Submitted to the Annals of Statistics

DETECTING MULTIPLE REPLICATING SIGNALS USING ADAPTIVE FILTERING PROCEDURES

BY JINGSHU WANG[1,*], LIN GUI[1,†], WEIJIE J. SU[2], CHIARA SABATTI[3,‡] AND ART B. OWEN[3,§]

[1]*Department of Statistics, The University of Chicago,* [*]*jingshuw@uchicago.edu;* [†]*glin6@uchicago.edu*

[2]*Department of Statistics and Data Science, University of Pennsylvania, suw@wharton.upenn.edu*

[3]*Department of Statistics, Stanford University,* [‡]*sabatti@stanford.edu;* [§]*owen@stanford.edu*

Replicability is a fundamental quality of scientific discoveries: we are interested in those signals that are detectable in different laboratories, different 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.

1. Introduction. Replication is “the cornerstone of science" [34]. 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 [29] and the scientific press [3, 1] 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 [4], provides such a framework. Given n null hypotheses (base nulls) and a number $r \in \{2, 3, \dots, n\}$, the PC null states that there are fewer than r base non-nulls. In the setting where each base 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 [36], to detect genes that show consistent effects across genetic experiments [21], and recently to study mediation effects [32] and find evidence factors [27] in causal inference.

MSC 2010 subject classifications: Primary 62J15; secondary 62P10

Keywords and phrases: Simultaneous signals, meta-analysis, high-throughput experiments, composite null, multiple hypotheses testing

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, 43]. With a growing trend in multi-omics data sharing [20], there is also active research in finding replicating signals across platforms [47], ethnic groups [33, 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 standard 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 [23, 41]. Both [23] and [7] suggest procedures to counter that power loss. Unfortunately, the appoach in [7] is designed only for $n = r = 2$ and the empirical Bayes approach `repfdr` in [23] 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 (j, j') where the base p-values p_j 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 is robust to dependence of p-values within each study and can have much higher power than the "direct approach" 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 lead 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 several real studies. Section 7 has conclusions. An R package implementing AdaFilter is available at `https://github.com/jingshuw/adaFilter`.

2. Multiple testing for partial conjunctions.

In this section, we provide a brief introduction of the partial conjunction hypotheses and the low power in detecting multiple PC hypotheses using the “direct approach”.

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 each base P-value is valid, satisfying $\mathbb{P}(P_{ij}\le\gamma)\le\gamma$ under its null. Also, let $P_{(1)j}\le P_{(2)j}\le\cdots\le P_{(n)j}$ be the sorted P-values for each $j=1,2,\dots,M$.

DEFINITION 2.1 (Partial Conjunction Hypothesis). For integers $n\ge r\ge 2$, the partial conjunction (PC) null hypothesis is:

$$H_0^{r/n}: \text{fewer than } r \text{ out of } n \text{ base hypotheses are non-null.}$$

When $r=1$, $H_0^{1/n}$ is the commonly tested global null for meta-analysis. Rejecting it would not guarantee replicability. In high-throughput experiments, for each DNA locus or gene $j\in\{1,2,\dots,M\}$, 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 $\{H_{01}^{r/n},\dots,H_{0M}^{r/n}\}$, 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

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

where $\text{FDP}=V/(R\vee 1)$ 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 [44] and [4]. Consider a single PC null $H_0^{r/n}$ with a vector of base P-values $(P_1,P_2,\dots,P_n)$ and let $P_{r/n}=f(P_1,P_2,\dots,P_n)$ be the combined P-value for $H_0^{r/n}$. Benjamini and Heller [4] discussed three approaches, which we report here, using the standard notation $(P_{(1)}\le P_{(2)}\le\cdots\le P_{(n)})$:

1. Simes’ method:

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

2. Fisher’s method:

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

3. Bonferroni’s method:

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

The idea is to apply meta-analysis to the largest $n-r+1$ base P-values. For instance, if $n=r=2$, then $P^S_{2/2} = P^F_{2/2} = P^B_{2/2} = \max(p_1, p_2)$. All three methods construct valid PC P-values for $H_0^{r/n}$ under independence, and [44] 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 jth 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 α, we could use the Bonferroni rule, rejecting $H_{0j}^{r/n}$ if $P_{r/n,j} \le \alpha/M$, which also controls the PFER at level α [42]. To control the FDR we could apply BH procedure [5] 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 $\mathcal{I}_k \subset \{1, \cdots, M\}$ such that

$$\mathcal{I}_k = \big\{ j \in \{1, \cdots, M\} \mid \text{exactly } k \text{ of } H_{01j}, \ldots, H_{0nj} \text{ are false} \big\} \tag{1}$$

for $k = 0, \ldots, n$. Sets $\{\mathcal{I}_k,\, k = 0, \ldots, n\}$ define a partition of $\{1, \ldots, M\}$. If a false rejection of $H_{0j}^{n/n}$ happens, then the jth 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 α, the true FWER instead satisfies

$$\begin{aligned} \text{FWER} \le \mathbb{E}(V) &= \sum_{k=0}^{n-1} \sum_{j \in \mathcal{I}_k} \mathbb{P}(P_{(n)j} \le \alpha/M) \\ &\le \sum_{k=0}^{n-1} \sum_{j \in \mathcal{I}_k} \frac{\alpha^{n-k}}{M^{n-k}} = \sum_{k=0}^{n-1} |\mathcal{I}_k| \frac{\alpha^{n-k}}{M^{n-k}}. \end{aligned}$$

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 we have

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

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 $\delta_{n-1} = 0$) for large M. 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 α 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 $r \le n$, the level of $\mathbb{E}(V)$ for Bonferroni correction will depend mainly on δ_{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 δ_k were known or if good estimates of δ_k can be obtained. This is what motivates the Bayesian methods [23, 14]. In this paper we take a frequentist perspective. Rather than estimating δ_k, AdaFilter works directly on an alternative estimation of V, implicitly and adaptively adjusting for the size of δ_{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 $\mathbb{P}(P_{r/n} \le \gamma) \le \gamma$ for a given γ is only tight for the least favorable null, while standard multiple testing procedures are designed to control error when $\mathbb{P}(P_{r/n} \le \gamma) = \gamma$ is always true. To overcome this, AdaFilter leverages a region $\mathcal{A}_\gamma \subset [0,1]^n$

such that the much tighter inequality

$$\mathbb{P}(P_{r/n,j} \le \gamma \mid (P_{1j}, \ldots, P_{nj}) \in \mathcal{A}_\gamma) \le \gamma$$

holds for any configuration in the PC null space.

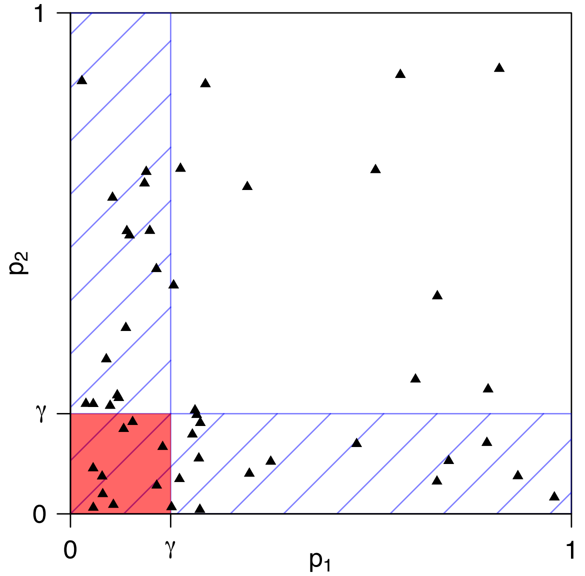


Fig 1: Illustration of the rejection (red) and filtering (L-shaped blue) regions at $\gamma = 0.2$ when $n = r = 2$. Each triangle corresponds to a pair of p-values.

Figure 1 illustrates the construction of the filtering region $\mathcal{A}_\gamma$ 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 $\mathbb{P}(P_{2/2,j} \le \gamma) \le \gamma^2$ under (True, True), while $\mathbb{P}(P_{2/2,j} \le \gamma)$ can be close to γ under the other two less favorable configuration. Let us consider, instead, conditioning on (P_{1j}, P_{2j}) being in the "L"-shaped filtering region $\mathcal{A}_\gamma = \{(p_1, p_2) \mid \min(p_1, p_2) \le \gamma\}$. We get $\mathbb{P}(P_{2/2,j} \le \gamma \mid (P_{1j}, P_{2j}) \in \mathcal{A}_\gamma) \le \gamma$ being true for all three null scenarios, which is much tighter than $\mathbb{P}(P_{r/n} \le \gamma) \le \gamma$. 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 γ: $\widehat{V}_\gamma = \gamma M$, we can improve their efficiency with a smaller estimate of $\widehat{V}_\gamma$ using the new inequality. Specifically, the estimated V is now $\widehat{V}_{\mathcal{A}_\gamma} = \gamma \times \sum_{j=1}^{M} 1_{(P_{1j}, P_{2j}) \in \mathcal{A}_\gamma}$, 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 $(1/M) \sum_{j=1}^{M} 1_{(P_{1j}, P_{2j}) \in \mathcal{A}_\gamma}$ is our "estimate" of δ_{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 α, we can adaptively choose the largest γ satisfying $\widehat{V}_{\mathcal{A}_\gamma} \le \alpha$. Similarly, to control the FDR at level α, we estimate the FDP as $\widehat{V}_{\mathcal{A}_\gamma}/(R \vee 1)$ and select the largest γ such that $\widehat{V}_{\mathcal{A}_\gamma}/(R \vee 1) \le \alpha$. These are essentially the Bonferroni or BH procedure with an alternative estimate of V.

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

$$F_j := (n-r+1)P_{(r-1)j}, \quad \text{and} \tag{3}$$

$$S_j := P^B_{r/n,j} = (n-r+1)P_{(r)j}, \tag{4}$$

respectively.

DEFINITION 3.1 (AdaFilter Bonferroni). For a level α, and with F_j and S_j given by (3) and (4) respectively, reject $H_{0j}^{r/n}$ if $S_j < \gamma_0^{\text{Bon}}$ where

$$\gamma_0^{\text{Bon}} = \sup\left\{\gamma \in [0,\alpha] \,\middle|\, \gamma \sum_{j=1}^{M} 1_{F_j<\gamma} \le \alpha\right\}.$$

DEFINITION 3.2 (AdaFilter BH). For a level α, and with F_j and S_j given by (3) and (4) respectively, reject $H_{0j}^{r/n}$ if $S_j < \gamma_0^{\text{BH}}$ where

$$\gamma_0^{\text{BH}} = \sup\left\{\gamma \in [0,\alpha] \,\middle|\, \frac{\gamma\sum_{j=1}^{M} 1_{F_j<\gamma}}{\sum_{j=1}^{M} 1_{S_j<\gamma} \vee 1} \le \alpha\right\}.$$

REMARK 3.1. We define the filtering region as $\{F_j < \gamma\}$ instead of $\{F_j \le \gamma\}$ to guarantee that γ_0^{Bon} and γ_0^{BH} themselves satisfy the corresponding inequalities. This is important for showing the theoretical properties of adaFilter procedures, especially when base p-values are discrete. The rejection criterion is set to $S_j < \gamma_0$ instead of $S_j \le \gamma_0$ where γ_0 is either γ_0^{Bon} or γ_0^{BH} accordingly (for Lemma 4.1).

We also introduce AdaFilter adjusted "p-values" like those commonly computed for standard Bonferroni and BH procedures. They provide equivalent sets of rejections as the above definitions, while can be more efficiently computed.

DEFINITION 3.3 (AdaFilter adjusted p-values). Rank the selection p-values as $S_{(1)} \le S_{(2)} \le \cdots \le S_{(M)}$ where $S_{(j)}$ is for the null hypothesis $H_{0(j)}^{r/n}$. For each j, define an AdaFilter adjustment number

$$m_{(j)}^{\text{AF}} := \sum_{h=1}^{M} 1_{F_h \le S_{(j)}}.$$

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

$$P_{(j)}^{\text{Bon}} = S_{(j)} m_{(j)}^{\text{AF}}$$

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

$$P_{(j)}^{\text{BH}} = \min\left\{\min_{h \ge j}\left\{S_{(h)}\frac{m_{(h)}^{\text{AF}}}{h}\right\}, 1\right\}.$$

For any level $\alpha > 0$, we reject the hypotheses whose AdaFilter adjusted p-values are smaller than α. We can verify that the AdaFilter adjusted p-values give the same set of rejections as Definition 3.1 and Definition 3.2.

PROPOSITION 3.4. *For any level $\alpha > 0$, the set of rejections defined as $\{j : P_j^{Bon} < \alpha\}$ is equivalent to the set of rejections from Definition 3.1. Similarly, the set of rejections defined as $\{j : P_j^{BH} < \alpha\}$ is equivalent to the set of rejections from Definition 3.2.*

In practice, the AdaFilter adjusted p-values can be more easily computed than finding γ_0^{Bon} and γ_0^{BH}. Our simulations and real data applications in Sections 5 and 6 also compute these adjusted p-values for getting the rejections of AdaFilter procedures.

3.2. *A heuristic comparison with the "direct approach".* Before we discuss the theoretical properties of AdaFilter procedures in Section 4, we revisit the case of $r = n$ in Section 2.2 to understand the level of power gain from AdaFilter procedures compared with the "direct approach". When $r = n$, the PC p-values for the "direct approach" are $P_{r/n,j} = P_{(n)j}$, which are the same as the selection p-values of AdaFilter procedures. As a consequence, AdaFilter procedures would not change the ordering/ranking of the individual PC hypotheses. AdaFilter gains power by selecting a much less conservative PC p-values threshold γ than the "direct approach" for the same nominal FWER/FDR level.

If one controls FWER at level α, then the PC p-value threshold from the "direct approach" using Bonferroni adjustment is α/M. We now give an approximation of the threshold from AdaFilter Bonferroni. When $r = n$, at any given threshold γ, the estimate of the number of false discoveries used in AdaFilter is

$$\hat{V}(\gamma) = \gamma \sum_{i=1}^{M} 1_{F_j < \gamma} = \gamma \sum_{i=1}^{M} 1_{P_{(n-1),j} < \gamma}.$$

AdaFilter Bonferroni finds the largest γ so that $\hat{V}(\gamma) \leq \alpha$. As defined in (1), let $\mathcal{I}_k \subset \{1, \cdots, M\}$ be the set of hypotheses with exactly k base non-nulls and let $\delta_k = |\mathcal{I}_k|/M$. When M is large, the expected value of $\hat{V}(\gamma)$ satisfies that

$$\begin{aligned}
\mathbb{E}\left(\hat{V}(\gamma)\right) &= \gamma \sum_{k=0}^{n} \sum_{j \in \mathcal{I}_k} \mathbb{P}\left(P_{(n-1),j} < \gamma\right) \\
&\leq \gamma \left(|\mathcal{I}_n| + |\mathcal{I}_{n-1}| + \sum_{k=0}^{n-2} |\mathcal{I}_k| \cdot \left((n-k)\gamma^{n-k-1}(1-\gamma) + \gamma^{n-k} \right) \right) \\
&\leq \gamma M(\delta_n + \delta_{n-1}) + MO(\gamma^2).
\end{aligned}$$

The first inequality is due to the fact that all base null p-values are independent and for each $j \in \mathcal{I}_k$, we can decompose $\mathbb{P}\left(P_{(n-1),j} < \gamma\right)$ into the events that all $n-k$ base nulls i satisfy $P_{ij} \leq \gamma$ and exactly $n-k-1$ base nulls satisfy this constraint. So roughly, the AdaFilter Bonferroni threshold γ^{Bon} will be around some value that is at least $\alpha/\big(M(\delta_n + \delta_{n-1})\big) + o(1/M)$. Compared with the Bonferroni threshold α/M in the "direct approach", AdaFilter Bonferroni increases this threshold by $1/(\delta_n + \delta_{n-1})$. In our motivating applications, both δ_n and δ_{n-1} are typically small, and so such an increase would be substantial. The resulting actual FWER is also less conservative. If we use a fixed threshold at $\gamma = \alpha/\big(M(\delta_n + \delta_{n-1})\big)$, then

$$\mathbb{E}(V) \leq \alpha \Big\{ \frac{\delta_{n-1}}{\delta_{n-1} + \delta_n} + O(\frac{1}{M}) \Big\}.$$

Compared to the bound $\alpha\delta_{n-1} + O(1/M)$ in (2) from the "direct approach", we can now be much less conservative especially when the proportion of least favorable PC nulls δ_{n-1} is small.

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 $\gamma > 0$*

$$\mathbb{P}\big(S_j < \gamma \mid F_j < \gamma\big) \leq \gamma \tag{5}$$

holds whenever $\mathbb{P}\big(F_j < \gamma\big) > 0$. Here F_j and S_j are given by (3) and (4), respectively.

Inequality (5) can be equivalently written as $\mathbb{P}\big(S_j < \gamma\big) \leq \gamma\mathbb{P}(F_j < \gamma)$, which holds even when $\mathbb{P}(F_j < \gamma) = 0$ as $S_j \geq F_j$ is always true. Intuitively, the "conditional validity" guarantees that for a fixed threshold γ, the estimated upper bound on the number of false rejections V is $\gamma \sum_j 1_{F_j<\gamma}$. However, AdaFilter uses a data-dependent γ, 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 3.1 controls FWER and PFER at level α for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \ldots, M\}$.*

REMARK 4.1. Though we name our method AdaFilter Bonferroni, we can only prove FWER/PFER control under independence of the p-values within each study, though simulations in Section 5 show that FWER/PFER control can also be achieved in practice for dependent p-values within each study.

REMARK 4.2. For controlling for FWER, one can combine adaFilter Bonferroni with the sequential rejection principle [18] to further increase the number of rejections while controlling for FWER at the same level. Intuitively, this is similar to improving the standard Bonferroni procedure with Holm's procedure. For a more detailed discussion, see Section S1.

For AdaFilter BH, however, we can only prove that it controls FDR at the nominal level of $\alpha C(M)$ where $C(M) = \sum_{j=1}^{M} 1/j \approx \log M$. In other words, adjusting the threshold to be $\alpha/C(M)$ can guarantee control of the FDR at level α.

THEOREM 4.3. *Let $(P_{ij})_{n\times M}$ contain independent valid p-values. Then AdaFilter BH in Definition 3.2 controls FDR at level $\alpha C(M)$ where $C(M) = \sum_{j=1}^{M} 1/j$ for the null hypotheses $\{H_{0j}^{r/n} : j = 1, 2, \cdots, M\}$.*

The inflation factor $C(M)$ in Theorem 4.3 for the adaFilter BH procedure is due to a technical difficulty encountered when proving for FDR control for finite 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 actually not needed in practice. In Section 4.2, we will show that AdaFilter BH can 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 γ

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

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 $\mathcal{H}_0^{r/n} = \{j : H_{0j}^{r/n} \text{ is true}\}$ be the set of true PC nulls and M_0 be its cardinality. Similarly, define $\mathcal{H}_1^{r/n} = \{j : H_{1j}^{r/n} \text{ is true}\}$ 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:

$$\lim_{M\to\infty} \frac{M_0}{M} = \pi_0 \in (0, 1)$$

$$\lim_{M\to\infty} \frac{1}{M_0} \sum_{j \in \mathcal{H}_0^{r/n}} P(F_j < \gamma) = \tilde{F}_0(\gamma), \quad \lim_{M\to\infty} \frac{1}{M_1} \sum_{j \in \mathcal{H}_1^{r/n}} P(F_j < \gamma) = \tilde{F}_1(\gamma)$$

$$\lim_{M\to\infty} \frac{1}{M_0} \sum_{j \in \mathcal{H}_0^{r/n}} P(S_j < \gamma) = \tilde{S}_0(\gamma), \quad \lim_{M\to\infty} \frac{1}{M_1} \sum_{j \in \mathcal{H}_1^{r/n}} P(S_j < \gamma) = \tilde{S}_1(\gamma).$$

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 $\boldsymbol{c} \in \{0, 1\}^n$ representing one of the 2^n combinations, let $m_{\boldsymbol{c}}$ be the number of PC hypotheses that fall into this combination. Also, let $\mathcal{H}_{0i}$ and $\mathcal{H}_{1i}$ be the sets of true nulls and true non-nulls for the ith study. If (a) $\lim_{M\to\infty} m_{\boldsymbol{c}}/M$ exists for all $\boldsymbol{c}$ and, (b) for each i, $\{P_{ij} : j \in \mathcal{H}_{0i}\}$ have identical distributions across j and $\{P_{ij} : j \in \mathcal{H}_{1i}\}$ also have identical distributions across j, 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 γ as

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

and the largest γ_0^{∞} 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 γ_0^{∞}:

ASSUMPTION 3 (Technical conditions). The following two conditions hold:

(a) There exists $\delta > 0$ such that $f^{\infty}(\gamma)$ is monotonically increasing 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 γ_0^{∞}.

Intuitively, (a) guarantees that the limit of the AdaFilter threshold γ_0^{BH} is unique when $M \to \infty$ and (b) is satisfied if there are sufficient points (selection p-values) around γ_0^{∞} when M is large. Now we are ready to state the asymptotic FDR control of AdaFilter BH.

THEOREM 4.4. *Under Assumptions 1-3, the AdaFilter BH procedure of Definition 3.2 satisfies*

$$\gamma_0^{\mathrm{BH}} \xrightarrow{p} \gamma_0^{\infty}, \quad \textit{and}$$

$$\mathrm{FDP} \xrightarrow{p} \frac{\pi_0 \tilde{S}_0(\gamma_0^{\infty})}{\tilde{S}(\gamma_0^{\infty})} \leq \alpha$$

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

Notice that Assumption 3(a) implies that $f^{\infty}(\gamma_0^{\infty}) > 0$, thereby guaranteeing $\tilde{S}(\gamma_0^{\infty}) > 0$.

REMARK 4.3. 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 γ,

$$\frac{1}{M_s^2} \sum_{j \neq j' \in \mathcal{H}_s^{r/n}} \left| \mathbb{P}(P_{ij} < \gamma, P_{ij'} < \gamma) - \mathbb{P}(P_{ij} < \gamma)\mathbb{P}(P_{ij'} < \gamma) \right| \xrightarrow{M_s \to \infty} 0$$

for both $s = 0, 1$. We can not deal with $\pi_0 = 1$ as that would lead to $\tilde{S}(\gamma_0^{\infty}) = 0$ 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 an unexpected power gain when combining multiple similar studies. Suppose that we test the involvement of M genes in a disease with two 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 Table 1a 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 $H_{01}^{2/2}$ 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 base p-values can never cause any of the decisions on the null hypotheses to switch from 'reject' to 'accept'.

(a)

	Study			
j	1	2	F_j	S_j
1	0.04	0.03	0.03	0.04
2	0.5	0.9	0.5	0.9

(b)

	Study			
j	1	2	F_j	S_j
1	0.04	0.03	0.03	0.04
2	0.01	0.9	0.01	0.9

TABLE 1

(a) Toy example where AdaFilter is more efficient than testing for each study separately. Values are the p-values. (b) A counterexample to show that AdaFilter violates "complete monotonicity". The significance level is $\alpha = 0.05$.

DEFINITION 4.5 (Complete monotonicity). A multiple testing procedure has complete monotonicity if each decision function φ_j is a non-increasing function in all the elements of $(p_{ij})_{n\times M}$ for $j = 1, 2, \cdots, M$.

Simes', Fisher's and Bonferroni's meta-analyses have complete monotonicity. So does the BH procedure with $n = 1$. Heller, Bogomolov and Benjamini [21] call this property "stability" and it holds for the PC tests of [22]. However, AdaFilter do not satisfy complete monotonicity: lowering one of the p-values for gene j can change the rejection of $H_{0,j'}^{r/n}$ to acceptance for $j' \neq j$.

Table 1b shows how AdaFilter does not have complete monotonicity. Compared with Table 1a, the second hypothesis has a decreased p-value in study 1 while all other p-values are kept fixed. In Table 1a, both $\gamma_0^{\text{Bon}} = \gamma_0^{\text{BH}} = 0.05$ so the first PC hypothesis is rejected. In contrast, in Table 1b $\gamma_0^{\text{Bon}} = \gamma_0^{\text{BH}} = 0.03$ 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 base p-values may make the signal configuration across genes more similar among studies. 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 $H_{0,j}^{r/n}$ to accept.

DEFINITION 4.6 (Partial monotonicity). A multiple testing procedure has partial monotonicity if for all $j \in \{1, \cdots, M\}$, its decision function $\varphi_j(p_{\cdot 1}, \ldots, p_{\cdot M})$ 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 4.7. *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 4.7 indicates that AdaFilter is reasonable in a way that reducing the base p-values of the jth PC hypothesis indeed strengthens the evidence of replicability for the jth PC hypothesis, though possibly weakening the evidence of replicability for other PC hypotheses.

4.4. *Extensions and discussion of related literature.*

4.4.1. *Comparison with other strategies.* Two directly related methods to AdaFilter are [7] for $n = r = 2$ and the empirical Bayes approach in [23] 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 [7] but works for any n and r. It provides a frequentist approach comparable to and sometimes better than [23].

The procedures of [7] 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 γ_1 and γ_2 to satisfy

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

When $\gamma_1 \approx \gamma_2$, then

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

Thus $\gamma_0^{\text{Bon}} \approx \gamma_1 \approx \gamma_2$ and AdaFilter becomes similar to their procedure. The proposed method only applies for $n = r = 2$; this simplification makes the approach less widely applicable, despite its strong theoretical guarantees. In addition, for $n = r = 2$, some other methods [10, 9] have also discussed powerful multiple testing procedures controlling for FWER and in [**?**], the authors proposed a new procedure controlling for local FDR.

In `repfdr` [23], 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 [43, 45, 48], all of which use an empirical Bayes framework as in [23]. Compared to these methods, AdaFilter is typically faster, guarantees simultaneous error rate control and is more robust to the dependence of p-value within each study.

Finally, there has been much other recent literature on efficient FDR control by using some special data structure as prior knowledge [30, 31, 2, 6] 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.

4.4.2. *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 jth 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 (3) and (4) 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.

4.4.3. *Requiring sign replicability.* Partial conjunctions with two-sided test statistics can reject $H_{0j}^{r/n}$ 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 either 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 [35]. 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 $\{H_{01}^{r/n,+},\dots,H_{0M}^{r/n,+}\}$ and $\{H_{01}^{r/n,-},\dots,H_{0M}^{r/n,-}\}$. The PC null $H_{0j}^{r/n,+}$ is rejected if the signal j is positive in at least r studies, and $H_{0j}^{r/n,-}$ 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 $H_{0j}^{r/n,+}$ and $H_{0j}^{r/n,-}$ for the same j.

We can apply AdaFilter twice, separately on $\{H_{01}^{r/n,+},\dots,H_{0M}^{r/n,+}\}$ and $\{H_{01}^{r/n,-},\dots,H_{0M}^{r/n,-}\}$, controlling the simultaneous error rate (FWER, PFER or FDR) at levels α_1 and α_2 respectively, with $\alpha_1+\alpha_2=\alpha$ (ordinarily $\alpha_1=\alpha_2=\alpha/2$). Let the set of rejected PC nulls be $\mathcal{R}^{+}$ and $\mathcal{R}^{-}$, respectively. Rejecting the union of these two sets $\mathcal{R}^{\pm}=\mathcal{R}^{+}\cup\mathcal{R}^{-}$ controls the corresponding error rate at a level $\alpha=\alpha_1+\alpha_2$ for the null hypotheses $\{H_{01}^{r/n,\pm},\dots,H_{0M}^{r/n,\pm}\}$.

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

4.4.4. *Testing for all possible values of r.* The partial conjunction null $H_0^{r/n}$ can be meaningfully defined whenever $2 \le r \le n$, and sometimes it is of interest to test for all possible r values, adding another layer of multiplicity. In [4], it is shown that as the PC p-values $P_j^{r/n}$ are monotone increasing when r increases, the "direct approach" can control for multiple r values simultaneously, without any further multiplicity adjustment of r. Unfortunately, this is not true for AdaFilter. 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 the "direct approach" with the three forms of PC p-values in Section 2.2. For FDR control, we also include [23], 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 2a. 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: π_{00} is the probability of the global null combination and π_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 π_{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 $I_{b\times b} \otimes \Sigma_\rho$ where $\otimes$ is the Kronecker product. The covariance block $\Sigma_\rho \in \mathbb{R}^{M/b\times M/b}$ has 1s on the diagonal and common value $\rho = 0.5$ off the diagonal. We set the number of blocks $b = 100$ for weak dependence and $b = 10$ 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 three forms of the "direct approach". 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 2b shows the average PFER and recall over the six combinations of n and r for each setting of b and π_{00}. More detailed results for each n and r separately are shown in Figures 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 π_{00} is higher, which is expected in many genetics applications.

Table 2c shows the average FDR and recall over the six combinations of n and r for each setting of b and π_{00}. More detailed results for each n and r separately are shown in Figure S3–S4. AdaFilter BH, even not inflated, and the "direct approach" control FDR at the nominal level. However, similar to the PFER control, the "direct approach" procedures are too conservative. The inflated AdaFilter BH has lower power than AdaFilter BH, while its power still exceed the "direct approach", 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.

Finally, we point out that our simulations only compare different methods for a pre-defined r value. As discussed in Section 4.4.4, AdaFilter needs another layer of multiplicity adjustment if multiple r values are tested simultaneously. In practice, if one aims to testing for mulitple replicability levels or is interested in obtaining the lower bound of r for each hypotheses [26], the "direct approach" may still be a preferred method as it automatically controls for the error rates of multiple r values simultaneously.

(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 $\alpha = 1$

	$\pi_{00} = 0.8$				$\pi_{00} = 0.98$			
	$b = 100$		$b = 10$		$b = 100$		$b = 10$	
Method	PFER	Recall(%)	PFER	Recall(%)	PFER	Recall(%)	PFER	Recall(%)
Bon-$P^B_{r/n}$	0.04	14.72	0.05	14.87	0.00	14.72	0.00	14.83
Bon-$P^F_{r/n}$	0.05	19.30	0.06	19.50	0.01	19.18	0.00	19.38
Bon-$P^S_{r/n}$	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 $\alpha = 0.2$

	$\pi_{00} = 0.8$				$\pi_{00} = 0.98$			
	$b = 100$		$b = 10$		$b = 100$		$b = 10$	
Method	FDR	Recall(%)	FDR	Recall(%)	FDR	Recall(%)	FDR	Recall(%)
BH-$P^B_{r/n}$	0.01	29.50	0.01	29.55	0.00	29.04	0.00	29.10
BH-$P^F_{r/n}$	0.01	32.94	0.01	32.80	0.00	32.68	0.00	32.74
BH-$P^S_{r/n}$	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	11.56
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

TABLE 2

Simulation results. (a) lists 6 different n and r scenarios considered in the simulation. (b) and (c) compare the average error rates and recalls across all 6 n and r combinations under different b and π_{00} values. The results for each n and r are shown in Figure S1 - S4.

6. Case studies.

We apply AdaFilter to analyze two datasets: one investigates the replication of gene differential expression results in four microarray experiments of Duchenne muscular dystrophy and one focuses on identifying marker genes of one T cell subtype from lung cancer tumors using single-cell RNA-sequencing (scRNA-seq) data. In Section S2, we also discuss the application of AdaFilter BH to a third dataset, testing for consistently significant signals across different metabolic super-pathways within one study.

6.1. *Duchenne Muscular Dystrophy microarray studies.*

Following [28], 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 3a), 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` [25] for microarrays. Within each study, we find genes that are differentially expressed between the disease and healthy group, using a popular software `Limma` [40] 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. There are only $M = 1871$ genes present in all four studies, with $M = 9848$ genes shared in at least 3 studies and $M = 13912$

(a) GEO datasets information

GEO ID	Platform	Description	Source
GDS 214	custom Affymetrix	4 healthy, 26 DMD	Muscle
GDS 563	Affymmetrix 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-69	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

TABLE 3

Replicability analysis for DMD microarrays

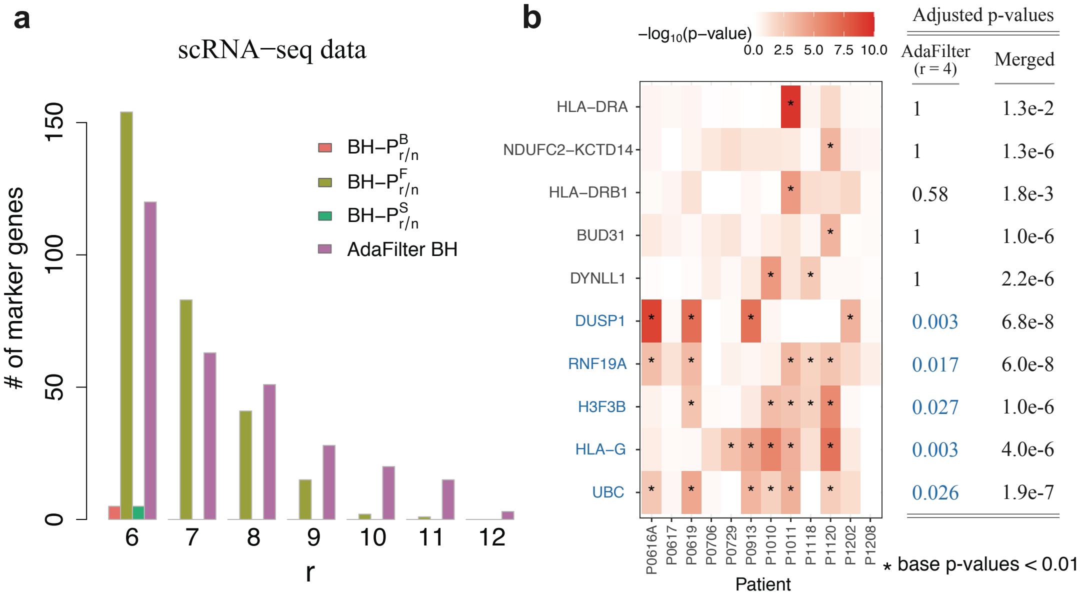
genes in at least two studies. As discussed in Section 4.4.2, AdaFilter can work with varying n_j thus allow missing entries in the p-value matrix.

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 3b). Specifically, at $r = 4$, AdaFilter BH finds 32 significant genes (Table S2). By contrast, a BH adjustment on the Fisher combined PC p-values ($P^F_{r/n,j}$) 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 3c 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 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 [19], the authors sequenced tumor T cells from 14 treatment-naïve non-small-cell lung cancer patients and one main finding is the discovery of a new subtype of the CD4+ regulatory T cells (Tregs), named the suppressive tumor-resident Tregs (CD4-C9-CTLA4), that is different from the normal Tregs (CD4-C8-FOXP3). We download data from the GEO database (GSE99254), where cell type labels are also provided.

In order to characterize the new cell type 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, which is the standard test for analyzing scRNA-seq. Two patients who have less than 10 Treg cells in either of the two groups 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 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, as the power gain of Fisher's combination to construct PC p-values may exceed the power gain using AdaFilter, whose selection p-values are from the Bonferroni's combination. The other two forms of "direct approach" show limited power for all r and `repfdr` fails to run with insufficient memory for $r \geq 6$ even with 300G of RAM. In Table S3, we list the 20 genes that are detected at $r = 10$, most of which are known to be linked to immunoresponse in tumors.



	Adjusted p-values	
	AdaFilter (r = 4)	Merged
HLA−DRA	1	1.3e-2
NDUFC2−KCTD14	1	1.3e-6
HLA−DRB1	0.58	1.8e-3
BUD31	1	1.0e-6
DYNLL1	1	2.2e-6
DUSP1	0.003	6.8e-8
RNF19A	0.017	6.0e-8
H3F3B	0.027	1.0e-6
HLA−G	0.003	4.0e-6
UBC	0.026	1.9e-7

Fig 2: (a) scRNA-seq data: the number of genes whose $H_{0j}^{r/n}$ were rejected by each of the compared procedures. FDR is controlled at $\alpha = 0.05$. (b) The left is a heatmap of each patient's one-sided Wilcoxon rank-sum p-values for 10 genes. The darker color represents a smaller p-value and a '*' label is added if it 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 standard BH adjusted merged p-values combining cells in all patients.

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). All 10 genes in Figure 2b would be selected in the original paper as their adjusted merged p-values are far less than 0.05. However, the top 5 genes only have one or two patients whose base p-values are less than 0.01. Intuitively, they are less convincing markers as there is no replicability across patients. While the merged p-values can not distinguish the more convincing markers, they can easily be separated with their AdaFilter BH adjusted p-values.

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. In our simulations, we demonstrated that both AdaFilter Bonferroni and AdaFilter BH are robust to the dependence of p-values within each study in practice, even when such dependence is not weak. On the other hand, the validity of AdaFilter does need independence of the base p-values across different studies, as Lemma 4.1 can be easily violated when these base p-values are dependent.

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 [24], "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. This work was supported by the National Science Foundation under grants IIS-1837931, DMS-1521145 and DMS-2113646, and by the National Institutes of Health under grants R01MH101782 and U01HG007419. We thank Y. Benjamini, M. Bogomolov, R. Heller and N. R. Zhang for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplementary information (SI). The SI includes supplementary text (Sections S1-S3), Table S1-S4 and Figure S1-S6.
().

REFERENCES

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

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

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

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

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

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

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

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

[9] Djordjilović, V., Hemerik, J. and Thoresen, M. (2020). On optimal two-stage testing of multiple mediators. *arXiv preprint arXiv:2007.02844*.

[10] Djordjilović, V., Page, C. M., Gran, J. M., Nøst, T. H., Sandanger, T. M., Veierød, M. B. and Thoresen, M. (2019). Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in medicine* **38** 3346–3360.

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

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

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

[14] FLUTRE, T., WEN, X., PRITCHARD, J. and STEPHENS, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* **9** e1003486.

[15] FRISTON, K. J., PENNY, W. D. and GLASER, D. E. (2005). Conjunction revisited. *NeuroImage* **25** 661–667.

[16] GIRI, A., HELLWEGE, J. N., KEATON, J. M., PARK, J., QIU, C., WARREN, H. R., TORSTENSON, E. S., KOVESDY, C. P., SUN, Y. V., WILSON, O. D. et al. (2019). Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nature genetics* **51** 51–62.

[17] GOEMAN, J. J., HEMERIK, J. and SOLARI, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics* **49** 1218–1238.

[18] GOEMAN, J. J. and SOLARI, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* 3782–3810.

[19] GUO, X., ZHANG, Y., ZHENG, L., ZHENG, C., SONG, J., ZHANG, Q., KANG, B., LIU, Z., JIN, L., XING, R. et al. (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature medicine* **24** 978–985.

[20] HASIN, Y., SELDIN, M. and LUSIS, A. (2017). Multi-omics approaches to disease. *Genome biology* **18** 83.

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

[22] HELLER, R., GOLLAND, Y., MALACH, R. and BENJAMINI, Y. (2007). Conjunction group analysis: an alternative to mixed/random effect analysis. *Neuroimage* **37** 1178–1185.

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

[24] HUME, D. (2003). *A treatise of human nature*. Courier Corporation.

[25] IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

[26] JALJULI, I., BENJAMINI, Y., SHENHAV, L., PANAGIOTOU, O. and HELLER, R. (2019). Quantifying replicability and consistency in systematic reviews. *arXiv preprint arXiv:1907.06856*.

[27] KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2021). 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* **116** 82–92.

[28] KOTELNIKOVA, E., SHKROB, M. A., PYATNITSKIY, M. A., FERLINI, A. and DARASELIA, N. (2012). Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in Duchenne muscular dystrophy. *PLoS Comput Biol* **8** e1002365.

[29] LEHRER, J. (2010). The truth wears off. *The New Yorker*.

[30] LEI, L. and FITHIAN, W. (2016). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B*.

[31] LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B* **81** 45–74.

[32] LIU, Z., SHEN, J., BARFIELD, R., SCHWARTZ, J., BACCARELLI, A. A. and LIN, X. (2021). Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. *Journal of the American Statistical Association* 1–15.

[33] MARIGORTA, U. M. and NAVARRO, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS genetics* **9**.

[34] MOONESINGHE, R., KHOURY, M. J. and JANSSENS, A. C. J. W. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Medicine* **4** e28.

[35] OWEN, A. B. (2009). Karl Pearson's Meta-Analysis Revisited. *Annals of Statistics* **37** 3867–3892.

[36] PRICE, C. J. and FRISTON, K. J. (1997). Cognitive Conjunction: A New Approach to Brain Activation Experiments. *NeuroImage* **5** 261–270.

[37] PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** 559–575.

[38] ROUSSEEUW, P. J. and DRIESSEN, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

[39] SHIN, S. Y., FAUMAN, E. B., PETERSEN, A. K., KRUMSIEK, J., SANTOS, R., HUANG, J. et al. (2014). An atlas of genetic influences on human blood metabolites. *Nature genetics* **46** 543–550.

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

[41] SUN, W., REICH, B. J., CAI, T. T., GUINDANI, M. and SCHWARTZMAN, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B* **77** 59–83.

[42] TUKEY, J. W. (1953). The Problem of Multiple Comparisons Technical Report, Princeton University.

[43] URBUT, S. M., WANG, G., CARBONETTO, P. and STEPHENS, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics* **51** 187–195.

[44] WANG, J. and OWEN, A. B. (2018). Admissibility in partial conjunction testing. *Journal of the American Statistical Association* 1–11.

[45] XIANG, D., ZHAO, S. D. and CAI, T. T. (2019). Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests. *Journal of the Royal Statistical Society: Series B* **81** 707–734.

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

[47] ZHANG, N. R., SENBABAOGLU, Y. and LI, J. Z. (2010). Joint estimation of DNA copy number from multiple platforms. *Bioinformatics* **26** 153–160.

[48] ZHAO, S. D., NGUYEN, Y. T. et al. (2020). Nonparametric false discovery rate control for identifying simultaneous signals. *Electronic Journal of Statistics* **14** 110–142.