamma \in \mathcal{I}_{\alpha,M} : \gamma \cdot \sum_{j=1}^{M} 1_{F_j \leq \gamma} \leq \alpha \cdot \sum_{j=1}^{M} 1_{S_j \leq \gamma}\Big\}$$

and

$$\mathcal{I}_{\alpha,M} = \Big\{\frac{k}{m} \cdot \alpha : k \in 1{:}M, m \in 1{:}M, k \leq m\Big\}.$$

The adaptive filtering Bonferroni procedure can be defined equivalently using a more programmable two-step approach, which is also illustrated in Figure 2.

**Definition 6** (Adaptive filtering Bonferroni, alternative form)**.** For a control level $\alpha$, and with $F_j$ and $S_j$ given by (4) and (5) respectively, proceed as follows. First sort $\{F_1, F_2, \cdots, F_M\}$ into $F_{(1)} \leq F_{(2)} \cdots \leq F_{(M)}$. Then let

$$m' = \min\Big\{j \in 1{:}M : \ \frac{\alpha}{j} < F_{(j)}\Big\} \tag{6}$$

and define

$$m = \begin{cases} m', & \text{if } F_{(m')} \leq \frac{\alpha}{m'-1} \\ m' - 1, & \text{otherwise.} \end{cases}$$

Finally, reject $H_{0j}^{r/n}$ if $S_j \leq \alpha/m$. If the set in (6) is $\varnothing$, then take $m' = m = M$.

**Proposition 4.1.** *The two definitions of the Adaptive filtering Bonferroni procedure in 4 and 6 are equivalent.*

The alternative form provides another view of the adaptive filtering procedures. We filter out hypotheses with large $F_j$ values and only adjust for the $m$ remaining hypotheses. If $m \ll M$ then adaptive filtering can reject many more hypotheses than direct Bonferroni or BH on $S_j$.

# 5 Theory

## 5.1 Validity of the adaptive filtering procedures

When the P-values are independent across both the studies and the multiple testing cases, we can prove that the adaptive filtering procedures control the corresponding multiple testing error under any configurations of the true individual hypotheses. In addition to independence we assume validity.

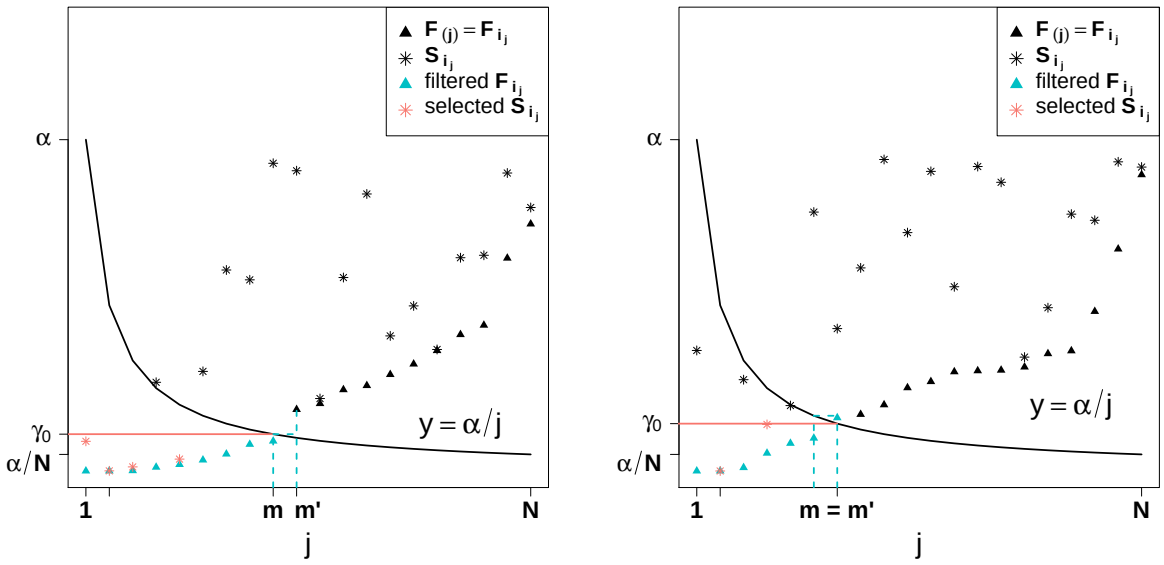


Figure 2: Adaptive filtering Bonferroni procedure. The left plot shows the case when $m = m' - 1$ and the right plot shows the case when $m = m'$. Define an order $(i_1, i_2, \cdots, i_M)$ where $i_j$ in the index corresponding to $F_{(j)}$. The points represent $(F_{i_j}, S_{i_j})$ pairs. The blue colored triangles represent the kept indice after the filtering step and the red stars represent the selected hypotheses for rejection in the selection step.

**Definition 7** (Valid components)**.** If $\mathbb{P}(P_{ij} \leq \alpha) \leq \alpha$ under $H_{0ij}$ for any $\alpha \in [0, 1]$, then $P_{ij}$ is a valid $P$-value.

The next two theorems state the validity of adaptive filtering procedures: the adaptive filtering Bonferroni procedure controls PFER (and FWER) and the adaptive filtering BH procedure controls FDR at the nominal level under independence.

**Theorem 5.1.** *Let $(P_{ij})_{n\times M}$ contain independent valid p-values. Then the adaptive filtering Bonferroni procedure in 4 controls FWER and PFER at level $\alpha$ for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \cdots, M\}$.*

**Theorem 5.2.** *Let $(P_{ij})_{n\times M}$ contain independent valid p-values. Then the adaptive filtering BH procedure in 5 controls the FDR at level $\alpha$ for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \cdots, M\}$.*

The proofs of both Theorems 5.1 and 5.2, depend on Lemma 5.3 below. It gives a conditional validity property of the selection statistics $S_j$ after they are filtered via $F_j$.

**Lemma 5.3.** *Let $(P_{ij})_{n\times M}$ contain independent valid p-values and assume that $H_{0j}^{r/n}$ is true for some $j \in 1{:}M$. Then*

$$\mathbb{P}\big(S_j \le \beta \mid F_j \le \beta\big) \le \beta \tag{7}$$

*holds whenever $\beta > 0$ and $\mathbb{P}\big(F_j \le \beta\big) > 0$. Here $F_j$ and $S_j$ are given by (4) and (5) respectively.*

Equation (7) can be equivalently written as $\mathbb{P}\big(S_j \le \beta\big) \le \beta\mathbb{P}(F_j \le \beta)$. It holds also when $\mathbb{P}(F_j \le \beta) = 0$ as $S_j \ge F_j$ is always true.

*Remark* 5.1. The independence assumption on $(P_{ij})$ in Theorems 5.1 and 5.2 can be relaxed a little bit. In the proof, we only require that if $v_j = 0$ ($H_{0j}^{r/n}$ is true), then the P-values $(P_{1j}, P_{2j}, \cdots, P_{nj})$ are independent of each other and of the other entries in $(P_{ij})_{n\times M}$.

## 5.2 Partial monotonicity

Comparing the adaptive filtering procedures with directly applying Bonferroni correction or BH on the Bonferroni style PC P-values $P^B_{r/n,j}$, the adaptive filtering procedures can reject many more PC null hypotheses as it only corrects for candidate hypotheses with $F_j \le \gamma_0$. However, as shown in Wang and Owen (2015), we need monotonicity constraints when discussing about efficiency of tests for PC hypotheses. To show that the adaptive filtering procedure is reasonable, we prove that it satisfies partial monotonicity.

**Corollary 5.4.** *Let $(P_{ij})$ be a matrix of valid p-values. Then both the adaptive Bonferroni procedure of 4 and the adaptive BH procedure of 5 satisfy partial monotonicity for each null hypotheses $H_{0j}^{r/n}$, $j = 1, 2, \cdots, M$.*

Notice that these adaptive filtering procedures do not satisfy complete monotonicity. For a case $j$ with $F_j > \gamma_0$, when the elements in $P_{\cdot j}$ get smaller, $\tilde{\gamma}_j$ can get smaller, and potentially reverse the rejection of $H_{0k}^{r/n}$ for some $k \ne j$. As we have discussed, the complete monotonicity requirement, though suitable for most multiple testing problems, is too stringent for an efficient multiple testing procedure of PC hypotheses.

# 6 Discussions

## 6.1 Extension to heterogeneous PC hypotheses

In some applications, the $M$ PC null hypotheses can be heterogeneous and have their unique $r_j$ or $n_j$. Then the $j$th PC null hypothesis becomes:

$$H_{0j}^{r_j/n_j} : \text{fewer than } r_j \text{ out of } n_j \text{ hypotheses are nonnull.}$$

For example in genetic analysis, if each $P_{ij}$ is the P-value testing for the null hypothesis related to gene $j$ in experiment $i$, then gene $j$ might not be present in every single experiment. The adaptive filtering procedures still work in this scenario. We replace formulas (4) and (5) by

$$F_j = (n_j - r_j + 1)P_{(r_j-1)j}, \quad \text{and} \quad S_j = (n_j - r_j + 1)P_{(r_j)j}$$

respectively. Adaptive filtering still controls the required error rate here because the conditional validity 5.3 still holds. Thus Theorems 5.1 and 5.2 still apply.

## 6.2 Comparison with other literature

Two methods that share related ideas with our adaptive filtering procedures for PC hypotheses are the multiple testing procedure developed in Bogomolov and Heller (2015) for replicability analysis where $n = r = 2$ and the empirical Bayes approach in Heller and Yekutieli (2014) for controlling for the Bayes FDR for multiple PC hypotheses. Both methods are developed to improve the efficiency of the original procedures solely based on BHPC P-values in Benjamini and Heller (2008). To some extent, our adaptive filtering procedures generalizes the procedures in Bogomolov and Heller (2015) to arbitrary $n$ and $r$ and provides a frequentist approach comparable to the empirical Bayes method in Heller and Yekutieli (2014).

Bogomolov and Heller (2015) provide procedures for controlling either FWER or FDR when $n = r = 2$. Their procedures select candidate hypotheses for each study if the corresponding p-value from the other study passes some threshold. Then from the candidates, a hypothesis is rejected for $n = r = 2$ if its BHPC p-value (for $n = r = 2$, it's the maximum of the two p-values) passes some criterion. To maximize the efficiency of the procedures, the authors suggest a data-adaptive threshold that is very similar to our adaptive filtering procedures. For instance, to control FWER, they chose two thresholds $\gamma_1$ and $\gamma_2$ to satisfy

$$\gamma_1 \cdot \Big( \sum_{j=1}^{M} 1_{P_{2j} \le \gamma_2} \Big) \approx \alpha/2, \quad \text{and} \quad \gamma_2 \cdot \Big( \sum_{j=1}^{M} 1_{P_{1j} \le \gamma_1} \Big) \approx \alpha/2.$$

Compared to our adaptive filtering Bonferroni procedure for $r = n = 2$ in (2), the values of $\gamma_1$, $\gamma_2$ and $\gamma$ should be similar. In their selection step, $H_{0j}^{2/2}$ is rejected if both $p_{2j} \le \gamma_2$ and $p_{1j} \le \gamma_1$. Thus the selection step is comparable to the selection criterion in our adaptive filtering Bonferroni procedure. Our adaptive filtering BH procedure also has similarities to their procedure in controlling for FDR. Our contribution is that we propose a general adaptive filtering idea that can solve the situations for any combination of $r$ and $n$ easily, and provides simple proofs.

Heller and Yekutieli (2014) used an empirical Bayes approach for multiple hypotheses testing. They estimated the fractions for each of the $2^n$ configurations of individual hypotheses being null or

non-null, along with the distribution of the Z-values under each configuration. As discussed before, we utilize the filtering P-values to stabilize the efficiency of the PC p-value under any null PC configuration. As a consequence, there is no need to estimate the fractions of each configuration which is unavailable in a frequentist setting. Another advantage of the adaptive filtering procedures is that the computation cost is $O(1)$ in $n$ while the cost of the empirical Bayes method increases exponentially in $n$. Simulation results in Section 7 show that our procedure is more accurate in controlling for FDR when $n$ is large for PC hypotheses.

### 6.3 Behavior when $r$ changes

The partial conjunction null $H_0^{r/n}$ has the flexibility to test for any $1 \leq r \leq n$, thus sometimes it is of interest to test for all possible $r$ values. When used in such manner, the adaptive filtering procedure may not have a monotonicity property, in the sense that the rejection set of the PC nonnull hypotheses is decreasing in monotone when $r$ increases. The explanation is that the filtering information based on the whole P-value matrix can be different for different $r$. The expected number of false discoveries accumulates when different $r$ values are considered at the same time.

## 7 Simulation

We use simulated data to explore the efficiency of adaptive filtering procedures. Our procedures are compared with using Bonferroni or BH correction on three types of PC P-values from the Simes, Fisher and Bonferroni examples of Section 3. For FDR control, we also computed the empirical Bayes method in Heller and Yekutieli (2014). We used the R package repfdr of the original authors. Besides the independence scenario, we also simulate P-values with some local dependency to check whether the simultaneous errors are still controlled at the nominal level when independence is violated.

### 7.1 Simulation setup

We set $M = 10000$ and control PFER at the nominal level $\alpha = 1$ and FDR at the nominal level $\alpha = 0.2$. For the repfdr method, we control the Bayes FDR at the same level $\alpha = 0.2$. Bayes FDR is the posterior probability of the hypothesis being null given the rejection region and has been shown to be more conservative then frequentist FDR under independence (Efron, 2012). For a given $n$, there are $2^n$ combinations of individual hypotheses being null ($\omega_{ij} = 0$) or non-null ($\omega_{ij} = 1$) for each PC hypothesis. We use two parameters to control the probability of each combination: $\pi_0$ is

| | | | | | | |
|---|---|---|---|---|---|---|
| n | 2 | 4 | 8 | 4 | 8 | 8 |
| r | 2 | 2 | 2 | 4 | 4 | 8 |

Table 1: Combinations of $n$ and $r$ in our simulation.

the probability of the complete null combination $(0, 0, \cdots, 0)$ and $\pi_{r/n}$ is the total probability of the combinations not belonging to $H_{0j}^{r/n}$. Then we generate each PC hypothesis combination $\boldsymbol{\omega}_{\cdot j}$ independently following the distribution that for a vector $v \in \{0,1\}^n$,

$$\mathbb{P}(\boldsymbol{\omega}_{\cdot j} = v) = \begin{cases} \pi_0, & \text{if } \|v\|_1 = 0, \\ \frac{\pi_{r/n}}{|\{v \in \{0,1\}^n : \|v\|_0 \geq r\}|}, & \text{if } \|v\|_1 \geq r, \\ \frac{1-\pi_0-\pi_{r/n}}{|\{v \in \{0,1\}^n : \|v\|_0 < r\}|} - 1, & \text{else.} \end{cases}$$

We set $\pi_{r/n} = 0.02$ and consider two values of $\pi_0$: $\pi_0 = 0.5$ or $0.9$.

Next, we generate the Z-value matrix, which will be used later to get individual p-values. We select four values $\mathcal{I} = \{\mu_1, \mu_2, \mu_3, \mu_4\}$ for which the detection powers of $Z \sim \mathcal{N}(\mu_i, 1)$ are $0.02, 0.2, 0.5, 0.95$ for $i = 1, 2, 3, 4$ respectively. The detection power is calculated assuming that the individual null hypothesis $H_{0ij}$ is rejected when the individual p-value $P_{ij}$ is less than the Bonferonni threshold $\alpha/M$. We then generate the effect sizes $U_{ij}$ independently following:

$$U_{ij} = \begin{cases} 0, & \text{if } \omega_{ij} = 0 \\ \text{Uniform}(\pm\mathcal{I}), & \text{if } \omega_{ij} = 1. \end{cases}$$

Then, we generate the noise matrix $E \in \mathbb{R}^{n \times M}$. The rows of $E$ correspond to different studies and they are always independent. For a given study (row of $E$), we consider both independent noise and a block dependence correlation structure. The block structure has $b$ blocks of size $m = M/b$, and then $E_{i\cdot} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\rho \otimes I_{b \times b})$ where $\Sigma_\rho \in \mathbb{R}^{m \times m}$ has 1s on the diagonal and common value $\rho \geq 0$ off the diagonal. We set $m = b = 100$ for our simulation and set $\rho$ to either 0 (independent) or 0.8 (block dependent). The Z-value matrix is then $Z = U + E$, and we construct the two-sided p-value matrix $(P_{ij})_{n \times M}$ from the Z-value matrix.

The values of $n$ and $r$ that we used are listed in Table 1. We simulate parameter combination $B = 100$ times and calculate the average power, number of false discoveries and false discovery proportions.

### 7.2 Simulation results

For PFER control, we compare the adaptive filtering Bonferroni procedure with other three classical procedures. Those procedures use Simes or Fisher or Bonferroni derived partial conjunction hypothesis tests over the $n$ studies, and then apply Bonferroni to the resulting $M$ $p$-values. All methods successfully control PFER at the nominal level under both the independence and block dependence cases, with the variance of $V$ generally smaller under independence (Figure 3). However, the classical procedures are too conservative and Figure 4 shows that the adaptive filtering Bonferroni procedure rejects more hypotheses, especially when both $n$ and $r$ are large. It has a greater power improvement when the complete null fraction is higher, which is expected in many genetics applications.

For FDR control, we compare the adaptive filtering BH procedure with the three classical procedures as well as the repfdr method in Heller and Yekutieli (2014). Again the three classical procedures use Simes or Fisher or Bonferroni derived partial conjunction hypothesis tests over the $n$ studies, and then apply BH procedures to the resulting $M$ $p$-values. The repfdr method fails to give an answer in some of the simulation cases and we then count both its number of false discoveries and total discoveries as 0. Similar to the PFER control, the classical procedures are too conservative. The adaptive filtering BH procedure has much more power than these methods in all cases. When $n = 8$, the parameter space for Repfdr is too large thus it does not control FDR at the nominal level. In the cases where repfdr controls FDR, it has comparable power with our adaptive filtering BH.

## 8 Real data example

The data set we use here is the AGEMAP data that has been studied for the LEAPP method in Sun et al. (2012). The AGEMAP study (Zahn et al., 2007) investigated age-related gene expression in mice. Ten mice at each of four age groups were investigated. From these 40 mice, samples were taken of 16 different tissues, resulting in 640 microarray data sets. A small number of those 640 microarrays were missing. From each microarray, 8932 probes were sampled. We use the robust-regression approach in Wang et al. (2015) which provides a simplified version of LEAPP to calculate confounder-adjusted p-values of age effect for each gene and each tissue. It has been shown in Wang et al. (2015) that the adjusted p-values of the genes should be asymptotically independent within each tissue after the confounder adjustment. Figure 7 gives the histogram of adjusted p-values for each tissue, which apprears uniform except for when p-values are very small. Correlation of p-values across tissues are considered as negligible.

Figure 8 compares the number of significant genes when PFER is controlled at 1 and $r$ ranges from 2 to 7. For all $r$ values, the adaptive filtering Bonferroni procedure is much more powerful than

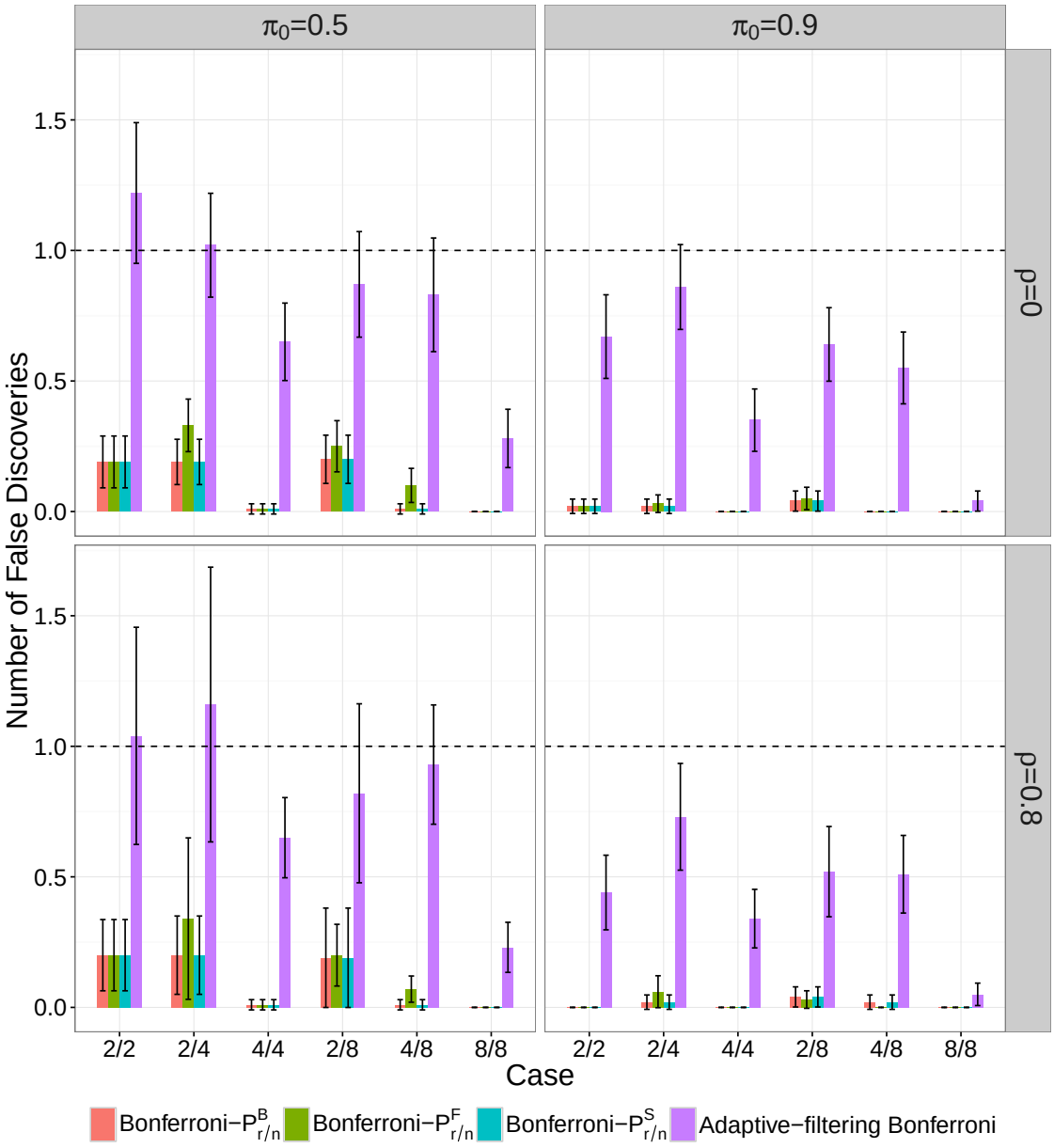


Figure 3: Comparison of expected number of false discoveries $E(V)$ (PFER). The dotted line indicates the nominal level $\alpha = 1$. The error bars are the 95% CI of estimated PFER.

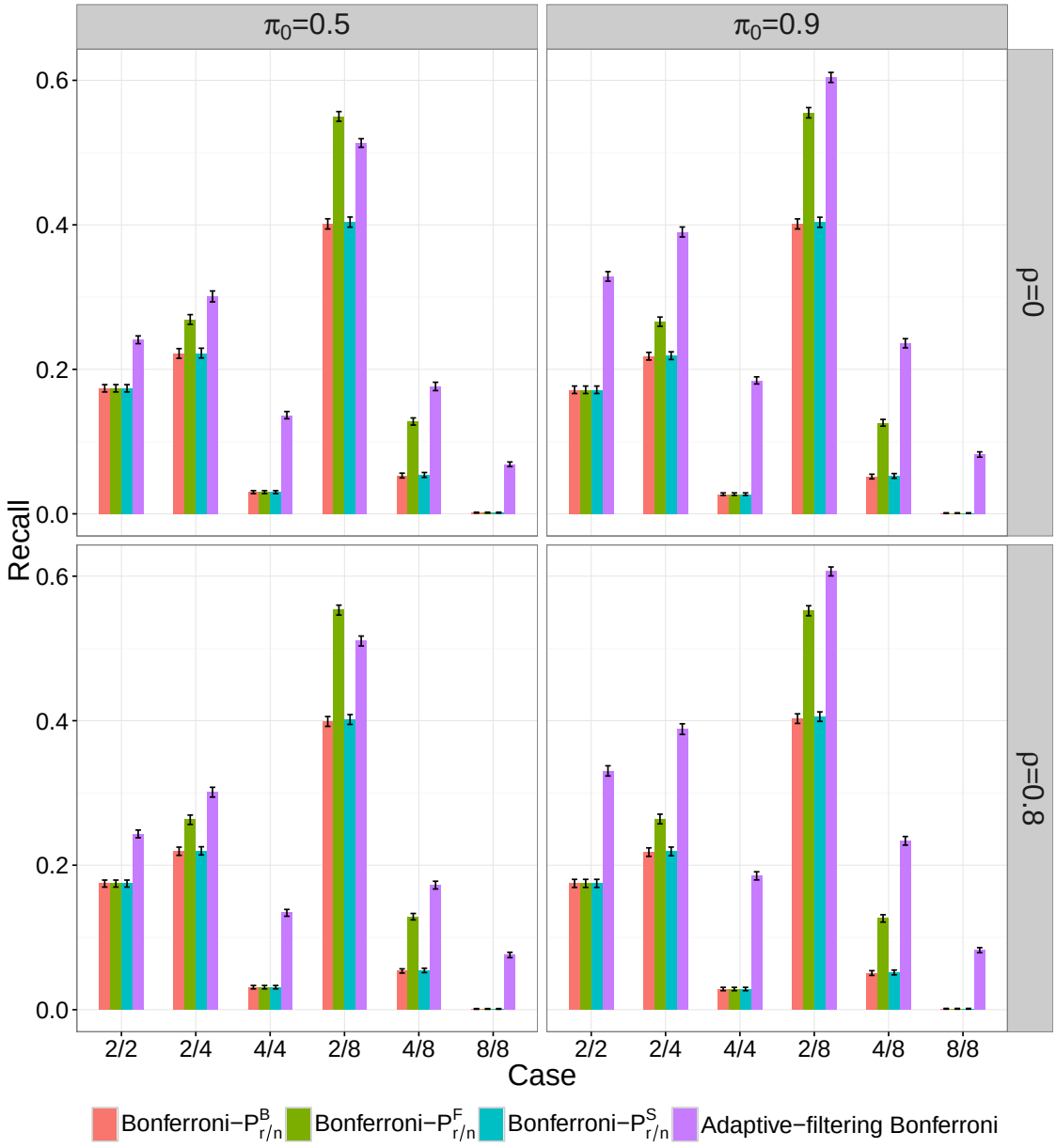


Figure 4: Comparison of power (recall or sensitivity) when PFER is controlled at $\alpha = 1$. The error bars are the 95% CI of the recall for $B = 100$ experiments.

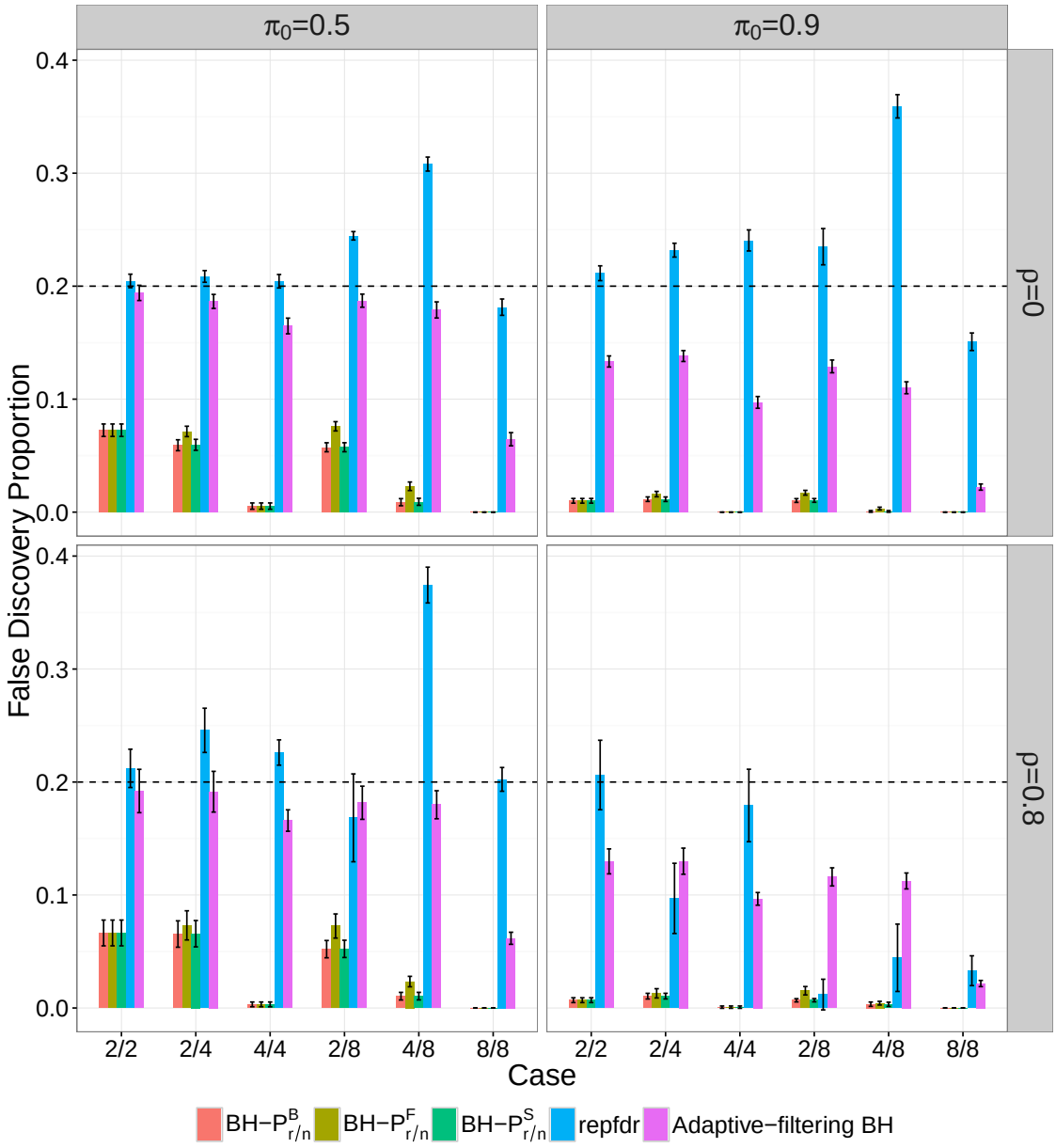


Figure 5: Comparison of false discoveries rate $E(V/R)$ (FDR). The dotted line indicates the nominal level $\alpha = 0.2$. The error bars are the 95% CI of estimated FDR.

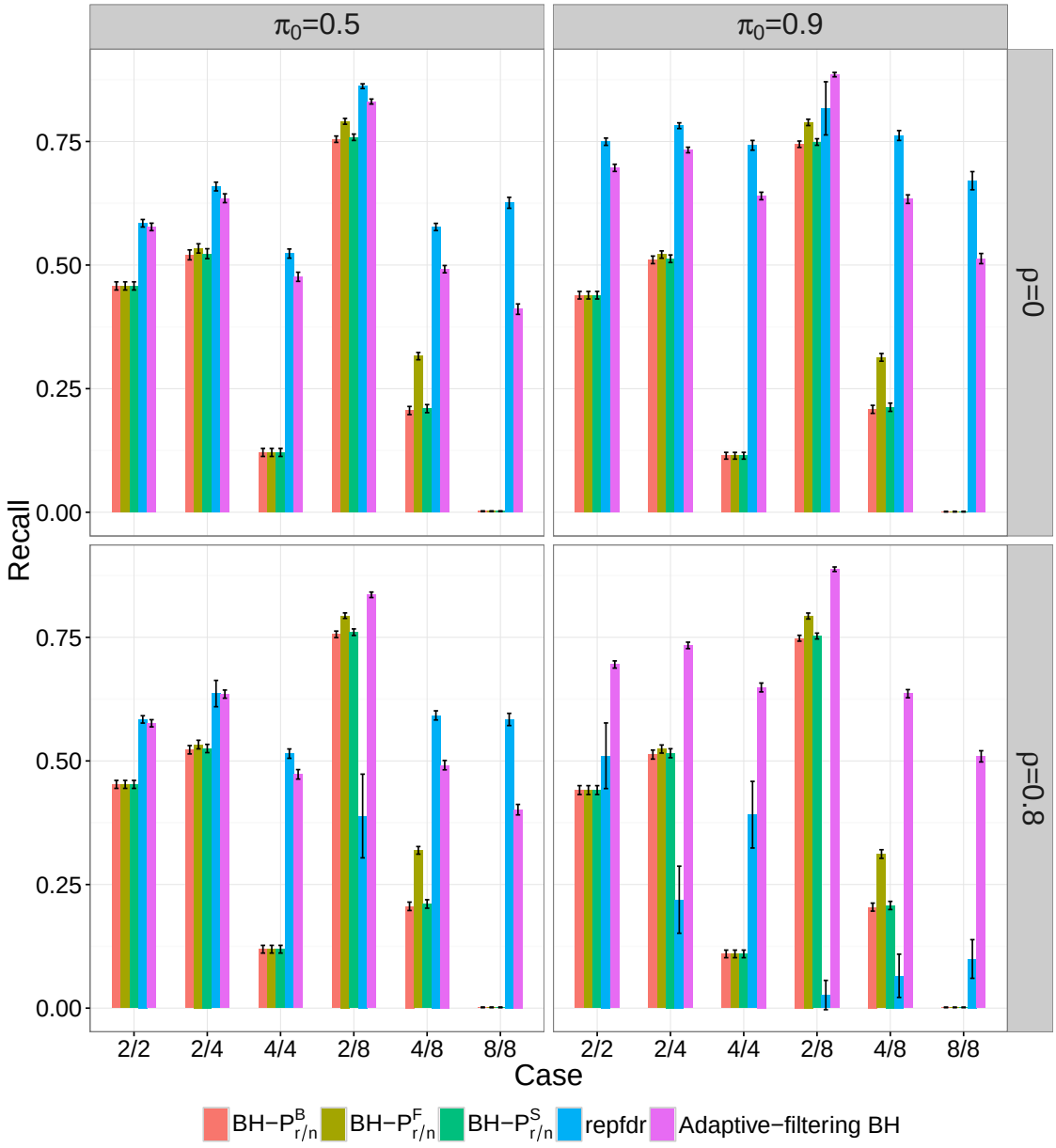


Figure 6: Comparison of power (recall or sensitivity) when FDR is controlled at level $\alpha = 0.2$. The error bars are the 95% CI of the recall for $B = 100$ experiments.

using the classical procedures. The power comparison between the adaptive filtering BH procedure and classical procedures under FDR control shows a very similar pattern. Unfortunately, it was not feasible to apply repfdr to this dataset as it needs to estimate the prevalance of $2^{16}-1$ different types of alternative hypothesis and there are only 8932 probes in the sample.

# A Appendix: Proofs

Here we provide proofs for all the theoretical results in Sections 4 and 5

## A.1 Proof of 4.1

To show that the two definitions are equivalent, we only need to show that $m = \gamma_0^{\text{Bon}}$ always holds. Define

$$\mathcal{S} = \left\{\gamma \in [0,1] : \gamma \cdot \sum_{j=1}^{M} 1_{F_j \le \gamma} \le \alpha\right\}$$

First consider the case when the set in (6) is not $\varnothing$. As $F_{(m')} > \alpha/m'$, it can be shown that for any $\gamma \ge F_{(m')}$ we have $\gamma \notin \mathcal{S}$. This is because both $\gamma > \alpha/m'$ and $\sum_{j=1}^{M} 1_{F_j \le \gamma} \ge m'$. Also, as $F_{(m'-1)} \le \alpha/(m'-1)$, it can be shown that for any $\gamma > \alpha/(m'-1)$ we have $\gamma \notin \mathcal{S}$. This is because both $\gamma > \alpha/(m'-1)$ and $\sum_{j=1}^{M} 1_{F_j \le \gamma} \ge m'-1$.

As a consequence, when $F_{(m')} > \alpha/(m'-1)$, it can be easily checked that $\alpha/(m'-1) \in \mathcal{S}$. Thus, $\gamma_0^{\text{Bon}} = \alpha/(m'-1) = \alpha/m$. On the other hand, when $F_{(m')} \le \alpha/(m'-1)$, we have $\gamma_0^{\text{Bon}} \le \alpha/m' = \alpha/m$. It can also be easily checked that $\alpha/m \in \mathcal{S}$. In summary, $\gamma_0^{\text{Bon}} = \alpha/m$.

Finally, when the set in (6) is $\varnothing$, it means that $\alpha/M \ge F_{(M)}$. In this scenario, $\gamma_0^{\text{Bon}} = \alpha/M = \alpha/m$ as we set $m = M$ in 6.

## A.2 Proof of 5.3

Choose $\beta > 0$ and let $\tilde{\beta} = \beta/(n-r+1)$. By independence of $P_{ij}$,

$$\mathbb{P}(P_{(r)j} \le \tilde{\beta}) = \sum_{k=r}^{n} \sum_{|u|=k} \prod_{i \in u} \mathbb{P}(P_{ij} \le \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}), \quad \text{and}$$

$$\mathbb{P}(P_{(r-1)j} \le \tilde{\beta}) = \sum_{k=r-1}^{n} \sum_{|u|=k} \prod_{i \in u} \mathbb{P}(P_{ij} \le \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}).$$

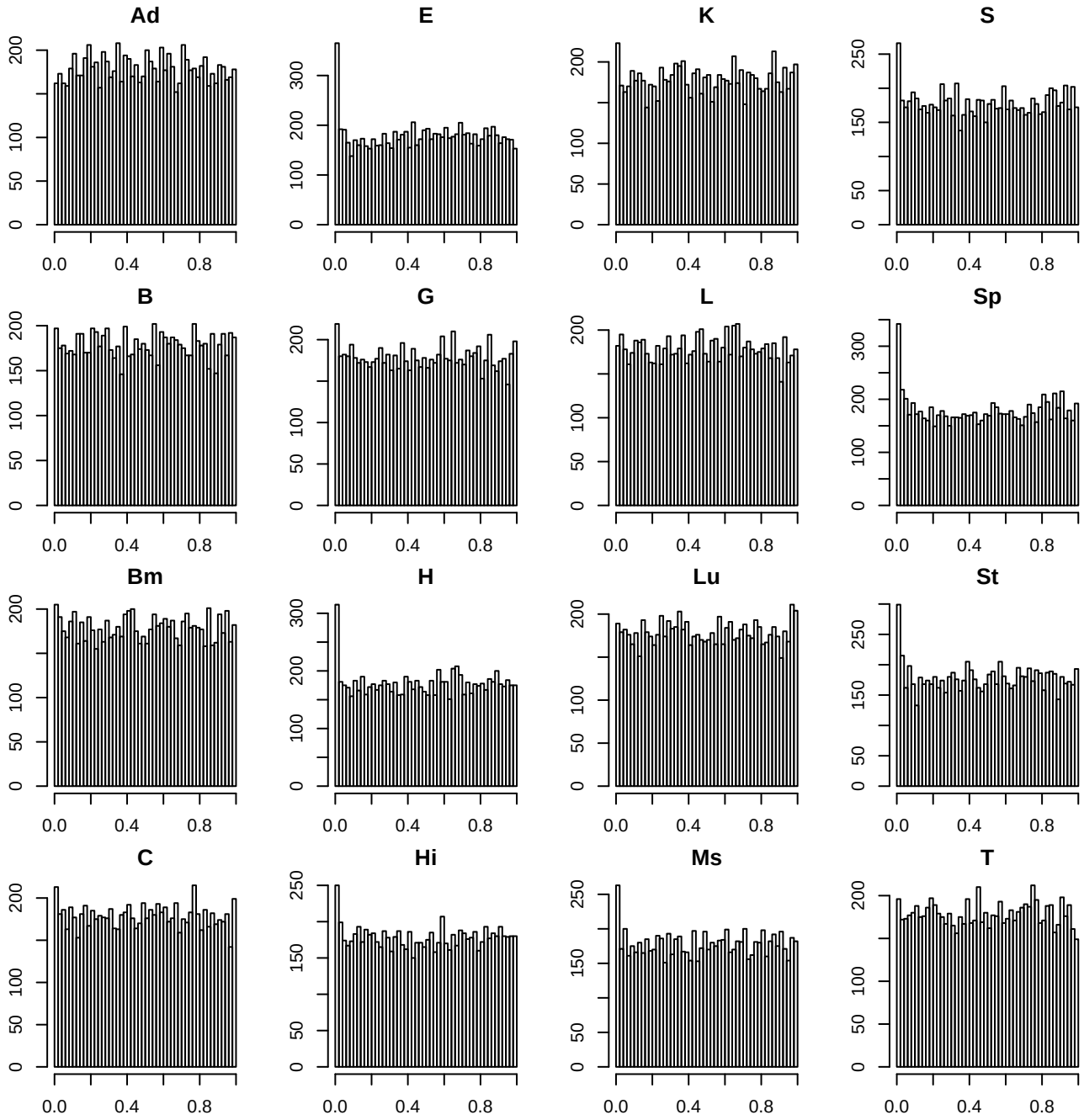


Figure 7: Adjusted p-values for each of the 16 tissues in the AGEMAP data.

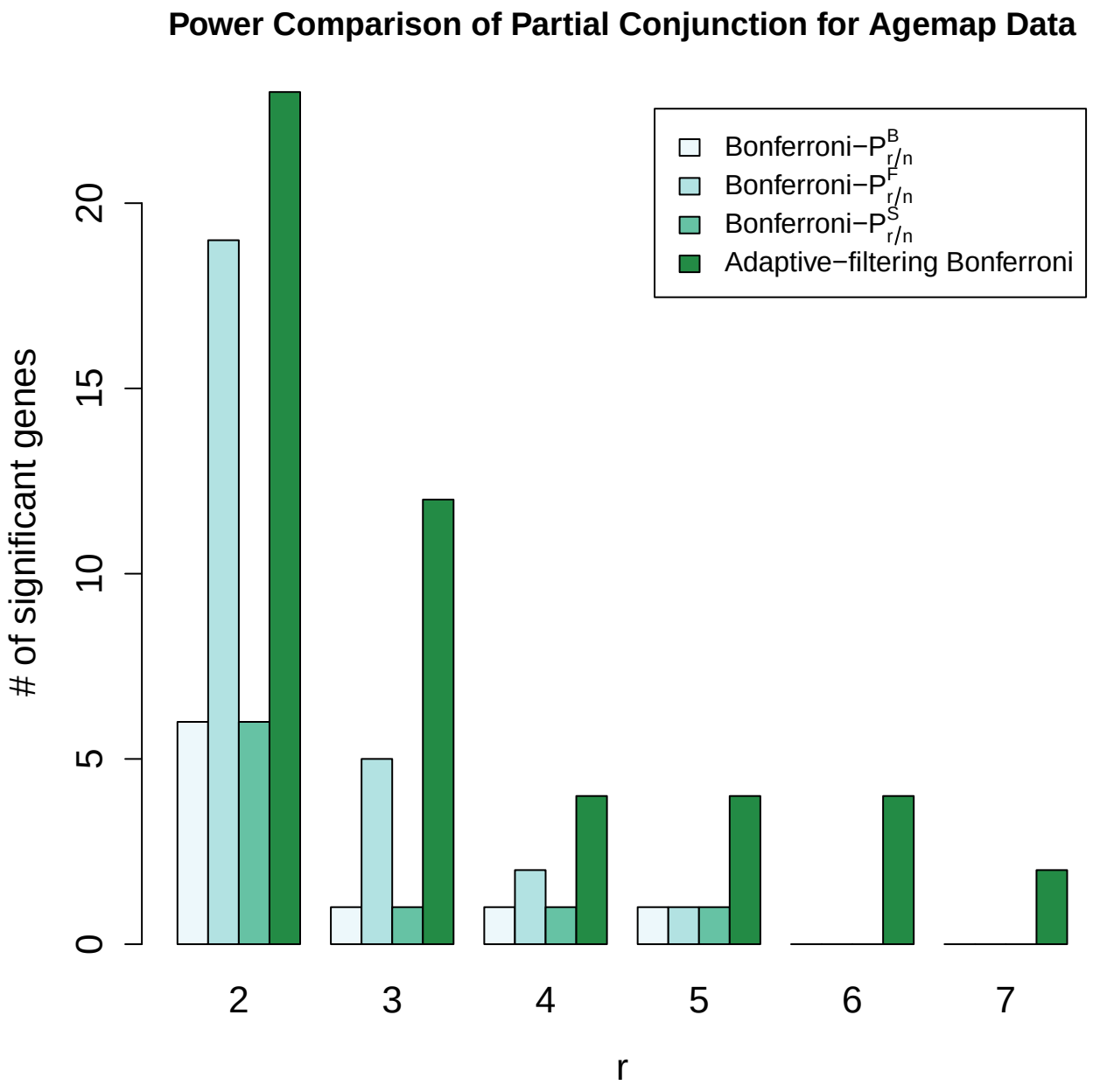


Figure 8: The number of genes whose null hypotheses were rejected by the four PC procedures. Here $r$ ranges from 2 to 7 and PFER is controlled at $\alpha = 1$

Next, because $H_{0j}^{r/n}$ is true, for any $u \subset 1{:}n$ with $|u| \geq r$ there is at least one index $i^* = i^*(u,j) \in u$ for which $\theta_{i^*j} \in \boldsymbol{\Theta}_{0i^*j}$. Because the $P_{ij}$ are valid,

$$\begin{aligned}
\mathbb{P}(S_j \leq \beta) &= \mathbb{P}(P_{(r)j} \leq \tilde{\beta}) \\
&= \sum_{k=r}^{n} \sum_{|u|=k} \Big( \prod_{i \in u} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&\leq \tilde{\beta} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \Big( \prod_{i \in u \setminus \{i^*\}} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&= \tilde{\beta} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \Big( \prod_{i \in u} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&+ \ \tilde{\beta} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \Big( \prod_{i \in u \setminus \{i^*\}} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u \cup \{i^*\}} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&\leq \tilde{\beta} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \Big( \prod_{i \in u} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&+ \ \tilde{\beta} \cdot \sum_{k=r-1}^{n-1} (n-k) \sum_{|u|=k} \Big( \prod_{i \in u} \mathbb{P}(P_{ij} \leq \tilde{\beta}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\beta}) \Big) \\
&\leq (n-r+1)\tilde{\beta} \cdot \mathbb{P}(P_{(r-1)j} \leq \tilde{\beta}).
\end{aligned}$$

Thus

$$\mathbb{P}\big(S_j \leq \beta \mid F_j \leq \beta\big) = \frac{\mathbb{P}(P_{(r)j} \leq \tilde{\beta})}{\mathbb{P}(P_{(r-1)j} \leq \tilde{\beta})} \leq \beta.$$

## A.3 Proof of Theorem 5.1

For $j = 1, 2, \cdots, M$, define

$$\gamma_j = \max\Big\{\gamma \in \big\{\alpha, \frac{\alpha}{2}, \cdots, \frac{\alpha}{M}\big\} : \gamma \cdot (1 + \sum_{s \neq j} 1_{F_s \leq \gamma}) \leq \alpha\Big\}.$$

It is obvious that if $F_j \leq \gamma_0^{\text{Bon}}$, then $\gamma_j = \gamma_0^{\text{Bon}}$, otherwise $\gamma_j \leq \gamma_0^{\text{Bon}}$. Then the PFER is,

$$\mathbb{E}(V) = \mathbb{E}\Big( \sum_{j=1}^{M} 1_{S_j \leq \gamma_0^{\text{Bon}}} \cdot 1_{v_j = 0} \Big)$$

$$= \mathbb{E}\Big( \sum_{j=1}^{M} 1_{S_j \le \gamma_0^{\mathrm{Bon}}} 1_{F_j \le \gamma_0^{\mathrm{Bon}}} \cdot 1_{v_j=0} \Big)$$
$$= \sum_{j=1}^{M} \mathbb{E}\Big( 1_{S_j \le \gamma_j} \cdot 1_{F_j \le \gamma_0^{\mathrm{Bon}}} \cdot 1_{v_j=0} \Big)$$
$$= \sum_{j=1}^{M} \mathbb{E}\Big( 1_{S_j \le \gamma_j} \cdot 1_{F_j \le \gamma_j} \Big) \cdot 1_{v_j=0}$$

Recall that $\gamma_j$ does not depend on $P_{\cdot j}$ while $(F_j, S_j)$ only depends on $P_{\cdot j}$. Therefore $\gamma_j$ is independent of $(F_j, S_j)$ by our assumption on $(P_{ij})$. Now using the conditional validity 5.3,

$$\mathbb{E}(V) = \mathbb{E}\left( \sum_{j=1}^{M} \mathbb{E}\Big[ 1_{S_j \le \gamma_j} \mid \gamma_j, 1_{F_j \le \gamma_j} \Big] \cdot 1_{F_j \le \gamma_j} \cdot 1_{v_j=0} \right)$$
$$\le \mathbb{E}\Big( \sum_{j=1}^{M} \gamma_j \cdot 1_{F_j \le \gamma_j} \cdot 1_{v_j=0} \Big)$$
$$\le \mathbb{E}\Big( \gamma_0^{\mathrm{Bon}} \cdot \sum_{j=1}^{M} 1_{F_j \le \gamma_0^{\mathrm{Bon}}} \cdot 1_{v_j=0} \Big) \le \alpha.$$

## A.4 Proof of Theorem 5.2

For each $j = 1, 2, \cdots, M$ define

$$\gamma_j = \max\Big\{ \gamma \in \mathcal{I}_{\alpha,M} : \gamma \cdot (1 + \sum_{k \ne j}^{M} 1_{F_k \le \gamma}) \le \alpha \cdot (1 + \sum_{k \ne j}^{M} 1_{S_k \le \gamma}) \Big\}.$$

It's obvious that if $S_j \le \gamma_0^{\mathrm{BH}}$ then also $F_j \le S_j \le \gamma_0^{\mathrm{BH}}$, thus $\gamma_j = \gamma_0^{\mathrm{BH}}$. If $F_j \le \gamma_0^{\mathrm{BH}} < S_j$, then $\gamma_j \ge \gamma_0^{\mathrm{BH}}$. Finally as $\gamma_0^{\mathrm{BH}} \le \alpha$, if $F_j > \gamma_0^{\mathrm{BH}}$, then also $\gamma_j \ge \gamma_0^{\mathrm{BH}}$. In summary, $\gamma_j \ge \gamma_0^{\mathrm{BH}}$ and when $S_j \le \gamma_0^{\mathrm{BH}}$ the equality holds.

Then the FDR is,

$$\begin{aligned}
\mathbb{E}\left(\frac{V}{\max(R,1)}\right) &= \mathbb{E}\left(\frac{\sum_{j=1}^{M} 1_{S_j \le \gamma_0^{\text{BH}}} \cdot 1_{v_j=0}}{\max\left(\sum_{j=1}^{M} 1_{S_j \le \gamma_0^{\text{BH}}}, 1\right)}\right) \\
&\le \alpha \cdot \mathbb{E}\left(\frac{\sum_{j=1}^{M} 1_{S_j \le \gamma_0^{\text{BH}}} \cdot 1_{v_j=0}}{\max\left(\gamma_0^{\text{BH}} \cdot \sum_{j=1}^{M} 1_{F_j \le \gamma_0^{\text{BH}}}, \alpha\right)}\right) \\
&= \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{1_{S_j \le \gamma_0^{\text{BH}}}}{\max\left(\gamma_0^{\text{BH}} \cdot \sum_{k=1}^{M} 1_{F_k \le \gamma_0^{\text{BH}}}, \alpha\right)}\right) 1_{v_j=0} \\
&\le \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{1_{S_j \le \gamma_0^{\text{BH}}} \cdot 1_{S_j \le \gamma_j}}{\max\left(\gamma_j \cdot (1 + \sum_{k \ne j}^{M} 1_{F_k \le \gamma_j}), \alpha\right)}\right) \\
&\le \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{1_{S_j \le \gamma_j}}{\max\left(\gamma_j \cdot (1 + \sum_{k \ne j}^{M} 1_{F_k \le \gamma_j}), \alpha\right)}\right) 1_{v_j=0}
\end{aligned}$$

As $\gamma_j$ does not depend on $P_{\cdot j}$ while $(F_j, S_j)$ only depend on $P_{\cdot j}$, for $v_j = 0$ we have that $(F_j, S_j)$ are independent from $\gamma_j$ and $P_{\cdot(-j)}$ under the independence assumption of the individual p-values. Thus using 5.3,

$$\begin{aligned}
\mathbb{E}\left(\frac{V}{\max(R,1)}\right) &\le \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\mathbb{E}\left[\frac{1_{S_j \le \gamma_j}}{\max\left(\gamma_j \cdot (1 + \sum_{k \ne j}^{M} 1_{F_k \le \gamma_j}), \alpha\right)} \mid P_{\cdot(-j)}\right]\right) 1_{v_j=0} \\
&= \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{\mathbb{E}[1_{S_j \le \gamma_j} \mid P_{\cdot(-j)}]}{\max\left(\gamma_j \cdot (1 + \sum_{k \ne j} 1_{F_k \le \gamma_j}), \alpha\right)}\right) 1_{v_j=0} \\
&\le \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{\gamma_j \cdot \mathbb{E}[1_{F_j \le \gamma_j} \mid P_{\cdot(-j)}]}{\max\left(\gamma_j \cdot (1 + \sum_{k \ne j} 1_{F_k \le \gamma_j}), \alpha\right)}\right) 1_{v_j=0} \\
&\le \alpha \cdot \sum_{j=1}^{M} \mathbb{E}\left(\frac{1_{F_j \le \gamma_j}}{1 + \sum_{k \ne j} 1_{F_k \le \gamma_j}}\right) 1_{v_j=0} \\
&= \alpha \cdot \mathbb{E}\left(\frac{\sum_{j=1}^{M} 1_{F_j \le \gamma_j} 1_{v_j=0}}{\sum_{k=1}^{M} 1_{F_k \le \gamma_j}}\right) \le \alpha.
\end{aligned}$$

## A.5 Proof of 5.4

For some $j$, let $\tilde{p}_{\cdot j} = (\tilde{p}_{1j}, \cdots, \tilde{p}_{nj})$ satisfy $\tilde{p}_{ij} \leq p_{ij}$ for $i = 1, 2, \cdots, n$. Now construct a new $N \times n$ $P$-value matrix $\tilde{P}$ with the given row $\tilde{P}_{\cdot j}$ and all other rows $\tilde{P}_{\cdot k} = P_{\cdot k}$ for $k \neq j$. Define $(\tilde{F}_1, \cdots, \tilde{F}_M)$ as the corresponding filtering statistics (4) and $(\tilde{S}_1, \cdots, \tilde{S}_M)$ as the corresponding selection statistics (5) with $\tilde{P}$ replacing $P$. Then $\tilde{F}_k = F_k$ and $\tilde{S}_k = S_k$ for $k \neq j$ and $\tilde{F}_j \leq F_j$ with $\tilde{S}_j \leq S_j$.

For the adaptive filtering Bonferroni procedure, let $\tilde{\gamma}_j^{\text{Bon}}$ be the new $\gamma_0^{\text{Bon}}$ using the new individual P-values. For the adaptive filtering BH procedure, let $\tilde{\gamma}_j^{\text{BH}}$ be the new $\gamma_0^{\text{BH}}$ using the new individual p-values. Then to show that the procedures satisfy partial monotonicity, we only need to show that if $S_j \leq \gamma_0$, then $\tilde{S}_j \leq \tilde{\gamma}_j$ for both the Bonferroni correction and BH.

For adaptive filtering BH procedure, if $S_j \leq \gamma_0$, then $\tilde{S}_j \leq \gamma_0$, thus

$$\gamma_0^{\text{Bon}} \cdot \sum_{k=1}^{M} 1_{\tilde{F}_k \leq \gamma_0^{\text{Bon}}} \leq \alpha$$

which means that $\tilde{\gamma}_j^{\text{Bon}} \geq \gamma_0^{\text{Bon}}$. Similarly, for adaptive filtering BH procedure using the same argument, we have $\tilde{\gamma}_j^{\text{BH}} \geq \gamma_0^{\text{BH}}$ when $S_j \leq \gamma_0$. As a consequence, for both adaptive filtering procedures, we have $\tilde{S}_j \leq S_j \leq \gamma_0 \leq \tilde{\gamma}_j$.

# References


Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 533*(7604), 452–454.

Benjamini, Y. and R. Heller (2008). Screening for partial conjunction hypotheses. *Biometrics 64*(4), 1215–1222.

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

Bogomolov, M. and R. Heller (2015). Assessing replicability of findings across two studies of multiple features. Technical report, arXiv:1504.00534.

Dudoit, S. and M. J. Van Der Laan (2007). *Multiple testing procedures with applications to genomics*. Springer Science & Business Media.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics 9*(5), e1003486.

Friston, K. J., W. D. Penny, and D. E. Glaser (2005). Conjunction revisited. *NeuroImage 25*(3), 661–667.

Heller, R., Y. Golland, R. Malach, and Y. Benjamini (2007). Conjunction group analysis: an alternative to mixed/random effect analysis. *Neuroimage 37*(4), 1178–1185.

Heller, R. and D. Yekutieli (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics 8*(1), 481–498.

Moonesinghe, R., M. J. Khoury, and A. C. J. W. Janssens (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Medicine 4*(2), e28.

Price, C. J. and K. J. Friston (1997). Cognitive conjunction: A new approach to brain activation experiments. *NeuroImage 5*(4), 261–270.

Sun, W., B. J. Reich, T. Cai, M. Guindani, and A. Schwartzman (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B 77*(1), 59–83.

Sun, W. and Z. Wei (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association 106*(493), 73–88.

Sun, Y., N. R. Zhang, and A. B. Owen (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *The Annals of Applied Statistics 6*(4), 1664–1688.

Tukey, J. W. (1953). The problem of multiple comparisons. Technical report, Princeton University.

Wang, J. and A. B. Owen (2015). Admissibility in partial conjunction testing. Technical report, arXiv:1508.00934.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2015). Confounder adjustment in multiple hypotheses testing. *Annals of Statistics 44*, (to appear).

Zahn, J., S. Poosala, A. B. Owen, D. K. Ingram, A. Lustig, A. Carter, A. T. Weeratna, D. D. Taub, M. Gorospe, K. Mazan-Mamczarz, E. G. Lakatta, K. R. Boheler, X. Xu, M. P. Mattson, G. Falco, M. S. H. Ko, D. Schlessinger, J. Firman, S. K. Kummerfeld, W. H. W. III, A. B.

Zonderman, S. K. Kim, and K. G. Becker (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genetics 3*(11), 2326–2337.