Detecting Multiple Replicating Signals using Adaptive Filtering Procedures

Jingshu Wang∗1, Lin Gui1, Weijie J. Su2, Chiara Sabatti3, and Art B. Owen3

1Department of Statistics, The University of Chicago
2Department of Statistics, University of Pennsylvania
3Department of Statistics, Stanford University

May 2020

Abstract

Replicability is a fundamental quality of scientific discoveries: we are interested in those signals that are detectable in different laboratories, study populations, across time etc. Unlike meta-analysis which accounts for experimental variability but does not guarantee replicability, testing a partial conjunction (PC) null aims specifically to identify the signals that are discovered in multiple studies. In many contemporary applications, e.g., comparing multiple high-throughput genetic experiments, a large number $M$ of PC nulls need to be tested simultaneously, calling for a multiple comparisons correction. However, standard multiple testing adjustments on the $M$ PC $p$-values can be severely conservative, especially when $M$ is large and the signals are sparse. We introduce AdaFilter, a new multiple testing procedure that increases power by adaptively filtering out unlikely candidates of PC nulls. We prove that AdaFilter can control FWER and FDR as long as data across studies are independent, and has much higher power than other existing methods. We illustrate the application of AdaFilter with three examples: microarray studies of Duchenne muscular dystrophy, single-cell RNA sequencing of T cells in lung cancer tumors and GWAS for metabolomics.

Keywords: simultaneous signals, meta-analysis, high-throughput experiments, partial conjunction, composite null, multiple hypotheses testing

∗jingshuw@uchicago.edu
1 Introduction

Replication is “the cornerstone of science” [31]. An important scientific finding should be supported by further evidence from similar conditions, by other researchers or with new samples. In the last decade, however, both the popular [27] and the scientific press [4, 2] have reported the lack of replicability in modern research. While there are many reasons behind this phenomenon, one important factor is that many scientific discoveries are obtained from complicated large-scale experiments with biases from various sources. Even when the data are carefully analyzed, idiosyncratic aspects of a single experiment can fail to extend to other settings, and any finding from just one study can easily lack external validity. Thus, it is crucial to have a statistical framework to objectively and precisely evaluate the consistency of scientific discoveries across multiple studies, while properly accounting for experimental heterogeneity.

The partial conjunction (PC) test, which was introduced by [15] and further studied in [5], provides such a framework. Given \( n \) null hypotheses (base nulls) and a number \( r \in \{2, 3, \ldots, n\} \), the PC null states that there are fewer than \( r \) base non-nulls. In the setting where each hypothesis represents a test from one study, rejecting a PC null explicitly guarantees that the signal replicates at least \( r \) times. The PC framework has been used to identify replicating signals in neuroimaging [33, 5], across genetic experiments [19], and recently to find evidence factors in causal inference [25].

In high-throughput genetic experiments, there is a special need to identify replicating signals across multiple studies. For instance, for gene expression data, it is important to find stable gene markers for a disease or cell type, which remain differentially expressed across similar experiments or in multiple patients. In multi-tissue expression quantitative trait loci (eQTL) studies, scientists are interested in identifying DNA loci with consistent regulation over tissues [14, 41]. With a growing trend in multi-omics data sharing [18], there is also active research in finding replicating signals across platforms [44], ethnic groups [30, 16] and even species. Though the PC framework fits all above scenarios, finding multiple replicating signals by simultaneously performing a large number of PC tests for thousands of genes or millions of DNA loci, however, typically suffers from extremely low power.

Specifically, let \( M \) denote the number of hypotheses in one study and suppose that we compare across \( n \) related studies. Then, to find replicating signals across the \( n \) studies, we have \( M \) PC nulls to test, each with \( n \) base nulls. The above framework gives us an \( n \times M \) matrix of base \( p \)-values, with one column per PC null and one row per study. Now, as we want to identify signals whose PC nulls are false, a “direct approach” is to first get a combined \( p \)-value for each PC null and then apply multiple testing adjustment to the \( M \) PC \( p \)-values. However, this “direct approach” for testing multiple PC tests has been shown to have extremely low power [21, 39]. Both [21] and [9] suggest procedures to counter that power loss. Unfortunately, the approach in [9] is designed only for \( n = r = 2 \) and the empirical Bayes approach
repfdr in [21] encounters both accuracy and computational barriers for \( n \) as large as 8, as shown in our simulations. There is thus a need for a powerful and fast method that can guarantee simultaneous error control and can handle a larger number of studies.

In this paper, we introduce AdaFilter, an adaptive filtering multiple testing procedure for multiple PC hypotheses. We propose different versions of AdaFilter to control simultaneous error rates including FDR (false discovery rate) and FWER (familywise error rate). AdaFilter can control FWER and FDR when all \( nM \) base \( p \)-values are independent. In addition, it asymptotically controls FDR when \( M \) goes to infinity, allowing base \( p \)-values to be weakly associated within each study. The weak dependence only assumes that within each study, the number of pairs \((i, j)\) where the base \( p \)-values \( P_i \) and \( P_j \) are dependent is \( o(M^2) \), which is reasonable for most genetics and genomics data. Using simulations and real data applications, we show that AdaFilter can have much higher power than the “direct approaches” or using repfdr.

Deferring precise statements to later sections, we give an intuitive explanation for how AdaFilter gains power. The low power of the “direct approach” is due to the fact that partial conjunction has a composite null. AdaFilter’s power gain is linked to its ability to borrow information across studies and learn from the data which PC hypotheses are likely to be least favorable nulls. Intuitively, AdaFilter filters the set of hypotheses down to a number \( m < M \) of candidate least favorable nulls, which are the nulls that have exactly \( r - 1 \) base non-nulls. The PC \( p \)-values are still “valid” conditioning on filtering and the decreased number of hypotheses lowers multiplicity burden. More surprisingly, the power gain also links to a lack of “monotonicity” of the number rejections in the base \( p \)-values, where increasing some base \( p \)-values can result in more rejections. In the extreme case, combining multiple studies while requiring replicability can even leads to more rejections than the union of rejections by testing each individual study separately.

The structure of the paper is as follows. Section 2 precisely defines the PC framework, and illustrates the power limitation of the “direct approach”. Section 3 introduces our AdaFilter procedures. Section 4 discusses theoretical properties of AdaFilter. Section 5 explores the performance with simulations. Section 6 applies AdaFilter to three studies: replication in microarray experiments from different platforms, identifying consistent marker genes in single-cell RNA-sequencing (scRNA-seq), and signal consistency across different Genome-wide association study (GWAS) subgroups. Section 7 has conclusions. An R package implementing AdaFilter is available at: https://github.com/jingshuw/adaFilter.
2 Multiple testing for partial conjunctions

2.1 Problem setup

We consider the problem where $M$ null hypotheses are tested in $n$ studies. The base null hypotheses are $(H_{0ij})_{n \times M}$. In high-throughput experiments, $M$ is the number of genes or DNA loci. We work with summary statistics that are base p-values $(p_{ij})_{n \times M}$ for $(H_{0ij})_{n \times M}$. Each $p_{ij}$ is the realization of a random variable $P_{ij}$. We assume that individual P-values are valid, satisfying $\Pr(P \leq \gamma) \leq \gamma$ under its null. Also, let $P_{(1)j} \leq P_{(2)j} \leq \cdots \leq P_{(nj)}$ be the sorted P-values for each $j = 1, 2, \ldots, M$.

Definition 1. (Partial Conjunction Hypothesis) For integers $n \geq r \geq 2$, the partial conjunction (PC) null hypothesis is:

$$H_{0r/n}^r : \text{fewer than } r \text{ out of } n \text{ base hypotheses are non-null.}$$

When $r = 1$, $H_{01/n}^1$ is the commonly tested null for meta-analysis. Rejecting it would not guarantee replicability. In high-throughput experiments, for each DNA locus or gene $j \in 1:M \equiv \{1, 2, \ldots, M\}$, we test for a PC null $H_{0j}^{r/n}$ to evaluate if genetic signals have been replicated at least $r$ times across $n$ studies. Throughout the paper, we assume that p-values across studies are independent. This can be assumed when samples do not overlap across studies.

For a multiple testing procedure on $\{H_{01}^{r/n}, \ldots, H_{0M}^{r/n}\}$, denote the decision function as $\varphi_j = 1$ if we reject $H_{0j}^{r/n}$ and $\varphi_j = 0$ otherwise. The total number of discoveries is then $R = \sum_{j=1}^{M} \varphi_j$. Among these, the number of false discoveries is $V = \sum_{j=1}^{M} \varphi_j 1_{v_j=0}$ where $v_j = 0$ if $H_{0j}^{r/n}$ is true and $v_j = 1$ otherwise.

There are many measures of the simultaneous error rate [11], with FWER and FDR being the most common ones. In addition, we consider the per-family error rate (PFER), as it provides a motivation for our procedures. With the notation introduced, we have

$$\text{FWER} := \Pr(V \geq 1), \quad \text{PFER} := \mathbb{E}(V), \quad \text{FDR} := \mathbb{E}(\text{FDP}).$$

where $\text{FDP} = V/(R \lor 1)$ is the false discovery proportion.

2.2 The “direct approach”

We start with a brief review of p-value construction for a single PC null, while more details can be found in [42] and [5]. Consider a single PC null $H_{0j}^{r/n}$ with a vector of base P-values $(P_1, P_2, \ldots, P_n)$ and let $P_{r/n} = f(P_1, P_2, \ldots, P_n)$ be the combined P-value for $H_{0j}^{r/n}$. Benjamini and Heller [5] discuss three approaches, which we report here, using the standard notation $(P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)})$:
1. Simes’ method:

\[ P_{r/n}^S = \min_{r \leq i \leq n} \left\{ \frac{n-r+1}{i-r+1} P(i) \right\}, \]

2. Fisher’s method:

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

3. Bonferroni’s method:

\[ P_{r/n}^B = (n-r+1) P(r). \]

The idea is to apply meta-analysis to the largest \( n - r + 1 \) individual P-values. For instance, if \( n = r = 2 \), then \( P_{2/2}^S = P_{2/2}^F = P_{2/2}^B = \max(p_1, p_2) \). All three methods construct valid PC P-values for \( H_{r/n}^{0} \) under independence, and [42] showed that they also provide the most powerful tests for a single PC null. For \( M \) hypotheses, we denote \( P_{r/n,j} \) as the PC p-value for the \( j \)th PC null.

The “direct approach” is to simply apply standard multiple testing adjustment procedures to the \( M \) PC P-values. For example, to control the FWER at level \( \alpha \), we could use the Bonferroni rule, rejecting \( H_{0,j}^{n} \) if \( P_{r/n,j} \leq \alpha/M \), which also controls the PFER at level \( \alpha \) [40]. To control the FDR we could apply BH procedure [6] on \( \{P_{r/n,j}, j = 1, \cdots, M\} \).

However, this direct approach is often too conservative, as we illustrate now for the case \( r = n \). To quantify how the performance associates with the composite nature of a PC null, define sets \( \mathcal{I}_k \subset 1:M \) such that

\[ \mathcal{I}_k = \{ j \in 1:M \mid \text{exactly } k \text{ of } H_{01j}, \ldots, H_{0nj} \text{ are false} \} \]

for \( k = 0, \ldots, n \). If a false rejection of \( H_{0j}^{n} \) happens, then the \( j \)th column must belong to one of \( \mathcal{I}_k \) where \( k = 0, 1, \cdots, n - 1 \). Thus, if we use Bonferroni to control for FWER at a nominal level \( \alpha \), the true FWER instead satisfies

\[ \text{FWER} \leq \mathbb{E}(V) = \sum_{k=0}^{n-1} \sum_{j \in \mathcal{I}_k} \mathbb{P}(P(j) \leq \alpha/M) \]

\[ \leq \sum_{k=0}^{n-1} \sum_{j \in \mathcal{I}_k} \frac{\alpha^{n-k}}{M^{n-k}} = \sum_{k=0}^{n-1} |\mathcal{I}_k| \frac{\alpha^{n-k}}{M^{n-k}}. \]

where the second inequality is close to an equality when all the tests for non-nulls \( H_{1ij} \) have high power. Let \( \delta_k = |\mathcal{I}_k|/M \) be the proportion of hypotheses in each partition. Then when \( M \) is large, the above bound is approximately

\[ \mathbb{E}(V) \approx \alpha \left\{ \delta_{n-1} + \delta_{n-2} \frac{\alpha}{M} + \delta_{n-3} \left( \frac{\alpha}{M} \right)^2 + \cdots + \delta_0 \left( \frac{\alpha}{M} \right)^n \right\} \]
which in the limit is dominated by \( \delta_{n-1} \alpha \) (when \( \delta_{n-1} \neq 0 \)) or is of order \( O(M^{-1}) \) (when \( \delta_{n-1} = 0 \)). Thus, when \( \delta_{n-1} \approx 0 \), a typical scenario in genetics problems with sparse signal, the expected number of rejections \( \mathbb{E}(V) \) would be much smaller than \( \alpha \) and the “direct approach” can become highly deficient, in fact much more conservative than Bonferroni usually is.

The point is that if we do not account for the fact that the PC null is composite, we will control the simultaneous error rates under the worst case scenario (\( \delta_{n-1} = 1 \)), which is unnecessary. For general \( r \leq n \), the level of \( \mathbb{E}(V) \) for Bonferroni correction will depend mainly on \( \delta_{r-1} \) in the large \( M \) setting. So does the BH control for FDR.

It is clear that there can be more efficient procedures if the fractions \( \delta_k \) were known or if good estimates of \( \delta_k \) can be obtained. This is what motivates the Bayesian methods [21, 14]. In this paper we take a frequentist perspective. Rather than estimating \( \delta_k \), AdaFilter works directly on an alternative estimation of \( V \), implicitly and adaptively adjusting for the size of \( \delta_{r-1} \), the fraction of the least favorable nulls.

3 The idea of AdaFilter

In Section 2.2, we showed that a PC null hypothesis is composite, thus the inequality \( \mathbb{P}(P_{r/n} \leq \gamma) \leq \gamma \) for a given \( \gamma \) is only tight for the least favorable null, while standard multiple testing procedures are designed to control error when \( \mathbb{P}(P_{r/n} \leq \gamma) = \gamma \) is always true. To overcome this, AdaFilter leverages a region \( A_\gamma \subset [0,1]^n \) such that the much tighter inequality

\[
\mathbb{P}(P_{r/n,j} \leq \gamma \mid (P_{1j}, \ldots, P_{nj}) \in A_\gamma) \leq \gamma
\]

holds for any configuration in the PC null space.

Figure 1a illustrates the construction of the filtering region \( A_\gamma \) for \( r = n = 2 \). The PC test \( j \) has base \( p \)-values \( P_{1j} \) and \( P_{2j} \), and its PC \( p \)-value is \( P_{2/2,j} = \max(P_{1j}, P_{2j}) \).

The null \( H_{j0}^{2/2} \) contains three configurations: \( (H_{01j}, H_{02j}) \) being (True, True), (True, False) or (False, True). It is easy to see that \( \mathbb{P}(P_{2/2,j} \leq \gamma) = \gamma^2 \) under (True, True), while \( \mathbb{P}(P_{2/2,j} \leq \gamma) \) can be close of \( \gamma \) under the other two less favorable configuration. Let us consider, instead, conditioning on \( (P_{1j}, P_{2j}) \) being in the “L”-shaped filtering region \( A_\gamma = \{(p_1, p_2) \mid \min(p_1, p_2) \leq \gamma \} \). We get \( \mathbb{P}(P_{2/2,j} \leq \gamma \mid (P_{1j}, P_{2j}) \in A_\gamma) \leq \gamma \) being true for all three null scenarios, which is much tighter than \( \mathbb{P}(P_{r/n} \leq \gamma) \leq \gamma \).

The inequality holds since at least one of \( P_{1j} \) and \( P_{2j} \) is stochastically greater than uniform under all three configurations.

Since Bonferroni and BH procedures are based on an implicit estimate of the number of false rejections \( V \) associated with a threshold \( \gamma \): \( \hat{V}_\gamma = \gamma M \), we can improve their efficiency with a smaller estimate of \( \hat{V}_\gamma \) using the new inequality. Specifically, the estimated \( V \) is now \( \hat{V}_{A_\gamma} = \gamma \times \sum_{j=1}^{M} 1_{(P_{1j}, P_{2j}) \in A_\gamma} \), where \( M \) is replaced by the number
of hypotheses falling into the L shaped region, a possibly much smaller number than $M$. Alternatively, the quantity \( \frac{1}{M} \sum_{j=1}^{M} 1_{(p_{1j}, p_{2j}) \in A_{\gamma}} \) is our “estimate” of \( \delta_{r-1} \), the fraction of least favorable nulls. Hypotheses that fall outside of the “L”-shaped filtering region are not counted towards the multiplicity of the PC hypotheses.

To control the FWER (and PFER) at level $\alpha$, we can adaptively choose $\gamma$ as the largest satisfying $\widehat{V}_{A_{\gamma}} \leq \alpha$. Similarly, to control the FDR at level $\alpha$, we estimate the FDP as $\widehat{V}_{A_{\gamma}}/(R \lor 1)$ and select the largest $\gamma$ such that $\widehat{V}_{A_{\gamma}}/(R \lor 1) \leq \alpha$. These are essentially the Bonferroni or BH procedure with an alternative estimate of $V$. Figure 1b illustrates how the estimates $\widehat{V}_{A_{\gamma}}$ and $\widehat{V}_{A_{\gamma}}/(R \lor 1)$ change as functions of $\gamma$ under the complete null scenario (when all $H_{j0}^2$ are true).
3.1 Definition of AdaFilter procedures

Now we formally define AdaFilter for general \( n \) and \( r \). It is convenient to first introduce the notion of filtering and selection \( P \)-values. These are

\[
F_j := (n - r + 1)P_{(r-1)j}, \quad \text{and} \\
S_j := P^B_{r/n,j} = (n - r + 1)P_{(r)j},
\]

respectively. The filtering region is defined as:

\[
\mathcal{A}_j := \{(p_1, p_2, \ldots, p_n) \mid (n - r + 1)p_{(r-1)} \leq \gamma\}.
\]

**Definition 2 (AdaFilter Bonferroni).** For a level \( \alpha \), and with \( F_j \) and \( S_j \) given by (1) and (2) respectively, reject \( H_{0j}^{r/n} \) if \( S_j \leq \gamma^\text{Bon}_0 \) where

\[
\gamma^\text{Bon}_0 = \max \left\{ \gamma \in \left\{ \frac{\alpha}{M}, \ldots, \frac{\alpha}{2}, \alpha \right\} \mid \gamma \sum_{j=1}^{M} 1_{F_j \leq \gamma} \leq \alpha \right\}.
\]

**Definition 3 (AdaFilter BH).** For a level \( \alpha \), and with \( F_j \) and \( S_j \) given by (1) and (2) respectively, reject \( H_{0j}^{r/n} \) if \( S_j \leq \gamma^\text{BH}_0 \) where

\[
\gamma^\text{BH}_0 = \max \left\{ \gamma \in \mathcal{I}_{\alpha,M} \mid \gamma \sum_{j=1}^{M} 1_{F_j \leq \gamma} \leq \alpha \right\}, \quad \text{and} \\
\mathcal{I}_{\alpha,M} = \left\{ \frac{k}{m} \alpha \mid k \in \{0, 1, \ldots, M\}, m \in \{1, \ldots, M\}, k \leq m \right\}.
\]

We can also define AdaFilter adjusted \( p \)-values that provide equivalent sets of rejections as the above definitions, while being more computationally convenient and efficient.

**Definition 4 (AdaFilter adjusted \( p \)-values).** Rank the selection \( p \)-values as \( S(1) \leq S(2) \leq \cdots \leq S(M) \) and denote the original index of \( S(j) \) as \( o_j \). For each \( j \), define an AdaFilter adjustment number

\[
m_j^{AF} := \sum_{h=1}^{M} 1_{F_h \leq S(j)},
\]

Then the AdaFilter Bonferroni adjusted \( P \)-value for \( H_{0o_j}^{r/n} \) is

\[
P_{o_j}^{\text{Bon}} = \min \left\{ \min_{h \geq j} \left\{ S(h) m_h^{AF} \right\}, 1 \right\}
\]

and the AdaFilter BH adjusted \( P \)-value for \( H_{0o_j}^{r/n} \) is

\[
P_{o_j}^{\text{BH}} = \min \left\{ \min_{h \geq j} \left\{ S(h) \frac{m_h^{AF}}{h} \right\}, 1 \right\}.
\]
For any level $\alpha > 0$, AdaFilter Bonferroni rejects $H_{0j}^{r/n}$ if and only if $p_j^{Bon} \leq \alpha$. Similarly, AdaFilter BH rejects $H_{0j}^{r/n}$ if and only if $p_j^{BH} \leq \alpha$. We can verify that the AdaFilter adjusted p-values give the same set of rejections as Definition 2 and Definition 3.

4 Theoretical properties of AdaFilter

Now we prove that AdaFilter procedures control simultaneous error rates under various conditions. As stated in Section 2.1, all the following results assume that p-values across $n$ studies are independent. The key property that AdaFilter relies on is the following conditional validity lemma:

**Lemma 4.1. (Conditional validity)** When $H_{0j}^{r/n}$ is true, for any fixed $\gamma > 0$

$$\mathbb{P}(S_j \leq \gamma \mid F_j \leq \gamma) \leq \gamma$$

holds whenever $\mathbb{P}(F_j \leq \gamma) > 0$. Here $F_j$ and $S_j$ are given by (1) and (2) respectively.

Inequality (3) can be equivalently written as $\mathbb{P}(S_j \leq \gamma) \leq \gamma \mathbb{P}(F_j \leq \gamma)$, which holds even when $\mathbb{P}(F_j \leq \gamma) = 0$ as $S_j \geq F_j$ is always true. Intuitively, the “conditional validity” guarantees that for a fixed threshold $\gamma$, the estimated upper bound on the number of false rejections $V$ is $\gamma \sum_{j=1}^{M} 1_{F_j \leq \gamma}$. However, AdaFilter uses a data-dependent $\gamma$, so extra assumptions on the base p-values within one study are needed to prove simultaneous error control of AdaFilter.

4.1 Exact simultaneous error rates control for finite $M$

First, for a finite number of hypotheses $M$, we can show that AdaFilter Bonferroni controls FWER and PFER if we further assume independence of all $nM$ base p-values.

**Theorem 4.2.** Let $(P_{ij})_{n \times M}$ contain independent valid p-values. Then AdaFilter Bonferroni in Definition 2 controls FWER and PFER at level $\alpha$ for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \ldots, M\}$.

For AdaFilter BH, however, we can only prove that it controls FDR at the nominal level of $\alpha C(M)$ where $C(M) = \sum_{j=1}^{M} 1/j \approx \log M$ for a finite $M$. This is the same factor that [7] use to account for any dependence pattern in BH tests and [22] uses to account for dependence in FWER. In other words, adjusting the threshold to be $\alpha/C(M)$ can guarantee control of the FDR at level $\alpha$. 

9
Theorem 4.3. Let \((P_{ij})_{n \times M}\) contain independent valid \(p\)-values. Then AdaFilter BH in Definition 3 controls FDR at level \(\alpha C(M)\) where \(C(M) = \sum_{j=1}^{M} 1/j\) for the null hypotheses \(\{H_{0j}^{r/n} : j = 1, 2, \cdots, M\}\).

Though Theorem 4.3 is not very satisfying with an inflation factor \(C(M)\), in Section 5, we find in simulations that the AdaFilter BH procedure adjusted by \(C(M)\) still achieves higher power than other bench-marking approaches. Our simulations also suggest that the adjustment \(C(M)\) is in practice not needed. In Section 4.2, we will show that AdaFilter BH actually asymptotically controls FDR without using the inflation factor \(C(M)\) when \(M \to \infty\). The asymptotic results also do not require independence among \(p\)-values within each study.

4.2 Asymptotic FDR control when \(M \to \infty\)

Now we discuss FDR control of AdaFilter BH when the number of hypotheses \(M\) is very large, the usual case in high-throughput genetic experiments. Inspired by [13], we make the following three assumptions.

First, instead of requiring independent \(p\)-values within each study, we only assume a weak dependence structure among the \(p\)-values within each study.

Assumption 1 (Weak dependence). Within any study \(i\), the \(p\)-values \(P_{ij}\) for \(j = 1, 2, \cdots, M\) satisfy weak dependence where for any fixed \(\gamma\)

\[
\frac{1}{M^2} \sum_{j \neq j'} \left| \mathbb{P}(P_{ij} \leq \gamma, P_{ij'} \leq \gamma) - \mathbb{P}(P_{ij} \leq \gamma) \mathbb{P}(P_{ij'} \leq \gamma) \right| \to 0
\]

as \(M \to \infty\).

One scenario where the weak dependence holds is that, within each study \(i\), the number of pairs \((P_{ij}, P_{ij'})\) where \(P_{ij}\) and \(P_{ij'}\) are not independent is \(o(M^2)\). For microarrays or RNA-seq experiments, gene-gene networks are typically sparser than \(O(M^2)\). For GWAS or eQTLs, DNA loci are usually associated only when they are close enough along the DNA chain, say when \(|j - j'| < b\) for some constant \(b\). The weak dependence assumption is reasonable for both the above two scenarios.

Now let \(\mathcal{H}_{0j}^{r/n} = \{j : H_{0j}^{r/n} \text{ is true}\}\) be the set of true PC nulls and \(M_0\) be its cardinality. Similarly, define \(\mathcal{H}_{1j}^{r/n} = \{j : H_{1j}^{r/n} \text{ is true}\}\) to be the set of true PC non-nulls and let \(M_1\) be its cardinality. Besides weak dependence, we also assume that when \(M \to \infty\), the following limits exist:

Assumption 2 (Existence of limits). The following limits exist

\[
\lim_{M \to \infty} \frac{M_0}{M} = \pi_0 \in (0, 1)
\]
\[
\lim_{M \to \infty} \frac{1}{M_0} \sum_{j \in \mathcal{H}_0^{r/n}} P(F_j \leq \gamma) = \tilde{F}_0(\gamma), \quad \lim_{M \to \infty} \frac{1}{M_1} \sum_{j \in \mathcal{H}_1^{r/n}} P(F_j \leq \gamma) = \tilde{F}_1(\gamma)
\]
\[
\lim_{M \to \infty} \frac{1}{M_0} \sum_{j \in \mathcal{H}_0^{r/n}} P(S_j \leq \gamma) = \tilde{S}_0(\gamma), \quad \lim_{M \to \infty} \frac{1}{M_1} \sum_{j \in \mathcal{H}_1^{r/n}} P(S_j \leq \gamma) = \tilde{S}_1(\gamma).
\]

For a given \( n \), there are \( 2^n \) combinations of base hypotheses being null or non-null. A special case where Assumption 2 is satisfied is when each of these combinations has a limiting proportion and within each study, the base p-values have identical distributions under the null, and identical distributions under the non-null, such as a mixture driven by random underlying effect sizes. Specifically, for any \( c \in \{0, 1\}^n \) representing one of the \( 2^n \) combinations, let \( m_c \) be the number of PC hypotheses that fall into this combination. Also, let \( \mathcal{H}_0^i \) and \( \mathcal{H}_1^i \) be the sets of true nulls and true non-nulls for the \( i \)th study. If (a) \( \lim_{M \to \infty} m_c/M \) exists for all \( c \) and, (b) for all \( i \), \( \{P_{ij} : j \in \mathcal{H}_0^i\} \) have identical distributions and \( \{P_{ij} : j \in \mathcal{H}_1^i\} \) also have identical distributions, then Assumption 2 is satisfied.

Under Assumption 2, we denote

\[
\tilde{F}(\gamma) = \pi_0\tilde{F}_0(\gamma) + (1 - \pi_0)\tilde{F}_1(\gamma),
\]
\[
\tilde{S}(\gamma) = \pi_0\tilde{S}_0(\gamma) + (1 - \pi_0)\tilde{S}_1(\gamma),
\]

and further define the “asymptotic FDR” for a given \( \gamma \) as

\[
f^\infty(\gamma) = \begin{cases} 
\frac{\tilde{F}(\gamma)}{\tilde{S}(\gamma)}, & \text{if } \tilde{S}(\gamma) > 0 \\
0, & \text{otherwise},
\end{cases}
\]

and the largest \( \gamma_0^\infty \) such that \( f^\infty(\gamma) \leq \alpha \), i.e.,

\[
\gamma_0^\infty = \sup\{\gamma : f^\infty(\gamma) \leq \alpha\}.
\]

Then \( f^\infty(\gamma) \) is 0 when \( \gamma = 0 \) and exceeds 1 when \( \gamma = 1 \), thus the above set is not empty. We make a final technical assumption on the functions \( f^\infty(\cdot) \), \( \tilde{S}_0(\cdot) \) and \( \tilde{S}_1(\cdot) \) around \( \gamma_0^\infty \):

**Assumption 3 (Technical conditions).** The following two conditions hold:

(a) There exists \( \delta > 0 \) such that \( f^\infty(\gamma) \) is monotonically increasing and continuous in the interval \( (\gamma_0^\infty - \delta, \gamma_0^\infty] \), and

(b) \( \tilde{S}_0(\gamma) \) and \( \tilde{S}_1(\gamma) \) are both continuous at the point \( \gamma_0^\infty \).
Intuitively, (a) guarantees that the limit of the AdaFilter threshold $\gamma^{\text{BH}}_0$ is unique when $M \to \infty$ and (b) is satisfied if there are sufficient points (selection p-values) around $\gamma_0^\infty$ when $M$ is large. Now we are ready to state the asymptotic FDR control of AdaFilter BH.

**Theorem 4.4.** Under Assumptions 1 to 3, the AdaFilter BH procedure of Definition 3 satisfies

$$
\gamma^{\text{BH}}_0 \xrightarrow{p} \gamma^\infty_0, \quad \text{and} \quad \text{FDP} \xrightarrow{p} \frac{\pi_0 \bar{S}_0(\gamma^\infty_0)}{\bar{S}(\gamma^\infty_0)} \leq \alpha
$$

as $M \to \infty$. Thus, AdaFilter BH asymptotically controls FDR at the nominal level $\alpha$ for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \ldots, M\}$.

Notice that Assumption 3 implies that $f^\infty(\gamma^\infty_0) > 0$, guaranteeing that $\bar{S}(\gamma^\infty_0) > 0$.

**Remark 4.1.** Theorem 4.4 still holds if Assumption 2 is weakened to allow $\pi_0 = 0$ while $M_0 \to \infty$ and Assumption 1 is modified to: for any fixed $\gamma$,

$$
\frac{1}{M^2} \sum_{j \neq j' \in H_{s}^{r/n}} \left| \mathbb{P}(P_{ij} \leq \gamma, P_{ij'} \leq \gamma) - \mathbb{P}(P_{ij} \leq \gamma) \mathbb{P}(P_{ij'} \leq \gamma) \right| \xrightarrow{M \to \infty} 0
$$

for both $s = 0, 1$. We can not deal with $\pi_0 = 1$ as that would lead to $\bar{S}(\gamma^\infty_0) = 0$ and violates Assumption 3(a). In Section 5, we show with simulations that both simultaneous error rates can be controlled in practice even when $M_0/M = 0.99$.

### 4.3 Lack of complete monotonicity

The increased power of AdaFilter can lead to a striking efficiency. Suppose that we test the involvement of $M$ genes in a disease with 2 studies. One researcher uses BH or Bonferroni separately on the $M$ base $p$-values in each study and claims that a gene is important for the pathology if it is rejected in any of the two studies. Another researcher runs AdaFilter with $r = 2$ on the same data while claiming that a gene is selected only when its nulls are false in both studies. The second researcher has a stricter goal, however, it is possible that she makes more discoveries than the first.

To see how this could happen, consider the toy example in Figure 1c where $M = 2$. In both studies, neither of the two hypotheses can be rejected at significance level $\alpha = 0.05$ when using either Bonferroni or BH on each study separately. However, both AdaFilter Bonferroni and AdaFilter BH can reject $H_{02}^{2/2}$ at the same nominal level. This interesting phenomenon arises from the lack of monotonicity.
of the number of rejections in the base p-values. A multiple testing procedure has “complete monotonicity” if reducing any individual p-values can never cause any of the decisions on the null hypotheses to switch from ‘reject’ to ‘accept’.

**Definition 5** (Complete monotonicity). A multiple testing procedure has complete monotonicity if each decision function $\varphi_j$ is a non-increasing function in all the elements of $(p_{ij})_{n \times M}$ for $j = 1, 2, \ldots, M$.

Simes’, Fisher’s and Bonferroni’s meta-analyses have complete monotonicity. So does the BH procedure with $n = 1$. Heller, Bogomolov and Benjamini [19] call this property “stability” and it holds for the PC tests of [20]. However, different versions of AdaFilter do not satisfy complete monotonicity: lowering one of the p-values for gene $j$ can change the rejection of $H_{0j}^{r/n}$ to acceptance for $j' \neq j$.

Figure 1d shows how AdaFilter does not have complete monotonicity. Compared with Figure 1c, the second hypothesis has a decreased p-value in study 1 while all other p-values are kept fixed. In Figure 1c, both $\gamma_{0}^{Bon} = \gamma_{0}^{BH} = 0.05$ so the first PC hypothesis is rejected. In contrast, in Figure 1d $\gamma_{0}^{Bon} = 0.025$ and $\gamma_{0}^{BH} = 0$ so that none of the hypotheses can be rejected though it has a smaller p-value matrix.

This lack of complete monotonicity, which might appear undesirable, in fact is at the core of the efficiency of AdaFilter. A larger $P_{ij}$ can increase $F_j$ to reduce the multiplicity burden. When only a few hypotheses are non-null—as in a sparse genomics setting—we expect lots of large $P_{ij}$. This gives AdaFilter a substantial advantage in identifying the few non-null PC hypotheses. From another perspective, increased individual p-values may make the signal configuration across genes more similar among studies (such as Figure 1c compared to Figure 1d). AdaFilter can implicitly learn such similarity and utilize it to allow more rejections.

Though lacking “complete monotonicity”, AdaFilter retains a “partial monotonicity” property: reducing one of the $n$ base p-values for test $j$ can never change the decision from reject $H_{0j}^{r/n}$ to accept.

**Definition 6** (Partial monotonicity). A multiple testing procedure has partial monotonicity if for all $j \in 1:M$, its decision function $\varphi_j(p_1, \ldots, p_M)$ is non-increasing in all elements of $(p_{1j}, p_{2j}, \ldots, p_{nj})$.

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 $H_{0j}^{r/n}$. We have the following result:

**Corollary 4.5.** Both the AdaFilter Bonferroni and the AdaFilter BH procedures satisfy partial monotonicity for all null hypotheses $H_{0j}^{r/n}$, $j = 1, 2, \ldots, M$.

Corollary 4.5 indicates that AdaFilter is reasonable in a way that reducing the base p-values of the $j$th PC hypothesis indeed strengthens the evidence of replicability for the $j$th PC hypothesis, though possibly weakening the evidence of replicability for other PC hypotheses.
4.4 Extensions and relation to literature

**Remark 1: Variable r and n** In many genetic problems, the \( M \) genes or DNA loci can have varying \( r_j \) or \( n_j \) as they may not be present in every experiment. Then the \( j \)-th PC null hypothesis is \( H_{0j}^{r_j/n_j} \). AdaFilter procedures still work in this scenario because Lemma 4.1 still holds. We only need to replace formulas (1) and (2) 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.

**Remark 2: Requiring sign replicability** Partial conjunctions with two-sided test statistics can reject \( H_{0j}^{r_j/n_j} \) in settings where some of the significant findings have test statistics with positive signs and others negative. It is more natural to think of replication as having concordant signs, be they consistently positive or consistently negative. In meta-analysis, one can pool \( n \) one-sided tests for positive alternatives, repeat that for negative alternatives and double the smaller of the resulting one-sided \( p \)-values [32]. This approach is very effective when either the most likely or most useful alternatives to the null have concordant signs. We can adapt this approach to PC tests and AdaFilter as follows.

We start with two base \( P \)-value matrices, \((P_{ij}^+)_{n \times M}\) and \((P_{ij}^-)_{n \times M}\), for null hypotheses \((H_{0ij}^+)_{n \times M}\) and \((H_{0ij}^-)_{n \times M}\) respectively. The rejection of \( H_{0ij}^+ \) is for a positive sign of the signal and the rejection of \( H_{0ij}^- \) is for a negative sign. We also define two vectors of PC hypotheses \( \{H_{r/n}^{r/n,+}, \ldots, H_{0M}^{r/n,+}\} \) and \( \{H_{r/n}^{r/n,-}, \ldots, H_{0M}^{r/n,-}\} \). The PC null \( H_{0ij}^{r/n,+} \) is rejected if the signal \( j \) is positive in at least \( r \) studies, and \( H_{0ij}^{r/n,-} \) is rejected if the signal \( j \) is negative in at least \( r \) studies. If \( r > n/2 \) then it will be impossible to reject both \( H_{0ij}^{r/n,+} \) and \( H_{0ij}^{r/n,-} \) for the same \( j \).

We can apply AdaFilter twice, separately on \( \{H_{01}^{r/n,+}, \ldots, H_{0M}^{r/n,+}\} \) and \( \{H_{01}^{r/n,-}, \ldots, H_{0M}^{r/n,-}\} \), controlling the simultaneous error rate (FWER, PFER or FDR) at levels \( \alpha_1 \) and \( \alpha_2 \) respectively, with \( \alpha_1 + \alpha_2 = \alpha \) (ordinarily \( \alpha_1 = \alpha_2 = \alpha/2 \)). Let the set of rejected PC nulls be \( R^+ \) and \( R^- \), respectively. Rejecting the union of these two sets \( R^\pm = R^+ \cup R^- \) controls the corresponding error rate at a level \( \alpha = \alpha_1 + \alpha_2 \) for the null hypotheses \( \{H_{01}^{r/n,\pm}, \ldots, H_{0M}^{r/n,\pm}\} \).

If \( r \leq n/2 \), then there might be some \( j \in R^+ \cap R^- \). While such findings are not what we usually have in mind with replication they could nonetheless be scientifically interesting.

**Remark 3: Comparison with other strategies** Two directly related methods to AdaFilter are [9] for \( n = r = 2 \) and the empirical Bayes approach in [21] for controlling the Bayes FDR, both of which are designed to test for multiple PC nulls. Both methods were developed to improve the efficiency of the “direct approach” we described. AdaFilter is similar to the method of [9] but works for any \( n \) and \( r \). It provides a frequentist approach comparable to and sometimes better than [21].
The procedures of [9] use a filtering step for each study based on the p-values in the other study and a selection step that rejects hypotheses that have small enough p-values in both studies. To maximize the efficiency, the authors suggest a data-adaptive threshold. For instance, to control FWER, they chose two thresholds $\gamma_1$ and $\gamma_2$ to satisfy

$$
\gamma_1 \times \sum_{j=1}^{M} 1_{P_{2j} \leq \gamma_2} \approx \frac{\alpha}{2} \quad \text{and} \quad \gamma_2 \times \sum_{j=1}^{M} 1_{P_{1j} \leq \gamma_1} \approx \frac{\alpha}{2}.
$$

When $\gamma_1 \approx \gamma_2$, then

$$
\gamma_1 \times \sum_{j=1}^{M} 1_{\min(P_{1j}, P_{2j}) \leq \gamma_1} \leq \gamma_1 \times \sum_{j=1}^{M} (1_{P_{1j} \leq \gamma_1} + 1_{P_{2j} \leq \gamma_1}) \approx \alpha.
$$

Thus $\gamma_{\text{Bon}} \approx \gamma_1 \approx \gamma_2$ and AdaFilter becomes similar to theirs. The proposed method only applies for $n = r = 2$; this simplification makes the approach less widely applicable, despite its strong theoretical guarantees.

In [21], the authors tried to learn the proportion of each of the $2^n$ (or $3^n$ for sign replicability) configurations of base hypotheses, along with the distribution of some Z-values under each configuration. This has cost at least $O(M2^n)$ while AdaFilter has cost $O(Mn \log(n))$. There are other multiple testing procedures that aim to find consistent signals across conditions [41, 43, 45], all of which use an empirical Bayes framework as in [21]. Compared to these methods, AdaFilter is typically faster, guarantees simultaneous error rate control and does not require p-value independence within each study.

Finally, there has been much other recent literature on efficient FDR control by using some special data structure as prior knowledge [28, 29, 3, 8] and then adaptively determining the selection threshold. AdaFilter shares some similar adaptive filtering ideas, but works directly from an $n \times M$ matrix of $p$-values without assuming any special structure and is uniquely tailored to the special nature of the PC hypotheses.

**Remark 4: Testing for all possible values of $r$** The partial conjunction null $H_{0r/n}^r$ can be meaningfully defined whenever $2 \leq r \leq n$, and sometimes it is of interest to test for all possible $r$ values, adding another layer of multiplicity. See [36]. As the filtering information learnt by AdaFilter varies for different $r$ values, a signal that is rejected by a larger $r$ using AdaFilter is not guaranteed to also be rejected at a smaller replicability level. The current formulation of AdaFilter is therefore not suited to data dependent selection of the $r$ value, but requires this to be specified by the user.
5 Simulations

We benchmark the performance of AdaFilter versus “direct approaches” with the three forms of PC p-values in Section 2.2. For FDR control, we also include [21], using their R package repfdr. Within each study, we assume a block dependence structure while changing the block size to create two scenarios, weak dependence with a small block size and strong dependence with a large block size.

We set $M = 10,000$ and consider six different configurations of $n$ and $r$, as listed in Table 1a. For a given $n$, there are $2^n$ combinations of base hypotheses. In generating different configurations of the truth, we use two parameters to control the probability of each combination: $\pi_{00}$ is the probability of the global null combination and $\pi_1$ is the probability of the combinations not belonging to $H_{0j}^{r/n}$. We set $\pi_1 = 0.01$ and consider two values for $\pi_{00}$: 0.8 or 0.98, to mimic the signal sparsity in gene expression and genetic regulation studies. All PC null combinations except for the global null have equal probabilities adding up to $1 - \pi_{00} - \pi_1$. All non-null PC combinations also have equal probabilities.

We assume that p-values belonging to different studies are independent and, within one study, the correlation of the $M$ Z-values is $\Sigma_\rho \otimes I_{b \times b}$ where $\otimes$ is the Kronecker product. The covariance $\Sigma_\rho \in \mathbb{R}^{m \times m}$ has 1s on the diagonal and common value $\rho = 0.5$ off the diagonal. We set the block size $b = 100$ for weak dependence and $b = 1000$ for strong dependence, which should cover the spectrum of what is typically expected in genomics. When the base hypothesis is non-null, we sample the mean of its z-value uniformly and independently from $\mathcal{I} = \{\pm \mu_1, \pm \mu_2, \pm \mu_3, \pm \mu_4\}$ where the four levels of signals $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ correspond to detection power of $0.02, 0.2, 0.5, 0.95$ respectively.

In the analysis, we target controlling PFER at the nominal level $\alpha = 1$, FDR at the nominal level $\alpha = 0.2$, and Bayes FDR at the same level $\alpha = 0.2$ for repfdr. Bayes FDR corresponds to the posterior probability of a null hypothesis given the test statistics falling into the rejection region, which has been shown to be similar to the frequentist FDR under independence [12]. Studying PFER control, we compare four methods: AdaFilter Bonferroni and the three “direct approaches”. For FDR control, we compare 6 methods: AdaFilter BH, AdaFilter BH with the inflation factor $C(M) = \sum_{j=1}^{M} 1/j \approx \log M$, repfdr and the “direct approaches”. For each parameter configuration, we run $B = 100$ random experiments and calculate the average power, number of false discoveries and false discovery proportions of each procedure.

Table 1b shows the average PFER and recall over the six combinations of $n$ and $r$ for each setting of $b$ and $\pi_{00}$. More detailed results for each $n$ and $r$ separately are shown in Figure S1-S2. All methods that target PFER successfully control it at the nominal level, while the direct approaches are much more conservative, especially when both $n$ and $r$ are large. The gain in power is more pronounced when $\pi_{00}$ is
Table 1: Simulation results

(a) Configurations of $n$ and $r$

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

(b) Comparison of methods targeting a nominal PFER of $\alpha = 1$

|     | $\pi_{00} = 0.8$ | $\pi_{00} = 0.98$ | $\pi_{00} = 0.8$ | $\pi_{00} = 0.98$ |
|-----|------------------|-------------------|------------------|-------------------|
|     | $b = 100$       | $b = 1000$        | $b = 100$       | $b = 1000$        |
| Bon-$P_B$ | 0.04 | 14.72 | 0.05 | 14.87 | 0.00 | 14.72 | 0.00 | 14.83 |
| Bon-$P_F$ | 0.05 | 19.30 | 0.06 | 19.50 | 0.01 | 19.18 | 0.00 | 19.38 |
| Bon-$P_S$ | 0.04 | 14.80 | 0.05 | 14.93 | 0.00 | 14.78 | 0.00 | 14.88 |
| AdaFilter Bonferroni | 0.73 | 28.71 | 0.76 | 28.93 | 0.29 | 38.10 | 0.21 | 38.25 |

(c) Comparison of methods targeting a nominal FDR of $\alpha = 0.2$

|     | $\pi_{00} = 0.8$ | $\pi_{00} = 0.98$ | $\pi_{00} = 0.8$ | $\pi_{00} = 0.98$ |
|-----|------------------|-------------------|------------------|-------------------|
|     | $b = 100$       | $b = 1000$        | $b = 100$       | $b = 1000$        |
| BH-$P_B$ | 0.01 | 29.50 | 0.01 | 29.55 | 0.00 | 29.04 | 0.00 | 29.10 |
| BH-$P_F$ | 0.01 | 32.94 | 0.01 | 32.80 | 0.00 | 32.68 | 0.00 | 32.74 |
| BH-$P_S$ | 0.01 | 29.68 | 0.01 | 29.70 | 0.00 | 29.16 | 0.00 | 29.28 |
| repfdr | 0.33 | 59.39 | 0.29 | 23.53 | 0.14 | 24.31 | 0.13 | 46.17 |
| AdaFilter BH | 0.15 | 58.64 | 0.14 | 58.71 | 0.06 | 71.27 | 0.06 | 71.49 |
| Inflated AdaFilter BH | 0.02 | 34.39 | 0.01 | 34.22 | 0.01 | 45.70 | 0.01 | 46.17 |

higher, which is expected in many genetics applications.

Table 1c shows the average FDR and recall over the six combinations of $n$ and $r$ for each setting of $b$ and $\pi_{00}$. More detailed results for each $n$ and $r$ separately are shown in Figure S3-S4. AdaFilter BH, even not inflated, and the three “direct approaches” control FDR at the nominal level. However, similar to the PFER control, the “direct approaches” are too conservative. The inflated AdaFilter BH has lower power than AdaFilter BH, while its power still exceed the “direct approaches”, especially for large $r$. The repfdr method fails to consistently control FDR especially when $n$ is large: we believe that this is due to the large number of parameters that need to be estimated in these scenarios. In the cases when repfdr does control FDR, its power is comparable to AdaFilter when $\pi_{00} = 0.8$ while is less when $\pi_{00} = 0.98$ is large and further reduces when dependence increases.
6 Case studies

We apply AdaFilter to analyze three datasets: one investigates the replication of gene differential expression results in four microarray experiments of Duchenne muscular dystrophy, one focuses on identifying marker genes of one T cell subtype from lung cancer tumors using single-cell RNA-sequencing (scRNA-seq) data and the last tests for consistently significant signals across different metabolic super-pathways within one study.

6.1 Duchenne Muscular Dystrophy microarray studies

Following [26], we investigate four independent Duchenne muscular dystrophy (DMD)-related microarray datasets in the Gene Expression Omnibus (GEO) database (GDS 214, GDS 563, GDS 1956 and GDS 3027, Table 2a), to understand the signature genes for the disease. The goal here is to find differentially expressed marker genes for DMD that show replicating signals in multiple datasets. For each experiment, the data is preprocessed using a standard data reprocessing tool RMA [24] for microarrays. Within each study, we find genes that are differentially expressed between the disease and healthy group, using a popular software Limma [38] and adjust for covariates like batch and patients' age and gender when they are available.

The four datasets are from three different microarray platforms where different probe-sets are used. In order to compare across platforms, we map probe-sets to common gene names. When multiple probe-sets map to the same gene, a Bonferroni rule is applied combining p-values of these probe sets into a single p-value for the gene. The p-values for the genes within each study achieve an approximately uniform empirical null distribution (Figure S5), consistent with our model of valid base p-values and sparse non-nulls. There are only $M = 1871$ genes present in all four studies, with $M = 9848$ genes shared in at least 3 studies and $M = 13912$ genes in at least two studies. As discussed in Section 4.4, AdaFilter can work with varying $n_j$ thus allow missing entries in the p-value matrix.

The application of AdaFilter BH at level $\alpha = 0.05$ leads to the discovery of many consistently differentially expressed genes at $r = 2, 3, 4$ (Table 2b). Specifically, at $r = 4$, AdaFilter BH finds 32 significant genes (Table S1). By contrast, a BH adjustment on the Fisher combined PC p-values ($P_{r/n,j}^F$) only detects two genes (MYH3 and S100A4) and repfdr reports no significant genes as it fails to perform the distribution estimation of p-values with $M = 1871$ being too small. Table 2c shows four of the 32 genes that are known to play important roles in muscle contraction (Table S1). Notice that besides MYH3, all three markers do not have a small enough p-value in the third study (GDS1956, which is the least powerful study) to be detected when BH is applied to the study alone with a nominal FDR level 0.05. However, AdaFilter can compensate for this deficiency by leveraging the overall similarity of the results.
Table 2: Replicability analysis for DMD microarrays

(a) GEO datasets information

| GEO ID  | Platform  | Description | Source       |
|---------|-----------|-------------|--------------|
| GDS 214 | custom Affymetrix | 4 healthy, 26 DMD | Muscle       |
| GDS 563 | Affymetrix U95A | 11 healthy, 12 DMD | Quadriceps Muscle |
| GDS 1956 | Affymetrix U133A | 18 healthy, 10 DMD | Muscle       |
| GDS 3027 | Affymetrix U133A | 14 healthy, 23 DMD | Quadriceps Muscle |

(b) AdaFilter BH rejections

| r | M   | Rejected |
|---|-----|----------|
| 2 | 13912 | 494      |
| 3 | 9848  | 142      |
| 4 | 1871  | 32       |

(c) Known marker genes detected by AdaFilter at $r = 4$

| Gene Symbol | GDS 214 | GDS 563 | GDS 1956 | GDS 3027 |
|-------------|---------|---------|---------|---------|
| MYH3        | 5.47e-14 | 2.18e-09 | 3.31e-07 | 2.49e-20 |
| MYH8        | 5.74e-06 | 9.09e-11 | 2.58e-03 | 5.16e-33 |
| MYL5        | 8.97e-04 | 3.06e-06 | 1.87e-03 | 6.63e-08 |
| MYL4        | 1.48e-06 | 7.94e-08 | 1.21e-02 | 2.66e-08 |

in this study compared with other studies.

6.2 scRNA-seq of T cells in lung cancer tumors

Understanding T cell heterogeneity in tumors brings in key information to cancer immunotherapies, and the recent single-cell RNA-sequencing (scRNA-seq) technology enables measurement of gene expression levels at the single cell resolution. In [17], the authors sequenced tumor T cells from 14 treatment-naïve non-small-cell lung cancer patients and one main finding is the heterogeneity of the CD4+ regulatory T cells (Tregs). Specifically, they discovered a subtype of Tregs, named the suppressive tumor-resident Tregs (CD4-C9-CTLA4), that is different from the normal Tregs (CD4-C8-FOXP3), which is potentially associated with the immunosuppressive functions and patient prognosis. We download data from the GEO database (GSE99254), where clustering labels are already provided.

In order to characterize CD4-C9-CTLA4, one need to identify a list of “reliable” marker genes that are consistently highly expressed in CD4-C9-CTLA4 across multiple patients. Thus we apply AdaFilter treating each patient as a “study”. For each patient, we obtain p-values of each gene for whether the gene expression is higher in CD4-C9-CTLA4 than in CD4-C8-FOXP3. These one-sided base p-values are calculated using the Wilcoxon rank-sum test on the single-cell data, which is the standard test adopted in scRNA-seq. Two patients who have less than 10 Treg cells in either of the two Treg subtypes are excluded from the analysis. In summary, we obtain a p-value matrix for 23459 genes and $n = 12$ patients.

We vary the replicability level $r$ and Figure 2a compares the number of non-
null genes detected using different methods. For large $r (r \geq 8)$, AdaFilter is more powerful than the “direct approach” with Fisher’s PC p-values. However, it is less powerful when $r$ is relatively small, which is mainly because in this example, the power gain of Fisher’s combination on the $n - r + 1$ largest p-values per gene exceeds the power gain using AdaFilter, whose selection p-values are the Bonferroni’s PC p-values. All other three methods show limited power for all $r$. In Table S2, we list the 18 genes that are detected at $r = 10$, most of which are known to be linked to immunoresponse in tumors.

Figure 2: (a) scRNA-seq data: the number of genes whose $H_{0j}$ were rejected by each of the compared procedures. FDR is controlled at $\alpha = 0.05$. (b) The left is a visualization of each patient’s log fold changes between CD4-C9-CTLA4 cells and CD4-C8-FOXP3 cells for 11 genes. The darker color represents a larger log fold change. A ‘*’ label is added if the base p-value for the gene on the particular patient is smaller than 0.01. The right table shows the adjusted p-values of each gene. The first column contains the adjusted AdaFilter BH p-values for $H_{4/12}$ and the second column contains the BH adjusted merged p-values combining cells in all patients. For both columns the p-values are adjusted so that rejecting hypotheses with adjusted p-values smaller than 0.05 controls FDR at 0.05.

To further show the benefit of requiring replicability on marker gene selection, we compare a list of genes on their base p-values per patient, their standard BH adjusted merged p-values and AdaFilter BH adjusted p-values at $r = 4$ (Figure 2b). As adopted for marker gene selection in the original paper, the adjusted merged p-value per gene is computed by applying the Wilcoxon rank-sum test on cells merged...
across all patients, which are further adjusted by the standard BH procedure. All 
genes in Figure 2b would be selected in the original paper as their adjusted merged 
p-values are far less than 0.05. However, taking a closer look, the last six genes 
(ACTR3C, PRF1, MIIP, LMNA, IL27RA, CKS2) only have one patient whose base 
p-value is less than 0.01. Moreover, the base p-values do not coordinate with the 
log-fold change (the log of the mean ratio comparing cells in the two clusters) in 
each patients, indicating that there is no convincing evidence for these six genes. In 
contrast, there are multiple patients with small base p-values for other five genes. The 
merged p-values are not enough to distinguish these genes, while the more convincing 
markers show a dramatically smaller AdaFilter BH adjusted p-values.

6.3 Metabolites super-pathways GWAS data

The multi-trait GWAS data from [37] is a comprehensive study of the genetic loci 
influencing human metabolism: in addition to DNA variation, it measured the lev-
els of 333 metabolites, categorized into 8 non-overlapping “super pathways”, and 
integrated this data with gene expression and other prior information. Shin et al. 
[37] strongly emphasize how distinct metabolic traits are linked through the effects 
of specific genes and indicates that the discovery of genes that affect a diverse class 
of metabolic measurements is particularly interesting as these genes are associated 
with complex trait/disease or drug responses.

Testing for partial conjunction is a means to discover such genes. Specifically, we 
apply AdaFilter to the tests for association between single nucleotide polymorphisms 
(SNPs) and “super-pathways” (each SNP is linked to a gene, and hence discovering 
a SNP points to a specific gene; super-pathways are defined in [37]). We calculate 
the individual p-value for each SNP and each super-pathway by combining p-values 
of individual metabolic traits, giving appropriate consideration of the correlation of 
traits within each super-pathway (see details in Appendix A). A simple estimation 
of the sample correlation matrix across all traits suggests that correlation between 
different super pathways is not large and could be safely ignored (Figure S1).

Figure 3a compares the number of significant SNPs when FDR is controlled at 
0.05 and r ranges from 2 to 5. Compared with other four methods, our AdaFilter 
BH is much more powerful for any value of r. The method repfdr rejects less than 
AdaFilter BH, which is a consistent result with the simulations that repfdr may 
suffer from a power deficiency under dependence structures.

Among the significant SNPs at r = 3, 14 different SNPs are detected after clump-
ing using PLINK 1.9 ([34], Table S2), representing 13 different genes (Figure 3b).
Many of these genes have important roles in complex disease. For instance, gene 
GCKR encodes a regulatory protein that inhibits glucokinase, which regulates car-
bohydrate metabolism, converting glucose to amino acid and fatty acids. It is also 
a potential drug target for diabetes. Several genes (SLC17A3, SLC2A9, SLC22A4,
Figure 3: (a) Metabolics GWAS data: the number of SNPs whose $H_{r/n}^{n}$ were rejected by each of the compared procedures. FDR is controlled at $\alpha = 0.05$. (b) Visualization of individual p-values of the super-pathways for the 13 clumped significant SNPs (their mapped genes are labeled) detected at $r = 3$ (Web Table 2). For the two SNPs that map to the same gene, only the more significant one is shown. The significant p-values have a blue color. The darker the color, the smaller the p-value is.

$SLCO1B1$ encoding the solute carrier (SLC) group of membrane transport proteins are also detected. This suggests that they might function to transport multiple solutes and could possibly be drug targets for diabetes, chronic kidney disease and various autoimmune diseases.

7 Conclusion

Testing PC hypotheses provides a framework to detect consistently significant signals across multiple studies, leading to an explicit assessment of the replicability of scientific findings. We introduced AdaFilter, a multiple testing procedure which greatly increases the power in simultaneous testing of PC hypotheses over other existing methods. AdaFilter implicitly learns and utilizes the overall similarity of results across studies and exhibits a lack of complete-monotonicity. We proved that
AdaFilter procedures control FWER and FDR under independence of all $p$-values for a given finite number of hypotheses, and further showed that AdaFilter BH asymptotically controls FDR allowing weak dependence within each study. The validity of AdaFilter does require independence of the results between studies.

We applied AdaFilter to three case studies, encompassing gene expression and genetic association. Other types of applications include eQTL studies and multi-ethnic GWAS (such as new Population Architecture using Genomics and Epidemiology (PAGE) study) where it is of great interest to understand which genetic regulations are shared and which are tissue / population specific. Actually, PC tests can be quite useful in even broader context. According to Hume [23], “constant conjunction” is a characteristic of causal effects. If some hypotheses are rejected repeatedly under various distinct settings, that can be supportive evidence for some causal mechanism instead simple associations. These directions can be further investigated in future research.

Acknowledgement

A. Owen by the US NSF grants IIS-1837931 and DMS-1521145; and C. Sabatti by NIH grants R01MH101782 and U01HG007419. We thank for helpful discussions Y. Benjamini, M. Bogomolov, R. Heller and N. R. Zhang.

References

[1] Introduction to Replicability in science. http://www.replicability.tau.ac.il/index.php/replicability-in-science.html. Accessed: 2018-08-29.

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

[3] R. F. Barber and A. Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. Journal of the Royal Statistical Society: Series B, 79(4):1247–1268, 2017.

[4] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. Nature, 483(7391):531–533, Mar 2012.

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

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B, pages 289–300, 1995.
[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[8] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

[9] M. Bogomolov and R. Heller. Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3):505–516, 2018.

[10] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

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

[12] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[13] J. Ferreira and A. Zwinderman. On the Benjamini–Hochberg method. *The Annals of Statistics*, 34(4):1827–1849, 2006.

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

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

[16] A. Giri, J. N. Hellwege, J. M. Keaton, J. Park, C. Qiu, H. R. Warren, E. S. Torstenson, C. P. Kovesdy, Y. V. Sun, O. D. Wilson, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nature genetics*, 51(1):51–62, 2019.

[17] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature medicine*, 24(7):978–985, 2018.

[18] Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

[19] R. Heller, M. Bogomolov, and Y. Benjamini. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111(46):16262–16267, 2014.
[20] R. Heller, Y. Golland, R. Malach, and Y. Benjamini. Conjunction group analysis: an alternative to mixed/random effect analysis. *Neuroimage*, 37(4):1178–1185, 2007.

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

[22] G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[23] D. Hume. *A treatise of human nature*. Courier Corporation, 2003.

[24] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[25] B. Karmakar, D. S. Small, and P. R. Rosenbaum. Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study of the effectiveness of catholic schools. *Journal of the American Statistical Association*, (just-accepted):1–33, 2020.

[26] E. Kotelnikova, M. A. Shkrob, M. A. Pyatnitskiy, A. Ferlini, and N. Daraselia. Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in duchenne muscular dystrophy. *PLoS Comput Biol*, 8(2):e1002365, 2012.

[27] J. Lehrer. The truth wears off. *The New Yorker*, December 13, 2010.

[28] L. Lei and W. Fithian. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B*, 2016.

[29] A. Li and R. F. Barber. Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *arXiv preprint arXiv:1606.07926*, 2016.

[30] U. M. Marigorta and A. Navarro. High trans-ethnic replicability of gwas results implies common causal variants. *PLoS genetics*, 9(6), 2013.

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

[32] A. B. Owen. Karl Pearson’s meta-analysis revisited. *Annals of Statistics*, 37(6):3867–3892, 2009.
[33] C. J. Price and K. J. Friston. Cognitive conjunction: A new approach to brain activation experiments. *NeuroImage*, 5(4):261–270, 1997.

[34] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[35] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[36] L. Shenhav, R. Heller, and Y. Benjamini. Quantifying replicability in systematic reviews: the r-value. Technical report, arXiv:1502.00088, 2015.

[37] S. Shin, E. Fauman, A. Petersen, J. Krumsiel, R. Santos, J. Huang, et al. An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550, 2014.

[38] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3, 2004.

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

[40] J. W. Tukey. The problem of multiple comparisons. Technical report, Princeton University, 1953.

[41] S. M. Urbut, G. Wang, P. Carbonetto, and M. Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, 2019.

[42] J. Wang and A. B. Owen. Admissibility in partial conjunction testing. *Journal of the American Statistical Association*, pages 1–11, 2018.

[43] D. Xiang, S. D. Zhao, and T. T. Cai. Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):707–734, 2019.

[44] N. R. Zhang, Y. Senbabaoglu, and J. Z. Li. Joint estimation of dna copy number from multiple platforms. *Bioinformatics*, 26(2):153–160, 2010.

[45] S. D. Zhao, Y. T. Nguyen, et al. Nonparametric false discovery rate control for identifying simultaneous signals. *Electronic Journal of Statistics*, 14(1):110–142, 2020.
S1 Appendix A: More details for the analysis of metabolites super-pathways GWAS data

In [37], a total of 7824 adult individuals from 2 European populations were recruited in the study, and $M = 2,182,555$ SNPs were recorded, either directly genotyped or imputed from the HapMap 2 panel. Out of the 333 annotated metabolite traits reported in the paper, only 275 have the summary statistics (t statistics and p-values for the association of each SNP and trait) publicly available at the Metabolomics GWAS Server http://mips.helmholtz-muenchen.de/proj/GWAS/gwas/index.php?task=download, which is the data we use for analysis.

To calculate individual p-values $p_{ij}$ for each marker $j$ and each super-pathway $i$, we start with the Z-values $Z_{sj}$ for test of association between each metabolite $s$ and marker $j$, which are given as summary statistics. For a super-pathway $i$, let $\{s_1, s_2, \ldots, s_n\}$ be the index set of metabolite measures that belong to it. We assume that $(Z_{s_1j}, Z_{s_2j}, \ldots, Z_{s_nj}) \sim \mathcal{N}(0, \Sigma_i)$. The covariance $\Sigma_i$ can be accurately estimated in principle since we have millions of markers. Most of the individual hypotheses are null and the noise of the estimates of the marginal effects of these SNPs should share a common correlation matrix [10]. We estimate $\Sigma_i$ using graphical Lasso, assuming that the precision matrix is sparse. To do this, we randomly sample 2000 SNPs (markers) that lie at least 1Mbp away from each other, so that they can be considered as independent SNPs. Then these SNPs are treated as samples in Graphical Lasso and the tuning parameters of the final estimates selected by cross-validation. The Graphical Lasso approach guarantees an accurate sparse inverse covariance matrix estimation, that is needed for computing the p-values for each super-pathway.

Let $p_{ij}$ be the p-value for the association between a super-pathway $i$ and marker $j$, in other words, the p-value for the null $H_{ij0}$: no metabolite measure of super pathway $i$ is associated with marker $j$.

Given $(Z_{s_1j}, Z_{s_2j}, \ldots, Z_{s_nj})$ and $\hat{\Sigma}_i$, we calculate p-values $p_{ij}$ from the chi-square test treating the estimated $\Sigma_i$ as known. These $p_{ij}$ serve as individual p-values which will be used in the partial conjunction testing. Web Figure 6 shows the estimated correlation across metabolites assuming $(Z_{1j}, Z_{2j}, \ldots, Z_{mj}) \sim \mathcal{N}(0, \Sigma)$ for $j = 1, 2, \ldots, M$. We estimate $\Sigma$ by applying the Minimum Covariance Determinant (MCD) [35] estimator to the 2000 randomly sampled SNPs, where MCD is a highly robust method to reduce the influence of the sparse non-null hypotheses. Notice that we choose MCD instead of graphical Lasso here as we do not need an estimate of the inverse of $\Sigma$. It is evident that most of the nonzero correlations are between traits within the same super-pathways: this allows us to apply the adaptive filtering procedures for PC hypotheses across super-pathways with confidence.
S2 Appendix B: proofs

Here we provide proofs for all the theoretical results in Section 3 and Section 4. Additional to the notations in the main text, we define \( p_{ij} = (p_{1j}, \ldots, p_{nj}) \) be the vector of p-values for individual hypotheses \( (H_{01j}, \ldots, H_{0nj}) \) involved in \( H_{0ij}^{r/n} \) for each \( j = 1, 2, \ldots, M \). Also, we use \( 1:M \) as a concise notation of the index set \( \{1, 2, \ldots, M\} \). For a set \( u \subset 1:n \), define \( P_{ui} = \{P_{ki} : k \in u\} \) and \( P_{ui} \preceq \lambda \) to be the event that all elements \( P_{ki} \in P_{ui} \) satisfy \( P_{ki} \leq \lambda \) for some scalar \( \lambda \).

S2.1 Proof of Lemma 4.1 (Conditional validity)

We use \( u \subseteq \{1, 2, \ldots, n\} \) to represent a subset of the studies. This set \( u \) has cardinality \( |u| \). We use \( -u \) to denote its complement \( \{1, 2, \ldots, n\}\backslash u \).

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

\[
P(P_{ij} \leq \tilde{\gamma}) = \sum_{k=r}^{n} \sum_{|u|=k} \prod_{i \in u} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}), \quad \text{and}
\]

\[
P(P_{(r-1)j} \leq \tilde{\gamma}) = \sum_{k=r-1}^{n} \sum_{|u|=k} \prod_{i \in u} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}).
\]

Next, because \( H_{0ij}^{r/n} \) 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 \( H_{0u^*j} \) is true. Because all the \( P_{ij} \) are valid,

\[
P(S_j \leq \gamma) = P(P_{(r)j} \leq \tilde{\gamma})
\]

\[
= \sum_{k=r}^{n} \sum_{|u|=k} \left( \prod_{i \in u} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}) \right)
\]

\[
\leq \tilde{\gamma} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \left( \prod_{i \in \{1, \ldots, n\}\backslash\{i^*\}} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}) \right)
\]

\[
= \tilde{\gamma} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \left( \prod_{i \in \{1, \ldots, n\}\backslash\{i^*\}} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}) \right)
\]

\[
+ \tilde{\gamma} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \left( \prod_{i \in \{1, \ldots, n\}\backslash\{i^*\}} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u \cup\{i^*\}} P(P_{ij} > \tilde{\gamma}) \right)
\]

\[
\leq \tilde{\gamma} \cdot \sum_{k=r}^{n} \sum_{|u|=k} \left( \prod_{i \in u} P(P_{ij} \leq \tilde{\gamma}) \prod_{i \notin u} P(P_{ij} > \tilde{\gamma}) \right)
\]

28
\[ + \tilde{\gamma} \cdot \sum_{k=r-1}^{n-1} (n-k) \sum_{|u|=k} \left( \prod_{i \in u} \mathbb{P}(P_{ij} \leq \tilde{\gamma}) \prod_{i \in -u} \mathbb{P}(P_{ij} > \tilde{\gamma}) \right) \]
\[ \leq (n-r+1)\tilde{\gamma} \cdot \mathbb{P}(P_{(r-1)j} \leq \tilde{\gamma}). \]

Thus
\[ \mathbb{P}(S_j \leq \gamma \mid F_j \leq \gamma) = \frac{\mathbb{P}(P_{(r)j} \leq \tilde{\gamma})}{\mathbb{P}(P_{(r-1)j} \leq \tilde{\gamma})} \leq \gamma. \]

### S2.2 Proof of Theorem 4.2

For \( j = 1, 2, \ldots, M \), define
\[ \gamma_j = \max \left\{ \gamma \in \{ \frac{\alpha_1}{2}, \ldots, \frac{\alpha_M}{2} \} : \gamma \cdot \left( 1 + \sum_{s \neq j} 1_{F_s \leq \gamma} \right) \leq \alpha \right\}. \]

It is obvious that if \( F_j \leq \gamma_0 \Bon \), then \( \gamma_j = \gamma_0 \Bon \), otherwise \( \gamma_j \leq \gamma_0 \Bon \). Then the PFER is
\[ \mathbb{E}(V) = \mathbb{E}\left( \sum_{j=1}^{M} 1_{S_j \leq \gamma_0 \Bon} \cdot 1_{v_j=0} \right) = \mathbb{E}\left( \sum_{j=1}^{M} 1_{S_j \leq \gamma_0 \Bon} 1_{F_j \leq \gamma_0 \Bon} \cdot 1_{v_j=0} \right) \]
\[ = \sum_{j=1}^{M} \mathbb{E}\left( 1_{S_j \leq \gamma_j} \cdot 1_{F_j \leq \gamma_0 \Bon} \cdot 1_{v_j=0} \right) \]
\[ = \sum_{j=1}^{M} \mathbb{E}\left( 1_{S_j \leq \gamma_j} \cdot 1_{F_j \leq \gamma_j} \right) \cdot 1_{v_j=0} \]

Recall that \( \gamma_j \) does not depend on \( P_j \) while \((F_j, S_j)\) only depends on \( P_j \). Therefore \( \gamma_j \) is independent of \((F_j, S_j)\) by our assumption on \((P_{ij})\). Now using the conditional validity in Lemma 4.1 of the main article,
\[ \mathbb{E}(V) = \mathbb{E}\left( \sum_{j=1}^{M} \mathbb{E}\left[ 1_{S_j \leq \gamma_j} \mid \gamma_j, 1_{F_j \leq \gamma_j} \right] \cdot 1_{F_j \leq \gamma_j} \cdot 1_{v_j=0} \right) \]
\[ \leq \mathbb{E}\left( \sum_{j=1}^{M} \gamma_j \cdot 1_{F_j \leq \gamma_j} \cdot 1_{v_j=0} \right) \]
\[ \leq \mathbb{E}\left( \gamma_0 \Bon \sum_{j=1}^{M} 1_{F_j \leq \gamma_0 \Bon} \cdot 1_{v_j=0} \right) \leq \alpha. \]
S2.3 Proof of Theorem 4.3

For each \( j = 1, 2, \ldots, M \) define

\[
\gamma_j = \max \left\{ \gamma \in \mathcal{I}_{a,M} : \gamma \cdot (1 + \sum_{k \neq j}^M 1_{F_k \leq \gamma}) \leq \alpha \cdot (1 + \sum_{k \neq j}^M 1_{S_k \leq \gamma}) \right\}. \tag{4}
\]

It’s obvious that if \( S_j \leq \gamma_0^{BH} \) then also \( F_j \leq S_j \leq \gamma_0^{BH} \), thus \( \gamma_j = \gamma_0^{BH} \). If \( F_j \leq \gamma_0^{BH} < S_j \), then \( \gamma_j \geq \gamma_0^{BH} \). Finally as \( \gamma_0^{BH} \leq \alpha \), if \( F_j > \gamma_0^{BH} \), then also \( \gamma_j \geq \gamma_0^{BH} \) as \( \gamma \leq \alpha \) for any \( \gamma \in \mathcal{I}_{a,M} \). In summary, \( \gamma_j \geq \gamma_0^{BH} \) and when \( S_j \leq \gamma_0^{BH} \) the equality holds.

Writing the FDR and using (4),

\[
\mathbb{E} \left( \frac{V}{R \vee 1} \right) = \mathbb{E} \left( \frac{\sum_{j=1}^M 1_{S_j \leq \gamma_0^{BH}} \cdot 1_{V_j = 0}}{\sum_{j=1}^M 1_{S_j \leq \gamma_0^{BH}} \vee 1} \right) \\
\leq \alpha \cdot \mathbb{E} \left( \frac{\sum_{j=1}^M 1_{S_j \leq \gamma_0^{BH}} 1_{V_j = 0}}{[\gamma_0^{BH} \sum_{j=1}^M 1_{F_j \leq \gamma_0^{BH}}] \vee \alpha} \right) \\
\leq \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{1_{S_j \leq \gamma_0^{BH}}}{[\gamma_0^{BH} \sum_{k=1}^M 1_{F_k \leq \gamma_0^{BH}}] \vee \alpha} \right) \\
\leq \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{1_{S_j \leq \gamma_j}}{[\gamma_j (1 + \sum_{k \neq j}^M 1_{F_k \leq \gamma_j})] \vee \alpha} \right),
\]

because \( \gamma_j \geq \gamma_0^{BH} \) and \( 1_{V_j = 0} \leq 1 \).

Now let \( P_{(-j)} \) contain all \( P_k \) for \( k \neq j \). This \( P_{(-j)} \) determines \( \gamma_j \) and all \( F_k \) for \( k \neq j \) while being independent of \( P_j \) which determines \( F_j \) and \( S_j \). Using this independence and the conditional validity Lemma 4.1

\[
\mathbb{E} \left( \frac{V}{R \vee 1} \right) \leq \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{1_{S_j \leq \gamma_j}}{[\gamma_j (1 + \sum_{k \neq j}^M 1_{F_k \leq \gamma_j})] \vee \alpha} \mid P_{(-j)} \right) \\
\leq \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{\gamma_j 1_{F_j \leq \gamma_j}}{[\gamma_j (1 + \sum_{k \neq j}^M 1_{F_k \leq \gamma_j})] \vee \alpha} \mid P_{(-j)} \right) \\
= \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{\gamma_j 1_{F_j \leq \gamma_j}}{[\gamma_j (1 + \sum_{k \neq j}^M 1_{F_k \leq \gamma_j})] \vee \alpha} \right).
\]

Next, because \( 1_{F_j \leq \gamma_j} \leq 1 \),

\[
\mathbb{E} \left( \frac{V}{R \vee 1} \right) \leq \alpha \cdot \sum_{j=1}^M \mathbb{E} \left( \frac{\gamma_j 1_{F_j \leq \gamma_j}}{[\gamma_j \sum_{k=1}^M 1_{F_k \leq \gamma_j}] \vee \alpha} \right) \leq \alpha \sum_{j=1}^M \mathbb{E} \left( \frac{1_{F_j \leq \gamma_j}}{\sum_{k=1}^M 1_{F_k \leq \gamma_j} \vee 1} \right).
\]
To complete the proof, note that whether $F_j \leq \gamma_j$ or not,

$$\frac{1_{F_j \leq \gamma_j}}{\sum_{k=1}^{M} 1_{F_k \leq \gamma_j} \lor 1} \leq \frac{1_{F_j \leq \gamma_j}}{\sum_{k=1}^{M} 1_{F_k \leq F_j}} \leq \frac{1}{\sum_{k=1}^{M} 1_{F_k \leq F_j}} \leq \frac{1}{j}.$$

### S2.4 Proof of Theorem 4.4

We separate the proof into three parts. The first part shows convergence of some empirical cumulative distribution functions (ECDFs). Then the next two parts establish the two claims in the theorem.

The first part of that proof requires weak dependence of the filtered $p$-values $F_i$.

Our next lemma extends weak dependence from the base $p$-values $P_{ij}$ to the $F_i$.

**Lemma S2.1.** Assumption 1 guarantees that for any fixed $\gamma$,

$$\frac{1}{M^2} \sum_{j \neq j'} \left| \mathbb{P}(F_j \leq \gamma, F_{j'} \leq \gamma) - \mathbb{P}(F_j \leq \gamma) \mathbb{P}(F_{j'} \leq \gamma) \right| \xrightarrow{M \to \infty} 0$$

**Proof.** Notice that $F_j = (n - r + 1)P_{(r-1)j}$, we only need to show

$$A = \frac{1}{M^2} \sum_{j \neq j'} \left[ \mathbb{P}(P_{(r-1)j} \leq \gamma, P_{(r-1)j'} \leq \gamma) - \mathbb{P}(P_{(r-1)j} \leq \gamma) \mathbb{P}(P_{(r-1)j'} \leq \gamma) \right] \to 0$$

With the following decomposition:

$$\mathbb{P}(P_{(r-1)j} \leq \gamma, P_{(r-1)j'} \leq \gamma)$$

$$= \sum_{k=r-1}^{n} \sum_{k'=r-1}^{n} \sum_{u \subset 1:n} \sum_{\tilde{u} \subset 1:n} \mathbb{P}(P_u \leq \gamma, P_{\tilde{u}^c} > \gamma, P_{\tilde{u}j} \leq \gamma, P_{\tilde{u}^c j'} \geq \gamma)$$

and

$$\mathbb{P}(P_{(r-1)j} \leq \gamma) \mathbb{P}(P_{(r-1)j'} \leq \gamma)$$

$$= \sum_{k=r-1}^{n} \sum_{k'=r-1}^{n} \sum_{u \subset 1:n} \sum_{\tilde{u} \subset 1:n} \mathbb{P}(P_u \leq \gamma, P_{\tilde{u}^c} > \gamma) \mathbb{P}(P_{\tilde{u}j} \leq \gamma, P_{\tilde{u}^c j'} \geq \gamma),$$

we further only need to show that for any sets $u, \tilde{u} \subset 1:n$ and any $i \neq j \in 1:M,$

$$\Delta_{u,\tilde{u},j,j'} = \left| \mathbb{P}(P_u \leq \gamma, P_{\tilde{u}^c} > \gamma) \mathbb{P}(P_{\tilde{u}j} \leq \gamma, P_{\tilde{u}^c j'} \geq \gamma) - \mathbb{P}(P_u \leq \gamma, P_{\tilde{u}^c} > \gamma) \mathbb{P}(P_{\tilde{u}j} \leq \gamma, P_{\tilde{u}^c j'} \geq \gamma) \right|$$
converges to 0 when \( M \to \infty \) and \( n \) is fixed. Since we assume that base p-values across studies are independent, we have

\[
\begin{align*}
\mathbb{P}(P_{ij} \leq \gamma, P_{i'j} > \gamma, P_{\tilde{a}j} \leq \gamma, P_{\tilde{a}'j} > \gamma) \\
= \prod_{t \in u} \mathbb{P}(P_{ij} \leq \gamma) \prod_{t' \in u} \mathbb{P}(P_{i'j} \leq \gamma) \prod_{o \in u} \mathbb{P}(P_{oj} > \gamma) \prod_{\tilde{a} \in u} \mathbb{P}(P_{\tilde{a}j} > \gamma) \\
\times \prod_{t' \in u} \mathbb{P}(P_{t'j} \leq \gamma, P_{t'j'} \leq \gamma) \prod_{o' \in u} \mathbb{P}(P_{o'j} > \gamma, P_{o'j'} > \gamma)
\end{align*}
\]

and for \( s = i \) or \( j \) and any \( u \),

\[
\mathbb{P}(P_{us} \leq \gamma, P_{us} \geq \gamma) = \prod_{t \in u} \mathbb{P}(P_{ts} \leq \gamma) \prod_{h \in u} \mathbb{P}(P_{hs} > \gamma).
\]

Thus, after merge the probability terms that are shared and bound them by 1, we have

\[
\Delta_{u,\tilde{a},jj'} \leq \left| \prod_{t' \in u} \mathbb{P}(P_{t'j} \leq \gamma, P_{t'j'} \leq \gamma) \prod_{o' \in u} \mathbb{P}(P_{o'j} > \gamma, P_{o'j'} > \gamma) \right| \\
- \left| \prod_{t' \in u} \mathbb{P}(P_{t'j} \leq \gamma) \mathbb{P}(P_{t'j'} \leq \gamma) \prod_{o' \in u} \mathbb{P}(P_{o'j} > \gamma) \mathbb{P}(P_{o'j'} > \gamma) \right|
\]

Using the fact that for any \( \{a_k, b_k, k = 1:n\} \) where \( a_k \in [0,1] \) and \( b_k \in [0,1] \), we have

\[
\left| \prod_{k} a_k - \prod_{k} b_k \right| \\
= \left| \prod_{k} a_k - \sum_{l=1}^{n-1} \left( \prod_{k \leq t \wedge k' > t} a_k \prod_{k \leq t \wedge k' > t} b_k' \right) \prod_{k \geq t \wedge k' > t} b_k \right| \\
\leq \sum_{k} |a_k - b_k|.
\]

As a consequence, we can derive that under Assumption 1

\[
\Delta_{u,\tilde{a},jj'} \leq \sum_{t' \in u} \left| \mathbb{P}(P_{t'j} \leq \gamma, P_{t'j'} \leq \gamma) - \mathbb{P}(P_{t'j} \leq \gamma) \mathbb{P}(P_{t'j'} \leq \gamma) \right| \\
+ \sum_{o' \in u} \left| \mathbb{P}(P_{o'j} > \gamma, P_{o'j'} > \gamma) - \mathbb{P}(P_{o'j} > \gamma) \mathbb{P}(P_{o'j'} > \gamma) \right| \\
\leq \sum_{l=1}^{n} \left| \mathbb{P}(P_{lj} \leq \gamma, P_{lj'} \leq \gamma) - \mathbb{P}(P_{lj} \leq \gamma) \mathbb{P}(P_{lj'} \leq \gamma) \right| \to 0
\]

When \( M \to 0 \). Thus, \( A \to \infty \) and the Lemma is proved. \( \Box \)
Now we are ready to prove the three parts.

**PART I: ECDF convergence.**

Define four ECDFs

\[
\begin{align*}
F_{0,M}(\gamma) &:= \frac{1}{M_0} \sum_{j \in H_0^r/n} 1_{F_j \leq \gamma}, & F_{1,M}(\gamma) &:= \frac{1}{M_1} \sum_{j \in H_1^r/n} 1_{F_j \leq \gamma}, \\
S_{0,M}(\gamma) &:= \frac{1}{M_0} \sum_{j \in H_0^r/n} 1_{S_j \leq \gamma}, & S_{1,M}(\gamma) &:= \frac{1}{M_1} \sum_{j \in H_1^r/n} 1_{S_j \leq \gamma}.
\end{align*}
\]

We will show that \( F_{0,M}(\gamma) \xrightarrow{p} \tilde{F}_0(\gamma) \) uniformly in \( 0 \leq \gamma \leq 1 \) under Assumptions 1 and 2. The same argument establishes uniform convergence of \( F_{1,M} \xrightarrow{p} \tilde{F}_{1,M}, \)
\( S_{0,M} \xrightarrow{p} \tilde{S}_{0,M} \) and \( S_{1,M} \xrightarrow{p} \tilde{S}_{1,M}. \)

First we write

\[
|F_{0,M}(\gamma) - \tilde{F}_{0,M}(\gamma)| \leq |F_{0,M}(\gamma) - \mathbb{E}(F_{0,M}(\gamma))| + \mathbb{E}(F_{0,M}(\gamma)) - \tilde{F}_{0,M}(\gamma)|. \tag{5}
\]

The second term in (5) vanishes pointwise in \( \gamma \) by Assumption 2. Next by Lemma S2.1,

\[
\text{var}(F_{0,M}(\gamma)) = \frac{\mathbb{P}(F_1 \leq \gamma)\mathbb{P}(F_1 > \gamma)}{M_0} + o(1) \to 0
\]

and so by Chebychev’s inequality, the first term in (5) also vanishes pointwise in \( \gamma \). This proves pointwise convergence of \( F_{0,M} \) to \( \tilde{F}_0 \). Then uniform convergence follows the same way it does in the Glivenko-Cantelli theorem.

**PART II: Proof that \( \gamma_0^{BH} \to \gamma_0^{\infty}, \) a constant.**

Define

\[
\begin{align*}
F_M(\gamma) &:= \frac{1}{M} \sum_{i=1}^M 1_{(F_i \leq \gamma)}, & S_M(\gamma) &:= \frac{1}{M} \sum_{i=1}^M 1_{(S_i \leq \gamma)}, \\
f_M(\gamma) &:= \frac{\gamma F_M(\gamma)}{S_M(\gamma) \vee \frac{1}{M}}
\end{align*}
\]

A direct conclusion from Part I is that

\[
\begin{align*}
F_M(\gamma) \xrightarrow{p} \tilde{F}(\gamma), & \quad S_M(\gamma) \xrightarrow{p} \tilde{S}(\gamma) \\
f_M(\gamma) \xrightarrow{p} f^\infty(\gamma)
\end{align*}
\]

hold uniformly in \( \gamma \in [0, \alpha] \).

33
Notice that for $\forall x \in [0, \alpha]$

$$
\inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f_M(\gamma) \geq \inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} [f_M(\gamma) - f^\infty(\gamma)] + \inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f^\infty(\gamma)
$$

$$
\geq - \sup_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} |f_M(\gamma) - f^\infty(\gamma)| + \inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f^\infty(\gamma)
$$

$$
= o_p(1) + \inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f^\infty(\gamma)
\xrightarrow{p} \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma)
$$

where the last limit holds as $f^\infty(\gamma)$ is right continuous (this is because both $\tilde{F}(\gamma)$ and $\tilde{S}(\gamma)$ are right continuous) and $\mathcal{I}_{a,M}$ contains all rational numbers between $[0, \alpha]$ in the limit. Similarly,

$$
\inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f_M(\gamma) \leq \sup_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} |f_M(\gamma) - f^\infty(\gamma)| + \inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f^\infty(\gamma)
\xrightarrow{p} \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma).
$$

Also,

$$
E[(\gamma^\text{BH}_0)^k] = \int_0^\alpha k x^{k-1} P(\gamma^\text{BH}_0 \geq x) dx
$$

$$
= \int_0^\alpha k x^{k-1} P(\inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f_M(\gamma) \leq \alpha) dx.
$$

Under Assumption 3(a), we have

$$
1_{[0, \alpha]}(\inf_{\gamma \in [x, \alpha]} f^\infty(\gamma)) \leq \liminf_{m \to \infty} P(\inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f_M(\gamma) \leq \alpha)
$$

$$
\leq \limsup_{m \to \infty} P(\inf_{\gamma \in [x, \alpha] \cap \mathcal{I}_{a,M}} f_M(\gamma) \leq \alpha)
$$

$$
\leq 1_{[0, \alpha]}(\inf_{\gamma \in [x, \alpha]} f^\infty(\gamma))
$$

for almost all $x$. Thus, by Fatou’s lemma

$$
\int_0^\alpha k x^{k-1} 1_{[0, \alpha]}(\inf_{\gamma \in [x, \alpha]} f^\infty(\gamma)) dx \leq \liminf_{m \to \infty} E[(\gamma^\text{BH}_0)^k]
\leq \limsup_{m \to \infty} E[(\gamma^\text{BH}_0)^k]
$$

$$
\leq \int_0^\alpha k x^{k-1} 1_{[0, \alpha]}(\inf_{\gamma \in [x, \alpha]} f^\infty(\gamma)) dx.
$$
In addition, we have $\tilde{F}(\gamma) \geq \tilde{S}(\gamma)$ as $F_j \leq S_j$ for any hypothesis $j$, thus

$$\gamma_0^\infty \leq \alpha.$$  

Thus Assumption 3(a) also guarantees that

$$\{x : \gamma_0^\infty \geq x \text{ and } x \leq \alpha\} = \left\{x : \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma) \leq \alpha\right\} = \left\{x : \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma) < \alpha\right\} \cup \gamma_0^\infty$$

Hence,

$$\int_0^\alpha k x^{k-1} 1_{[0, \alpha]}( \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma) ) dx = \int_0^\alpha k x^{k-1} 1_{[0, \alpha]}( \inf_{\gamma \in [x, \alpha]} f^\infty(\gamma) ) dx = \int_0^\gamma \infty k x^{k-1} dx = (\gamma_0^\infty)^k$$

Then

$$E|\gamma_0^{BH} - \gamma_0^\infty|^2 = E[(\gamma_0^{BH})^2] - 2\gamma_0^\infty \cdot E[\gamma_0^{BH}] + (\gamma_0^\infty)^2 \to 0$$

$$\Rightarrow P(|\gamma_0^{BH} - \gamma_0^\infty| \geq \epsilon) \leq \frac{E|\gamma_0^{BH} - \gamma_0^\infty|^2}{\epsilon^2} \to 0$$

$$\Rightarrow \gamma_0^{BH} \xrightarrow{p} \gamma_0^\infty$$

Where $\gamma_0^\infty$ is a constant.

**PART III: Convergence of FDP.**

Finally, we prove that if $\tilde{S}(\gamma_0^\infty) > 0$, then

$$\text{FDP} \xrightarrow{p} \frac{\pi_0 \tilde{S}(\gamma_0^\infty)}{\tilde{S}(\gamma_0^\infty)} \leq \alpha$$

Where FDP is the false discovery proportion of the AdaFilter BH procedure.

Since we have already shown in Part II that

$$\gamma_0^{BH} \xrightarrow{p} \gamma_0^\infty,$$

Then $\forall \epsilon \in (0, \gamma_0^\infty)$ and $\forall \eta \in (0, 1),$

$$\frac{M_0S_{0,M}(\gamma_0^\infty - \epsilon)/M}{S_M(\gamma_0^\infty + \epsilon) \lor 1} \leq \frac{M_0S_{0,M}(\gamma_0^{BH})}{M S_M(\gamma_0^{BH}) \lor 1} \leq \frac{M_0S_{0,M}(\gamma_0^\infty + \epsilon)/M}{S_M(\gamma_0^\infty - \epsilon) \lor 1}$$
hold with probability at least $1 - \eta$ when $M$ is sufficiently large.

Then as $M_0/M \to \pi_0$, $S_{0,M}(\gamma) \overset{p}{\to} \tilde{S}_0(\gamma)$, $S_M(\gamma) \overset{p}{\to} \tilde{S}(\gamma)$ uniformly in $\gamma \in [0,1]$, and $\tilde{S}_0, \tilde{S}_1$ are continuous at $\gamma^\infty_0$ in Assumption 3, let $\epsilon \to 0$, we can easily get,

$$FDP \overset{p}{\to} \frac{\pi_0 \tilde{F}_0(\gamma^\infty_0)}{\tilde{S}(\gamma^\infty_0)} \leq \frac{\gamma^\infty_0 \tilde{F}(\gamma^\infty_0)}{\tilde{S}(\gamma^\infty_0)} \leq \alpha.$$

The first inequality is due to Lemma 4.1, the second inequality is by definition and the last inequality is because of the continuity in Assumption 3(a).

### S2.5 Proof of Corollary 4.5

For some $j$, let $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n)$ 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 $\tilde{P}_j$ and all other rows $\tilde{P}_k = P_k$ for $k \neq j$. Define $(\tilde{F}_1, \cdots, \tilde{F}_M)$ as the corresponding filtering statistics (1) and $(\tilde{S}_1, \cdots, \tilde{S}_M)$ as the corresponding selection statistics (2) 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 AdaFilter Bonferroni procedure, let $\tilde{\gamma}_0^\text{Bon}_j$ be the new $\gamma_0^\text{Bon}$ using the new individual $P$-values. For the AdaFilter BH procedure, let $\tilde{\gamma}_0^\text{BH}_j$ 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 the AdaFilter Bonferroni procedure, if $S_j \leq \gamma_0$, then $\tilde{S}_j \leq \gamma_0$, thus

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

which means that $\tilde{\gamma}_0^\text{Bon}_j \geq \gamma_0^\text{Bon}$. Similarly, for the AdaFilter BH procedure using the same argument, we have $\tilde{\gamma}_0^\text{BH}_j \geq \gamma_0^\text{BH}$ when $S_j \leq \gamma_0$. As a consequence, for both AdaFilter procedures, we have $\tilde{S}_j \leq S_j \leq \gamma_0 \leq \tilde{\gamma}_j$. 


Table S1: 27 or the 32 AdaFilter selected genes for $r = 4$ with FDR controlled at $\alpha = 0.05$ where functional annotations are available in [26](Table S2). The AdaFilter selection p-values for these genes are also reported.
Table S2: scRNA-seq data analysis: marker genes at \( r = 10 \). Filtering, selection and adjusted AF p-values at \( r = 10 \) of marker genes are shown. FDR controlled at \( \alpha = 0.05 \).

| Gene | F value | S value | Adjusted selection p value |
|------|---------|---------|---------------------------|
| BATF | 3.48e-03| 4.10e-03| 9.03e-03 |
| CXCR6 | 1.81e-03 | 1.56e-02 | 2.87e-02 |
| LYST | 1.26e-02 | 2.35e-02 | 4.40e-02 |
| CCR8 | 3.03e-03 | 5.34e-03 | 1.14e-02 |
| CREM | 5.43e-03 | 1.57e-02 | 2.92e-02 |
| LAYN | 4.69e-03 | 6.93e-03 | 1.56e-02 |
| CTLA4 | 1.51e-02 | 2.54e-02 | 4.38e-02 |
| DUSP4 | 1.68e-03 | 4.90e-03 | 1.14e-02 |
| PHLDAA1 | 2.20e-02 | 2.41e-02 | 4.40e-02 |
| GAPDH | 1.90e-03 | 2.63e-03 | 1.19e-02 |
| ID2 | 2.15e-02 | 2.76e-02 | 4.41e-02 |
| TNFRSF9 | 4.61e-03 | 7.15e-03 | 1.43e-02 |
| FOXP3 | 4.57e-03 | 2.66e-02 | 4.34e-02 |
| SRGN | 8.63e-04 | 2.74e-03 | 8.21e-03 |
| PHTF2 | 5.21e-03 | 1.31e-02 | 2.30e-02 |
| SDC4 | 5.88e-03 | 9.99e-03 | 1.90e-02 |
| TNFRSF4 | 1.06e-02 | 1.19e-02 | 2.16e-02 |
| TNFRSF18 | 1.50e-04 | 1.97e-04 | 1.97e-04 |
| CD7 | 2.10e-03 | 2.90e-03 | 6.52e-03 |

Table S3: Metabalics GWAS data analysis: significant SNPs at \( r = 3 \) after clumping \((r^2 \text{ set to } 0.1 \text{ in PLINK})\). Individual p-value for each of the 8 super-pathways are shown. The significant ones for each marker are in bold. FDR controlled at \( \alpha = 0.05 \).
Figure S1: 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 S2: 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. $b = 100$ is the weak dependence scenario and $b = 1000$ is the strong dependence scenario.
Figure S3: 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 S4: 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.
Figure S5: Histogram of the individual p-values for each of the DMD datasets. The spike at 1 is due to the reason that we use Bonferroni combination rule to get a combined p-value when multiple probe-sets are referring to the same gene.
Figure S6: Correlation of test statistics of the $m = 275$ metabolites. The darker the color, the higher the absolute value of the correlation. The metabolite measurements are reordered so that metabolites in the same pathway are adjacent to each other. The red lines and blue texts label the eight super-pathways.