E-values as unnormalized weights in multiple testing

Nikolaos Ignatiadis* Ruodu Wang† Aaditya Ramdas‡

July 20, 2023

Abstract

We study how to combine p-values and e-values, and design multiple testing procedures where both p-values and e-values are available for every hypothesis. Our results provide a new perspective on multiple testing with data-driven weights: while standard weighted multiple testing methods require the weights to deterministically add up to the number of hypotheses being tested, we show that this normalization is not required when the weights are e-values that are independent of the p-values. Such e-values can be obtained in the meta-analysis setting wherein a primary dataset is used to compute p-values, and an independent secondary dataset is used to compute e-values. Going beyond meta-analysis, we showcase settings wherein independent e-values and p-values can be constructed on a single dataset itself. Our procedures can result in a substantial increase in power, especially if the non-null hypotheses have e-values much larger than one.

Keywords: weighted multiple testing, false discovery rate, p-values, e-values, normalization.

1 Introduction

The p-value is perhaps the most commonly used inferential device in statistical practice. Traditional procedures for multiple testing, such as the procedure of Benjamini and Hochberg [1995] for controlling the false discovery rate, begin with a list of p-values as the input. The e-value is an alternative inferential tool that encompasses betting scores, likelihood ratios, and stopped supermartingales, e.g., Shafer [2021], Vovk and Wang [2021], Grünwald et al. [2021], and Howard et al. [2020, 2021]. For example, the "universal inference" e-value has gained popularity, and has recently led to the first known valid tests for many composite null hypotheses, such as testing mixtures, e.g., testing if the data comes from a mixture of Gaussians [Wasserman et al., 2020], or testing for shape constraints, e.g., testing if the data distribution is log-concave [Dunn et al., 2021]. Lists of e-values can also serve as the input to multiple testing procedures [Wang and Ramdas, 2022, Xu et al., 2021].

In this paper, we design testing procedures for situations in which we have both a p-value and an e-value for each hypothesis. A first motivation for our proposed methods is the *meta-analysis* setting wherein we collect data from two distinct sources. Our contributions to meta-analysis acknowledge our anticipation that e-values will increasingly find adoption in applications without displacing p-values. Thus it is natural to develop procedures that can optimally combine the information available in an e-value and a p-value. What's more, we argue that our proposed meta-analysis methods are useful even when the analyst could in principle compute two separate p-values, one on each distinct dataset. Our methods provide an alternative to other existing meta-analysis methods [Heard and Rubin-Delanchy, 2018] that can be particularly powerful when one dataset (the primary dataset) is more informative than the secondary dataset. (We provide theoretical and empirical justification in Sections 3.3 and 7.3.)

*Department of Statistics, University of Chicago. E-mail: ignat@uchicago.edu.
†Department of Statistics and Actuarial Science, University of Waterloo. E-mail: wang@uwaterloo.ca.
‡Departments of Statistics and Machine Learning, Carnegie Mellon University. E-mail: aramdas@cmu.edu.

As a second contribution, our methods provide a new perspective on multiple testing with *data-driven hypothesis weights*. Weighted multiple testing procedures provide a flexible and convenient way of differentially prioritizing hypotheses by assigning a weight to each hypothesis and prioritizing hypotheses with large weights [Benjamini and Hochberg, 1997, Genovese et al., 2006, Blanchard and Roquain, 2008, Ramdas et al., 2019]. If the weight assignment is informative and correctly prioritizes alternatives, then weighted multiple testing procedures can lead to substantial power gains compared to unweighted procedures. Weighting methods have traditionally come with two requirements: first, the weights need to be deterministic, that is, they should not depend on the data used to compute the p-values, and second, they need to average to 1. Intuitively, the first requirement implies that the weights can only be a priori "guesses," and the second requirement enforces a constrained size budget to be split across hypotheses. A nascent literature including e.g., Westfall et al. [2004], Finos and Salmaso [2007], Roeder and Wasserman [2009], Ignatiadis et al. [2016], Durand [2019], Ignatiadis and Huber [2021], has dispensed with the first requirement: it is possible to construct data-driven weights and p-values based on the same dataset. In this paper, we demonstrate (for the first time, to our knowledge) that it is also simultaneously possible to dispense with the fixed weight budget requirement.

The key insight for our contributions to both meta-analysis and data-driven hypothesis weighting is the following: independent e-values can be directly used as weights for p-values in all standard multiple testing procedures, without needing to normalize them in any way. This can lead to huge increases in power relative to standard weighted procedures.

2 Multiple testing background

2.1 Terminology and notation

We first describe the basic setting. Let $H_1, \dots, H_K$ be K hypotheses, and write $\mathcal{K} = \{1, \dots, K\}$. Let the true (unknown) data-generating probability measure be denoted by $\mathbb{P}$. For each $k \in \mathcal{K}$, it is useful to think of hypothesis H_k as implicitly defining a set of joint probability measures, and H_k is called a true null hypothesis if $\mathbb{P} \in H_k$. A *p-value* P for a hypothesis H is a random variable that satisfies $Q(P \leq t) \leq t$ for all $t \in [0,1]$ and all $Q \in H$. In other words, a p-value is stochastically larger than $\mathrm{U}(0,1)$. An *e-value* E for a hypothesis H is a $[0,\infty]$-valued random variable satisfying $\mathbb{E}^Q(E) \leq 1$ for all $Q \in H$. Let $\mathcal{N} \subseteq \mathcal{K}$ be the (unknown to the decision maker) index set of true null hypotheses, $K_0 := |\mathcal{N}|$ the number of true null hypotheses, and $\pi_0 := K_0/K$ the proportion of true null hypotheses.

Two settings of testing multiple hypotheses were considered by Wang and Ramdas [2022]. In the first setting, for each $k \in \mathcal{K}$, P_k is a p-value for H_k. In the second setting, for each $k \in \mathcal{K}$, E_k is an e-value for H_k. In this paper we will consider the setting where both P_k and E_k are available for each H_k. Since we are testing whether $\mathbb{P} \in H_k$ for each k, we will only use the following (obvious) condition: if $k \in \mathcal{N}$, then $\mathbb{P}(P_k \leq t) \leq t$ for all $t \in [0,1]$ and $\mathbb{E}^{\mathbb{P}}(E_k) \leq 1$. There are no restrictions on P_k and E_k if $k \notin \mathcal{N}$. We will omit $\mathbb{P}$ in the statements (by simply calling them p-values and e-values) and the expectations. The terms p-values/e-values refer to both the random variables and their realized values (these should be clear from the context).

Now let $\mathcal{D}$ be a testing procedure, that is, a Borel mapping that produces a subset of $\mathcal{K}$ representing the indices of rejected hypotheses based on p-values (we write p-$\mathcal{D}$ to denote a procedure $\mathcal{D}$ that is based only on p-values), e-values (e-$\mathcal{D}$), or a combination of both as the input. The rejected hypotheses by $\mathcal{D}$ are called discoveries. We write $F_{\mathcal{D}} := |\mathcal{D} \cap \mathcal{N}|$ as the number of true null hypotheses that are rejected (i.e., false discoveries), and $R_{\mathcal{D}} := |\mathcal{D}|$ as the total number of discoveries. We are interested in controlling generalized type-I errors that are defined as expectations of the form $\mathbb{E}\{G(F_{\mathcal{D}}, R_{\mathcal{D}})\}$, where $G : \mathbb{N}_{\geq 0} \times \mathbb{N}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a fixed mapping.

One choice of particular interest is the choice $G(f,r) = f/r$, with the convention $0/0 = 0$. Then $G(F_{\mathcal{D}}, R_{\mathcal{D}}) = F_{\mathcal{D}}/R_{\mathcal{D}}$ is called the false discovery proportion, which is the ratio of the number

of false discoveries to that of all claimed discoveries. Benjamini and Hochberg [1995] proposed to control the false discovery rate, which is the expected value of the false discovery proportion, that is, $\mathrm{FDR}_{\mathcal{D}} := \mathbb{E}(F_{\mathcal{D}}/R_{\mathcal{D}})$. Further important generalized type-I errors are given by the choices $G(f,r) = f$ and $G(f,r) = \mathbb{1}(f \geq 1)$. These yield the per-family error rate of a procedure $\mathcal{D}$ which is defined as $\mathrm{PFER}_{\mathcal{D}} := \mathbb{E}(F_{\mathcal{D}})$, as well as the family-wise error rate, defined as $\mathrm{FWER}_{\mathcal{D}} := \mathbb{P}(F_{\mathcal{D}} \geq 1)$. The family-wise error rate is particularly relevant for testing the global null, and is identical to the false discovery rate if all hypotheses are true nulls.

We next turn to discuss the dependence structure among p-values. A common, albeit strong assumption that appears in the literature, e.g., in Liang and Nettleton [2012], is the following:

Definition 2.1 (P-Independence)**.** A vector $(P_k)_{k\in\mathcal{K}}$ of p-values satisfies the p-independence property if: (i) the null p-values $(P_k)_{k\in\mathcal{N}}$ are mutually independent, and (ii) the null p-values $(P_k)_{k\in\mathcal{N}}$ are independent of the non-null p-values $(P_k)_{k\notin\mathcal{N}}$.

To relax the above assumption, we rely on the notion of positive regression dependence on a subset in Finner et al. [2009, Section 4] and Barber and Ramdas [2017] which is slightly weaker than the original one used in Benjamini and Yekutieli [2001]. A set $A \subseteq \mathbb{R}^K$ is said to be *increasing* if $x \in A$ implies $y \in A$ for all $y \geq x$. The term "increasing" is in the non-strict sense, and inequalities should be interpreted component-wise when applied to vectors.

Definition 2.2 (Positive regression dependence on a subset)**.** A vector $(P_k)_{k\in\mathcal{K}}$ of p-values satisfies positive regression dependence on a subset if for any null index $k \in \mathcal{N}$ and increasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}\{(P_\ell)_{\ell\in\mathcal{K}} \in A \mid P_k \leq x\}$ is increasing on $[0,1]$.

A caveat of Definition 2.2 is that it enforces certain positive dependence between the nulls and non-nulls. To address this concern, Su [2018] proposed the following more general notion of dependence.

Definition 2.3 (Positive regression dependence within nulls)**.** A vector $(P_k)_{k\in\mathcal{K}}$ of p-values satisfies positive regression dependence within nulls if the subvector of null p-values, $(P_k)_{k\in\mathcal{N}}$, is positive regression dependent on a subset.

2.2 Unweighted and weighted multiple testing procedures

We now describe a few canonical procedures that control the generalized type-I errors introduced above. We start by describing the p-BH and e-BH procedures. These procedures use p-values, respectively e-values, and seek to control the false discovery rate at the target level α.

Definition 2.4 (p-BH procedure (Benjamini and Hochberg, 1995))**.** For $k \in \mathcal{K}$, let $P_{(k)}$ be the k-th order statistic of the p-values $P_1, \ldots, P_K$, from the smallest to the largest. The p-BH procedure rejects all hypotheses with the smallest k_p^* p-values, where

$$k_p^* := \max\left\{k \in \mathcal{K} : \frac{KP_{(k)}}{k} \leq \alpha\right\}, \tag{1}$$

with the convention $\max(\varnothing) = 0$.

Definition 2.5 (e-BH procedure (Wang and Ramdas, 2022))**.** For $k \in \mathcal{K}$, let $E_{[k]}$ be the k-th order statistic of the e-values $E_1, \ldots, E_K$, from the largest to the smallest. The e-BH procedure rejects all hypotheses with the largest k_e^* e-values, where

$$k_e^* := \max\left\{k \in \mathcal{K} : \frac{kE_{[k]}}{K} \geq \frac{1}{\alpha}\right\}. \tag{2}$$

An equivalent way to describe the e-BH procedure is to apply the p-BH procedure to $(E_1^{-1}, \ldots, E_K^{-1})$.

The p-BH procedure at level α has false discovery rate at most (i) $\pi_0\alpha$ when the p-values satisfy p-independence or positive regression dependence on a subset [Benjamini and Hochberg, 1995,

Benjamini and Yekutieli, 2001], (ii) $\pi_0\alpha \log\{e/(\pi_0\alpha)\}$ when the p-values satisfy positive regression dependence within nulls [Su, 2018], and (iii) $\ell_K \pi_0 \alpha$, where $\ell_K := \sum_{k=1}^K k^{-1} \approx \log K$, under arbitrary dependence [Benjamini and Yekutieli, 2001]. As for the e-BH procedure, Wang and Ramdas [2022] showed a surprising property that the base e-BH procedure controls the false discovery rate at α even under unknown arbitrary dependence between the e-values.

A procedure closely related to p-BH is the p-Simes procedure. This is not a multiple testing procedure per se, but instead, it is a test of the global null hypothesis $H := \bigcap_{k=1}^K H_k$.

Definition 2.6 (p-Simes procedure (Simes, 1986))**.** The p-Simes procedure rejects the global null $\bigcap_{k=1}^K H_k$ when the p-BH procedure applied to $(P_k)_{k\in\mathcal{K}}$ makes at least one discovery.

The p-Simes procedure has type-I error at most α when the p-values are positive regression dependent within nulls.

We next present the p-Bonferroni procedure to control the per-family error rate and the family-wise error rate.

Definition 2.7 (p-Bonferroni procedure (Bonferroni, 1935))**.** Let $P_1, \dots, P_K$ be the p-values. The p-Bonferroni procedure rejects all hypotheses with $P_k \le \alpha/K$.

The p-Bonferroni procedure controls the per-family error rate and the family-wise error rate at level α under arbitrary p-value dependence. The following procedure (p-Hochberg) controls the family-wise error rate under a stronger dependence assumption, namely, positive regression dependence within nulls, and is more powerful than p-Bonferroni.

Definition 2.8 (p-Hochberg procedure (Hochberg, 1988))**.** For $k \in \mathcal{K}$, let $P_{(k)}$ be the k-th order statistic of the p-values $P_1, \dots, P_K$, from the smallest to the largest. The p-Hochberg procedure rejects all hypotheses with the smallest k_h^* p-values, where

$$k_h^* := \max\left\{k \in \mathcal{K} : P_{(k)} \le \frac{\alpha}{K-k+1}\right\}.$$

In Supplement S2 we also discuss the procedures of Holm [1979] and Hommel [1988].

Many p-value based multiple testing procedures may be applied alongside a vector of weights. Two examples are weighted p-BH and weighted p-Bonferroni [Genovese et al., 2006].

Definition 2.9 (Weighted p-BH and weighted p-Bonferroni procedures)**.** Let $P_1, \dots, P_K$ be the p-values and let $(w_1, \dots, w_K) \in [0, \infty)^K$ be a pre-specified vector of weights. The weighted p-BH procedure (resp. p-Bonferroni procedure) is obtained by applying the p-BH (resp. p-Bonferroni procedure) to $(P_1/w_1, \dots, P_K/w_K)$.

For generalized type-I error control, classical thinking imposes the fixed weight budget requirement that the weights are normalized and average to 1, that is, $\sum_{k=1}^K w_k = K$. In that case, the weighted p-BH procedure controls the false discovery rate when the p-values are positive regression dependent on a subset [Blanchard and Roquain, 2008, Ramdas et al., 2019], and the weighted p-Bonferroni procedure controls the per-family error rate and the family-wise error rate under arbitrary dependence of the p-values [Genovese et al., 2006]. Later, we will see if the weights are obtained from e-values independent of the p-values, then normalization is not needed, and this can improve power substantially.

3 Combining a p-value and an e-value

3.1 Admissible p-value/e-value combiners

One of the main objectives of the paper is to design and understand procedures when both p-values and e-values are available. For this purpose, we first look at the single-hypothesis setting, in which case we drop the subscripts and use P for a p-value and E for an e-value.

We briefly review calibration between a p-value and an e-value as developed previously by Shafer et al. [2011] and Shafer and Vovk [2019, Chapter 11.5], amongst other sources. Denote by $\overline{\mathbb{R}}_+ = [0, \infty]$. First, an e-value E can be converted to a p-value $P = (1/E) \wedge 1$ (its validity follows from Markov's inequality). Further, the function $f : e \mapsto (1/e) \wedge 1$ is the unique admissible e/p calibrator [Vovk and Wang, 2021, Proposition 2.2].

A p-value P can also be converted to an e-value, but there are many admissible choices. One example is to set $E = P^{-1/2} - 1$. More generally, we speak of p/e calibrators. Small p-values correspond to large e-values, which represent stronger evidence against a null hypothesis. A p/e calibrator is a decreasing function $h : [0, 1] \to \overline{\mathbb{R}}_+$ satisfying $\int_0^1 h(u) \, \mathrm{d}u \le 1$. Then $h(P)$ is an e-value for any p-value P. Vovk and Wang [2021, Proposition 2.1] show that the set $\mathcal{C}^{\mathrm{p/e}}$ of all admissible p/e calibrators is

$$\mathcal{C}^{\mathrm{p/e}} = \left\{ h : [0, 1] \to \overline{\mathbb{R}}_+ \text{ decreasing \& upper semicontinuous } \mid h(0) = \infty, \int h(u) \, \mathrm{d}u = 1 \right\}.$$

In the above statements, admissibility of a calibrator (or a combiner below) means that it cannot be improved strictly, where improvement means obtaining a larger e-value or a smaller p-value.

Combining several p-values or e-values to form a new p-value or e-value is the main topic of Vovk and Wang [2020, 2021] and Vovk et al. [2022]. For the objective of this paper, we need to combine a p-value P and an e-value E, first in a single-hypothesis testing problem. We consider four cases. (i) If P and E are independent, how should we combine them to form an e-value? (ii) If P and E are independent, how should we combine them to form a p-value? (iii) If P and E are arbitrarily dependent, how should we combine them to form an e-value? (iv) If P and E are arbitrarily dependent, how should we combine them to form a p-value?

We use the following terminology, similar to Vovk and Wang [2021]. A function $f : [0, 1] \times \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+$ is called an i-pe/e combiner if $f(P, E)$ is an e-value for any independent p-value P and e-value E, and $(p, e) \mapsto f(p, e)$ is decreasing in p and increasing in e. Similarly, we define i-pe/p, pe/p, and pe/e combiners, where i indicates independence, and p and e are self-explanatory. If the output is a p-value, the combiner is increasing in p and decreasing in e.

We provide four natural answers to the above four questions, some relying on an admissible calibrator $h \in \mathcal{C}^{\mathrm{p/e}}$. (i) Return $h(P)E$ by using the function $\Pi_h(p, e) := h(p)e$. The convention here is $0 \times \infty = \infty$. (ii) Return P/E, capped at 1, by using the function $Q(p, e) := (p/e) \wedge 1$. (iii) Return $\lambda h(P) + (1 - \lambda)E$ by using the function $M_h^\lambda(p, e) := \lambda h(p) + (1 - \lambda)e$ for some $\lambda \in (0, 1)$. (iv) Return $2 \min(P, 1/E)$, capped at 1, by using the function $B(p, e) := \{2(p \wedge e^{-1})\} \wedge 1$.

The notation chosen for these functions is due to the initials of (i) product (but we avoid P which is reserved for p-values); (ii) quotient; (iii) mean; (iv) Bonferroni correction.

Π_h and M_h^λ depend on h whereas Q and B do not. For the function M_h^λ, it may be convenient to choose $\lambda = 1/2$, so that $M_h^\lambda(P, E)$ is the arithmetic average of two e-values $h(P)$ and E. As shown by Vovk and Wang [2021, Proposition 3.1], the arithmetic average essentially dominates, in a natural sense, all symmetric e-merging function. In our context, $\lambda = 1/2$ has no special role, since the positions of $h(P)$ and E are not symmetric.

Theorem 3.1. *For $h \in \mathcal{C}^{\mathrm{p/e}}$ and $\lambda \in (0, 1)$, Π_h is an admissible i-pe/e combiner, Q is an admissible i-pe/p combiner, M_h^λ is an admissible pe/e combiner, B is an admissible pe/p combiner.*

The proof can be found in Supplement S1.1. For the remainder of the paper, we pay particular attention to the i-pe/p combiner Q that forms a p-value based on independent P and E. We use the term Q-combiner to refer to both the mapping $(P, E) \mapsto Q(P, E) = (P/E) \wedge 1$ as well as the resulting p-value $Q(P, E)$. The Q-combiner typically leads to more powerful procedures compared to the other combiners and provides the foundation for our insight that e-values can act as unnormalized weights in multiple testing (see next sections). The Π_h i-pe/e combiner is also of interest, and we develop results for Π_h in the context of multiple testing in Supplement S3.

Remark 3.2. One consequence of Theorem 3.1 is as follows. Consider an e-value E and generate an independent uniform variable $U \sim \mathrm{U}(0,1)$. Then, $P' := Q(U,E)$ is a valid p-value that satisfies $\mathbb{P}\{Q(U,E) \le f(E)\} = 1$ and $\mathbb{P}\{Q(U,E) < f(E)\} > 0$, where $f : e \mapsto (1/e) \wedge 1$ is the unique admissible e/p calibrator. Hence, f is dominated by a randomized e/p calibrator. Although $Q(U,E)$ may not be practical in general due to external randomization, it becomes practical when applied as $Q(P,E)$ to a p-value P (computed from data) independent of E.

3.2 Q-combiner as a general-purpose method for meta-analysis from two studies

As mentioned above, for the remainder of the paper we consider procedures that build on the Q-combiner $(P,E) \mapsto Q(P,E) = (P/E) \wedge 1$. To start, we argue that the Q-combiner is a useful general-purpose method for meta-analysis from two independent datasets. The Q-combiner is immediately applicable when the researcher summarizes the first dataset as a single p-value, and the second dataset as a single e-value. Such a situation could occur when the second dataset is collected in such a way, e.g., with optional stopping and continuation, that inference is more natural with e-values; see Ramdas et al. [2022] for a survey of e-values and the inferential problems they solve. It could also be the case that one dataset comprises of a large sample size, allowing for asymptotic approximations to compute p-values, while the second dataset is smaller and may require finite-sample inference methods, e.g., universal inference [Wasserman et al., 2020], that lead to e-values.

Our claim, however, is stronger: the Q-combiner is also useful when the above data constraints are not in place and the researcher can in principle compute both a p-value P' and an e-value E on the second dataset, both of which are independent of the p-value P computed on the first dataset. In that case, the researcher could apply a p-value combination method based on P and P', e.g., Fisher's combination $P_{\mathrm{F}} := 1 - \chi_4\{-2\log(PP')\}$, where χ_4 is the chi-square distribution with 4 degrees of freedom. However, the researcher may still prefer to proceed with the Q-combiner $Q(P,E)$. We suggest the following rule of thumb.

The Fisher combination is preferable to the Q-combiner under dataset exchangeability: Suppose that the analyst considers the two datasets as a priori exchangeable. In that case, it may be undesirable to use an asymmetric combination rule such as $Q(P,E)$, and Fisher's combination P_{F} is preferable on conceptual grounds. If the two datasets are also exchangeable in terms of their statistical properties (i.e., they have similar power), then P_{F} will typically have higher power than $Q(P,E)$.

The Q-combiner is preferable to the Fisher combination for imbalanced datasets: When one dataset (the "primary" dataset) is substantially more well-powered (larger anticipated signal or sample size) than the secondary dataset, and the investigator knows which dataset is more well-powered, then the Q-combiner can often outperform Fisher's combination test in terms of power. A proviso is that the p-value is computed on the primary (more well-powered) dataset and the e-value on the secondary dataset.

In the next section, we provide theoretical and numerical evidence for the rule of thumb put forth in the preceding paragraph in a stylized example. We also provide further numerical evidence in the simulations of Section 7.3.

3.3 A stylized example: using two samples for the one-sided z-test via the Q-combiner

As a stylized example, suppose we have access to two independent samples of iid data points, $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$, both from a distribution $\mathbb{P}$, where $n \ge m \ge 1$. We seek to test $H_0 : \mathbb{P} = \mathrm{N}(0,1)$ against $H_1 : \mathbb{P} = \mathrm{N}(\delta,1)$, where $\delta > 0$ is known. The optimal p-value based on X is $P_X := 1 - \Phi(T_X)$, where Φ is the standard normal distribution function and $T_X := \sum_{i=1}^m X_i/\sqrt{m}$. Analogously we may compute p-values P_Y based on Y, as well as P_Z, where $Z = (X,Y)$ is the full dataset. The optimal e-value E_X based on X is the likelihood ratio of $\mathrm{N}(\delta,1)^m$ over $\mathrm{N}(0,1)^m$.

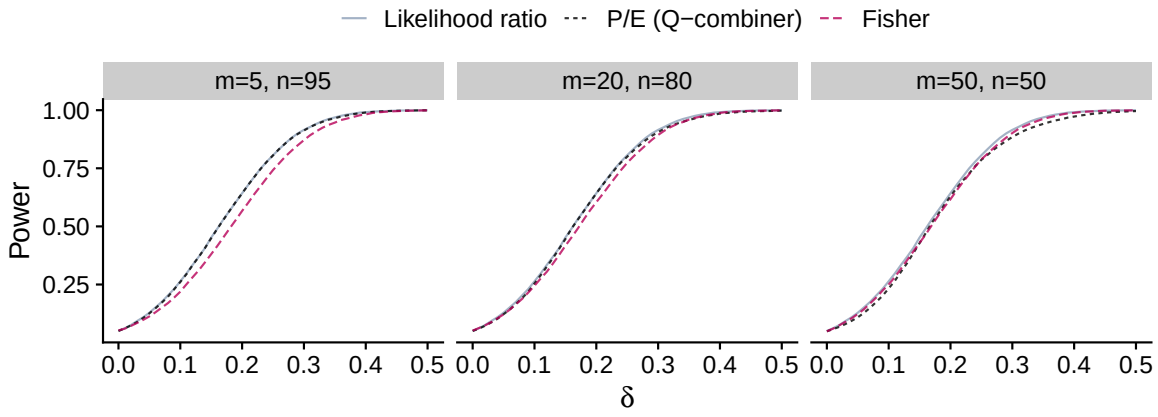

Figure 1: Simulation study for a meta-analysis combining two samples: We compare the likelihood ratio test, the Q-combiner, and the Fisher combination test, plotting power against signal strength δ. The panels correspond to different choices of the two sample sizes m and n. The Q-combiner is visibly more powerful on the left (matching the likelihood ratio), and Fisher's combination is marginally better on the right.

By the Neyman–Pearson lemma, the p-value P_Z leads to the most powerful test. We seek to compare P_Z against the Q-combiner $P_{\mathrm{E}} := Q(P_Y, E_X)$ by considering the hypothesis tests that reject H_0 when $P_Z \leq \alpha$, or when $P_{\mathrm{E}} \leq \alpha$, for $\alpha > 0$. We assume $m = \lfloor \theta^2 n \rfloor$ for some $\theta \in (0, 1]$ which measures the relative size of the two datasets.

In Supplement S4, we derive Pitman's asymptotic relative efficiency (Van der Vaart, 1998, Section 14.3; DasGupta, 2008, Section 22.1) between the two methods, which is the asymptotic ratio of the required sample size from P_Z to reach a fixed power, to that from P_{E}, as $\delta \downarrow 0$. We prove that the asymptotic relative efficiency converges to 1 in two different settings: as $\alpha \downarrow 0$, that is, when the type-I error is very stringent, and as $\theta \downarrow 0$, that is, when Y is substantially more well-powered than X. Our results can also be used to numerically compute the asymptotic relative efficiency for any choice of θ, α, and desired power, e.g., the asymptotic relative efficiency is (up to numerical rounding) equal to 0.989 when $\theta = 0.5$, $\alpha = 0.05$, and we seek a power of 50%.

We also conduct a small simulation study comparing (i) P_Z, (ii) P_{E}, and, (iii) the Fisher p-value $P_{\mathrm{F}} := 1 - \chi_4\{-2\log(P_Y P_X)\}$. Simulation results are reported in Fig. 1 based on the average of 10,000 runs. We take $\alpha = 0.05$, $m + n = 100$ and let the ratio m/n and $\delta > 0$ vary. We observe the following: if m/n is small (first two panels), meaning that X is less informative than Y, then the Q-combiner has almost the same power as the full likelihood ratio method, and both outperform the Fisher method. When $m = n$ (third panel), the Fisher test has more power than the Q-combiner, and both have (slightly) less power than the likelihood ratio test.

4 Using e-values as weights in multiple testing with p-values

4.1 General remarks

In this section, we move back to multiple testing by considering the setting where each hypothesis is associated with a p-value and an e-value. The generalized type-I error of testing procedures depends on the dependence amongst the e-values, the dependence amongst the p-values, and the dependence between the p-values and e-values. Throughout this section we make the following assumption:

Assumption 4.1. P_k is independent of E_k for all $k \in \mathcal{N}$.

Given any procedure p-$\mathcal{D}$ that is based on p-values, we can extend it to an e-weighted procedure that we call ep-$\mathcal{D}$ and which generalizes the concept of weighting for multiple testing.

Definition 4.2 (e-weighted p-value procedure (ep-$\mathcal{D}$))**.** Let p-$\mathcal{D}$ be a multiple testing procedure based on p-values. Given p-values $(P_k)_{k\in\mathcal{K}}$ and e-values $(E_k)_{k\in\mathcal{K}}$, we define the e-weighted p-value procedure ep-$\mathcal{D}$ which proceeds as follows: for $k \in \mathcal{K}$, compute the Q-combiner $P_k^* := Q(P_k, E_k) = (P_k/E_k) \wedge 1$, and then supply $(P_k^*)_{k\in\mathcal{K}}$ to p-$\mathcal{D}$.

Concretely, we define the (i) ep-BH, (ii) ep-Simes, (iii) ep-Bonferroni, resp. (iv) ep-Hochberg procedure by plugging in the (i) p-BH (Definition 2.4), (ii) p-Simes (Definition 2.6), (iii) p-Bonferroni (Definition 2.7), resp. (iv) p-Hochberg (Definition 2.8) procedure into Definition 4.2.

In view of Definition 2.9, ep-BH and ep-Bonferroni may be interpreted in two ways: (i) they are p-value based procedures applied to the p-value vector $(P_k^*)_{k\in\mathcal{K}}$, and (ii) they are weighted p-value based procedures with p-value vector $(P_k)_{k\in\mathcal{K}}$ and weight vector $(E_k)_{k\in\mathcal{K}}$. Both perspectives are useful in deriving guarantees on the control of generalized type-I error rates.

4.2 E-weighted p-value procedures as p-value procedures

We first present a general result under Assumption 4.1. Recall the generalized type-1 error mapping G from Section 2.1 whose expectation captures error metrics like the false discovery rate, per-family error rate, and the family-wise error rate, amongst others.

Theorem 4.3. *Let p-$\mathcal{D}$ be a p-value procedure such that $\mathbb{E}\{G(F_{p\text{-}\mathcal{D}}, R_{p\text{-}\mathcal{D}})\} \le \alpha'$ for any p-value vector that may be arbitrarily dependent, where $\alpha' > 0$. Suppose further that Assumption 4.1 holds. Then, the ep-$\mathcal{D}$ procedure applied to $(P_k)_{k\in\mathcal{K}}, (E_k)_{k\in\mathcal{K}}$ also satisfies $\mathbb{E}\{G(F_{ep\text{-}\mathcal{D}}, R_{ep\text{-}\mathcal{D}})\} \le \alpha'$. In particular, (i) the ep-BH procedure has false discovery rate at most $\ell_K \pi_0 \alpha$, where $\ell_K := \sum_{k=1}^K k^{-1}$, and (ii) the ep-Bonferroni procedure has per-family error rate and family-wise error rate at most α.*

Proof. By Theorem 3.1, P_k^* is a valid p-value under Assumption 4.1 for all $k \in \mathcal{N}$. Hence $(P_k^*)_{k\in\mathcal{K}}$ is a valid p-value vector (that may be arbitrarily dependent). □

The above argument can easily be generalized. For example, if Assumption 4.1 holds, and $(P_k^*)_{k\in\mathcal{K}}$ satisfies positive regression dependence on a subset, then ep-BH has false discovery rate at most $\pi_0\alpha$. Analogously, if Assumption 4.1 holds and $(P_k^*)_{k\in\mathcal{K}}$ is positive regression dependent within nulls, then ep-Hochberg has family-wise error rate at most α and so forth. The assumption that $(P_k^*)_{k\in\mathcal{K}}$ is positive regression dependent on a subset (or within nulls), however, may be difficult to interpret. Thus we prefer to directly impose assumptions on $(P_k)_{k\in\mathcal{K}}$, $(E_k)_{k\in\mathcal{K}}$, as well as the cross-dependence between $(P_k)_{k\in\mathcal{K}}$ and $(E_k)_{k\in\mathcal{K}}$.

We provide an example of the general approach by considering the following independence assumption. In the next subsection, we provide more elaborate results by turning to the perspective that ep-$\mathcal{D}$ procedures are weighted procedures.

Assumption 4.4. (i) The null (p-value, e-value) pairs $\{(P_k, E_k)\}_{k\in\mathcal{N}}$ are mutually independent, and (ii) $\{(P_k, E_k)\}_{k\in\mathcal{N}}$ is independent of $\{(P_k, E_k)\}_{k\notin\mathcal{N}}$.

Theorem 4.5. *Let p-$\mathcal{D}$ be a p-value procedure such that $\mathbb{E}\{G(F_{p\text{-}\mathcal{D}}, R_{p\text{-}\mathcal{D}})\} \le \alpha'$ for any p-value vector that satisfies p-independence (Definition 2.1), where $\alpha' > 0$. Suppose further that Assumptions 4.1 and 4.4 hold. Then, the ep-$\mathcal{D}$ procedure applied to $(P_k)_{k\in\mathcal{K}}, (E_k)_{k\in\mathcal{K}}$ also satisfies $\mathbb{E}\{G(F_{ep\text{-}\mathcal{D}}, R_{ep\text{-}\mathcal{D}})\} \le \alpha'$. In particular, (i) the ep-BH procedure has false discovery rate at most $\pi_0\alpha$, (ii) the ep-Hochberg procedure has family-wise error rate at most α.*

Proof. Assumptions 4.1 and 4.4 imply that $(P_k^*)_{k\in\mathcal{K}}$ satisfies p-independence (Definition 2.1). □

4.3 E-weighted p-value procedures as weighted p-value procedures

We now turn to the second perspective of e-weighted procedures: we interpret the e-values as weights for the p-values. Intuitively, if $E_k > 1$, then there is evidence against H_k being a null, and we have $P_k/E_k < P_k$ (assuming $P_k \neq 0$), that is, the weight strengthens the signal of P_k. Conversely, if $E_k < 1$, then there is no evidence against H_k being a null, and we have $P_k/E_k > P_k$. The above interpretation of e-values as weights is quite natural, and the perspective is useful in deriving guarantees for, e.g., ep-BH, that may be challenging to prove otherwise: below we prove that ep-BH controls the false discovery rate under the assumption that $(P_k)_{k\in\mathcal{K}}$ is positive regression dependent on a subset along with the following strengthening of Assumption 4.1.

Assumption 4.6. $(P_k)_{k\in\mathcal{K}}$ is independent of $(E_k)_{k\in\mathcal{K}}$.

Positive regression dependence on a subset of $(P_k)_{k\in\mathcal{K}}$, together with Assumption 4.6, does not imply positive regression dependence on a subset of $(P_k^*)_{k\in\mathcal{K}}$, and hence some arguments are needed to establish control of the false discovery rate by ep-BH. The following result integrates over the randomness in the weights (e-values). In contrast, weighted p-BH with normalized weights controls the false discovery rate conditionally on all the weights.

Theorem 4.7. *Suppose that Assumption 4.6 holds and that $(P_k)_{k\in\mathcal{K}}$ is positive regression dependent on a subset (Definition 2.2). Then, the ep-BH procedure has false discovery rate at most $\pi_0\alpha$.*

Proof. Let ep-$\mathcal{D}$ be the ep-BH procedure at level α. Since $(E_k)_{k\in\mathcal{K}}$ is independent of $(P_k)_{k\in\mathcal{K}}$, conditional on $(E_k)_{k\in\mathcal{K}}$, the ep-BH procedure becomes a weighted p-BH procedure with weight vector $(E_k)_{k\in\mathcal{K}}$ applied to the p-values $(P_k)_{k\in\mathcal{K}}$ that are positive regression dependent on a subset. Using well-known existing results on the false discovery rate of the weighted p-BH procedure (e.g., Ramdas et al., 2019, Theorem 1), we get

$$\mathbb{E}\left\{\frac{F_{\text{ep-}\mathcal{D}}}{R_{\text{ep-}\mathcal{D}}} \,\middle|\, (E_k)_{k\in\mathcal{K}}\right\} \leq \frac{1}{K}\sum_{k\in\mathcal{N}} E_k\alpha.$$

Hence, by iterated expectation, $\text{FDR}_{\text{ep-}\mathcal{D}} \leq \mathbb{E}(\sum_{k\in\mathcal{N}} E_k\alpha/K) \leq \pi_0\alpha$. □

Perhaps surprisingly, the above result does not require any assumption whatsoever about the dependence within $(E_k)_{k\in\mathcal{K}}$. In the case of p-values that are positive regression dependent within nulls, it is natural to posit the following dependence assumption on $(P_k)_{k\in\mathcal{K}}$ and $(E_k)_{k\in\mathcal{K}}$ (that is intermediate in strength compared to Assumptions 4.1 and 4.6).

Assumption 4.8. $(P_k)_{k\in\mathcal{N}}$ is independent of $(E_k)_{k\in\mathcal{N}}$.

Theorem 4.9. *Suppose that Assumption 4.8 holds, and that $(P_k)_{k\in\mathcal{K}}$ is positive regression dependent within nulls (Definition 2.3). Then, (i) the ep-Simes procedure has type-I error of at most α under the global null hypothesis, (ii) the ep-BH procedure has false discovery rate at most $\pi_0\alpha\log\{e/(\pi_0\alpha)\}$, and (iii) the ep-Hochberg procedure has family-wise error rate at most α.*

Proof. The result for ep-Simes follows from Theorem 4.7: under the global null hypothesis, it holds that $\mathcal{N} = \mathcal{K}$, and so the assumptions of Theorems 4.7 and 4.9 are identical. Hence by definition of the Simes procedure (and noting that FWER = FDR under the global null):

$$\mathbb{P}\left(\text{ep-Simes rejects } \textstyle\bigcap_{k=1}^K H_k\right) = \mathbb{P}(\text{ep-BH rejects at least one of } H_k) = \text{FDR}_{\text{ep-BH}} \leq \alpha.$$

The result on ep-BH follows from the false discovery rate linking theorem [Su, 2018, Theorem 1] which converts bounds on the false discovery rate of p-BH applied on the null p-values only to a bound on the false discovery rate of p-BH applied to all p-values. The false discovery rate linking theorem is also applicable to ep-BH once we interpret it as p-BH acting on $(P_k^*)_{k\in\mathcal{K}}$. Hence it suffices to bound

the false discovery rate of ep-BH applied on the null hypotheses only, and such a bound follows from Theorem 4.7.

The result for ep-Hochberg follows from the result for ep-Simes, since the p-Hochberg procedure [Hochberg, 1988] is a shortcut for closed testing [Marcus et al., 1976] based on p-Simes (and so ep-Hochberg is a shortcut for closed testing based on ep-Simes). □

4.4 Null proportion adaptivity: the e-weighted Storey procedure (ep-Storey)

As we explained above, weighted multiple testing procedures typically require normalized weights, that is, weights that satisfy $\sum_{k=1}^{K} w_k = K$. This constraint represents a fixed weight budget to be allocated across hypotheses. It is possible, however, to increase the budget in a data-driven way by adapting to the proportion of null hypotheses. For example, in the case of uniform weights (i.e., for unweighted multiple testing), Storey, Taylor, and Siegmund [2004] proposed to estimate the proportion of null hypotheses $\pi_0 = K_0/K$ by:

$$\widehat{\pi}_0 := \frac{1 + \sum_{k=1}^{K} \mathbb{1}(P_k > \tau)}{K(1-\tau)}, \tag{3}$$

for fixed $\tau \in (0,1)$, and then to apply the p-BH procedure with p-values P_k and weights $w_k := \mathbb{1}(P_k \leq \tau)/\widehat{\pi}_0$. Since hypotheses with $P_k > \tau$ would be unlikely to be rejected, the procedure of Storey et al. [2004] increases the weight budget (when $\widehat{\pi}_0 < 1$) from K to $K/\widehat{\pi}_0$.

The case of null-proportion adaptive procedures was addressed by Habiger [2017] and Ramdas et al. [2019] for arbitrary normalized weights, and by Li and Barber [2019] for a specific choice of data-driven weights. Here we propose *ep-Storey*, a null proportion adaptive version of ep-BH, which proceeds as follows: compute $\widehat{\pi}_0$ as in (3) with p-values P_k and then apply the weighted p-BH procedure (Definition 2.9) at level α with p-values P_k and weights $w_k := \mathbb{1}(P_k \leq \tau)E_k/\widehat{\pi}_0$.

Theorem 4.10. *Suppose that Assumption 4.6 holds, and that $(P_k)_{k\in\mathcal{K}}$ satisfies p-independence (Definition 2.1). Then the ep-Storey procedure has false discovery rate at most α.*

The proof (Supplement S1.2) proceeds as the proof of Theorem 4.7 by arguing conditionally on $(E_k)_{k\in\mathcal{K}}$. In defining ep-Storey, we abused terminology: ep-Storey is different than the procedure implied by Definition 4.2, that is, applying Storey's procedure to $(P_k^*)_{k\in\mathcal{K}}$ in which case one would estimate π_0 by $\hat{\pi}_0' = \{1 + \sum_{k=1}^{K} \mathbb{1}(P_k > \tau E_k)\}/\{K(1-\tau)\}$ instead of (3) and then weight hypotheses by $\mathbb{1}(P_k \leq \tau E_k)/\hat{\pi}_0'$. The latter procedure controls the false discovery rate under Assumptions 4.1 and 4.4 (by Theorem 4.5), but not necessarily under the assumptions of Theorem 4.10.

One shortcoming of ep-Storey occurs when some e-values are potentially very strong, say $1/E_k \approx \alpha/K$, but the corresponding p-values satisfy $P_k > \tau$. Such hypotheses would be discarded by ep-Storey, but would be rejected by e.g., e-BH that only uses the e-values. Hence it may be advisable to use larger values of τ for ep-Storey.

Remark 4.11. There is a further subtle benefit of ep-BH compared to weighted p-BH with normalized weights related to null proportion adaptivity. In many applications, the proportion of null hypotheses π_0 is close to 1. In such cases, it may not be worthwhile to apply ep-Storey compared to ep-BH, since their power will be comparable ($\alpha \approx \pi_0\alpha$) and the assumptions on the dependence of $(P_k)_{k\in\mathcal{K}}$ are stronger in Theorem 4.10 compared to Theorem 4.7. For weighted p-BH with normalized weights, however, the false discovery rate is controlled at $\alpha \sum_{k\in\mathcal{N}} w_k / \sum_{k=1}^{K} w_k$ [Ramdas et al., 2019]. Hence, for an informative weight assignment that prioritizes alternative hypotheses over null hypotheses, the gains from null proportion adaptivity can be substantial even when $\pi_0 \approx 1$. In other words: the more informative the weights are, the more conservative normalized weighted p-BH becomes in terms of the false discovery rate. ep-BH does not pay such a penalty for informative e-value weights as long as $\mathbb{E}(E_k) = 1$ for all $k \in \mathcal{N}$.

4.5 Robustness to misspecification

So far, we have presented all results under the assumption that we have access to e-values with the property that $\mathbb{E}(E_k) \leq 1$ for all $k \in \mathcal{N}$ and p-values with the property $\mathbb{P}(P_k \leq t) \leq t$ for all $t \in [0,1]$ and all $k \in \mathcal{N}$. Nevertheless, the guarantees are often robust to some deviations from these assumptions. We consider the following possible deviations.

Inflated (anticonservative) e-values or p-values: Suppose the e-values satisfy $\mathbb{E}(E_k) \leq 1+\eta$ for some $\eta > 0$ and all $k \in \mathcal{N}$ or alternatively that the p-values satisfy $\mathbb{P}(P_k \leq t) \leq (1+\eta)t$ for all $t \in [0,1]$ and all $k \in \mathcal{N}$. Then the generalized type-I error bounds for ep-BH, ep-Bonferroni, ep-Hochberg, and ep-Simes stated above hold with α replaced by $(1+\eta)\alpha$. For example, ep-BH controls the false discovery rate at level $(1+\eta)\ell_K \pi_0 \alpha$ in the setting of Theorem 4.3 and at level $(1+\eta)\pi_0\alpha$ in the setting of Theorem 4.7. (The proofs are simple and thus omitted.)

Compound e-values or p-values: Instead of $\mathbb{E}(E_k) \leq 1$ holding for all $k \in \mathcal{N}$, suppose this property holds on average over all k [Wang and Ramdas, 2022, Ren and Barber, 2022], that is,

$$\frac{1}{K_0}\sum_{k\in\mathcal{N}} \mathbb{E}(E_k) \leq 1. \tag{4}$$

Some results extend to this case as well. For example, under (4), ep-Bonferroni controls the family-wise error rate at α in the setting of Theorem 4.3, and ep-BH controls the false discovery rate at α in the setting of Theorem 4.7. Analogously to (4), one may consider compound p-values that satisfy $\sum_{k\in\mathcal{N}} \mathbb{P}(P_k \leq t) \leq K_0 t$ for all $t \in [0,1]$. See Armstrong [2022] for robustness guarantees in that setting.

5 Data-driven weighting with compound e-values

5.1 Simultaneous one-sample t-tests and means of squares as e-values

We now turn to our second contribution: we demonstrate the feasibility and practicality of weighted multiple testing procedures with unnormalized data-driven weights. For simplicity, we restrict attention to ep-BH and ep-Bonferroni. Our agenda will be to construct E_k such that the following properties hold approximately: E_k is independent of P_k for $k \in \mathcal{N}$ and (4) holds.

Throughout this section we study the problem of conducting simultaneous one-sample t-tests based on n observations per hypothesis (and we defer extensions to simultaneous two-sample t-tests to Supplement S5). To be concrete, for the k-th hypothesis we observe

$$Y_{kj} \sim \mathrm{N}(\mu_k, \sigma_k^2), \text{ for } j = 1, \ldots, n, \quad \mu_k \in \mathbb{R}, \quad \sigma_k > 0, \tag{5}$$

and we assume that all Y_{kj}, $1 \leq k \leq K$, $1 \leq j \leq n$ are mutually independent and $n \geq 2$. We seek to test $H_k : \mu_k = 0$ and to do so, we compute p-values based on the standard t-test,

$$\hat{\mu}_k := \frac{1}{n}\sum_{j=1}^{n} Y_{kj}, \quad \hat{\sigma}_k^2 := \frac{1}{n-1}\sum_{j=1}^{n}(Y_{kj} - \hat{\mu}_k)^2, \quad T_k := \frac{\sqrt{n}\hat{\mu}_k}{\hat{\sigma}_k}, \tag{6}$$

and $P_k := 2\{1 - F_{t,n-1}(|T_k|)\}$, where $F_{t,n-1}$ is the t-distribution with $n-1$ degrees of freedom. Multiple testing with simultaneous t-tests has been studied by several authors including Smyth [2004], Westfall et al. [2004], Finos and Salmaso [2007], Bourgon et al. [2010], Du and Zhang [2014], Lu and Stephens [2016], Guo and Romano [2017], Ignatiadis and Huber [2021], Hoff [2022], Ignatiadis and Sen [2023]. Albeit stylized, this problem has provided several new insights and has called attention to differences between single hypothesis testing and multiple testing. We contribute to the literature by demonstrating how unnormalized data-driven weights alongside $(P_k)_{k\in\mathcal{K}}$ can be constructed based on $(Y_{kj})_{k,j}$ while retaining type-I error control guarantees.

The key observation permitting the construction of data-driven weights is the following. Let $S_k^2 := \sum_{j=1}^n Y_{kj}^2/n$ be the mean of squares of the observations for the k-th hypothesis. When $k \in \mathcal{N}$, that is, when $\mu_k = 0$, then S_k^2 is complete and sufficient for σ_k^2. On the other hand, the t-statistic T_k is ancillary for σ_k^2. Hence, by Basu's theorem [Basu, 1955], it holds that T_k (and so P_k) is independent of S_k^2. To summarize:

$$S_k^2 \text{ is independent of } P_k \text{ for } k \in \mathcal{N}. \tag{7}$$

Motivated by the above, Westfall et al. [2004] considered the weights

$$w_k := K S_k^2 \Big/ \sum_{\ell=1}^K S_\ell^2, \tag{8}$$

and proved that weighted p-Bonferroni with p-values P_k and weights w_k controls the family-wise error rate. Similarly, weighted p-BH with the above p-values and weights controls the false discovery rate.

The perspective on e-values as weights in multiple testing along with the robustness guarantee for (4) suggests that instead of w_k we could have used $E_k^* := K_0 S_k^2 / \sum_{\ell \in \mathcal{N}} \sigma_\ell^2$, which satisfy $\sum_{k \in \mathcal{N}} \mathbb{E}(E_k^*) = K_0$ (since $\mathbb{E}(S_k^2) = \sigma_k^2$ for $k \in \mathcal{N}$). As explained in Section 4.5, ep-Bonferroni, resp. ep-BH with P_k and E_k^* control the family-wise error rate, resp. false discovery rate.

A caveat to the above argument is that the compound e-values E_k^* are not computable as they depend on the unknown data generating mechanism through $\sum_{\ell \in \mathcal{N}} \sigma_\ell^2$. Our proposal then is to conservatively estimate $\sum_{\ell \in \mathcal{N}} \sigma_\ell^2$ by $\sum_{\ell=1}^K \hat{\sigma}_\ell^2$, and to construct feasible approximations to E_k^*:

$$E_k := K S_k^2 \Big/ \sum_{\ell=1}^K \hat{\sigma}_\ell^2. \tag{9}$$

Since $\mathbb{E}(S_k^2) = \mu_k^2 + \sigma_k^2$, we see that by using E_k (9) in place of w_k (8) we increase the expected total weight budget by approximately the factor $\sum_{k=1}^K (\mu_k^2 + \sigma_k^2) / \sum_{k=1}^K \sigma_k^2$. Our proposal furthermore controls type-I error, as the following theorem demonstrates.

Theorem 5.1. *In the above setting, suppose there exist $\eta, \delta \in [0, 1)$ such that $\mathbb{P}(A_\delta) \le \eta$, where A_δ is the "bad" event $\{\sum_{\ell=1}^K \hat{\sigma}_\ell^2 < (1-\delta) \sum_{\ell \in \mathcal{N}} \sigma_\ell^2\}$. Then, (i) the ep-Bonferroni procedure with t-test p-values P_k and (approximate compound) e-values E_k (9) controls the family-wise error rate at $\alpha' := \alpha/(1-\delta) + \eta$, and (ii) the ep-BH procedure with P_k and E_k controls the false discovery rate at α'.*

Proof. We prove the result for ep-Bonferroni to highlight the core ideas and defer the proof for ep-BH to Supplement S1.3. Applying the union bound twice, we see that, $\mathrm{FWER}_{\mathcal{D}} = \mathbb{P}(F_{\mathcal{D}} \ge 1) \le \mathbb{P}(\{F_{\mathcal{D}} \ge 1\} \cap A_\delta^c) + \eta \le \sum_{k \in \mathcal{N}} \mathbb{P}(\{P_k \le \alpha E_k / K\} \cap A_\delta^c) + \eta$. Let $\tilde{E}_k := K S_k^2 / \{(1-\delta) \sum_{\ell \in \mathcal{N}} \sigma_\ell^2\}$, then on the complement of the event A_δ it holds that $E_k \le \tilde{E}_k$. Hence we may bound $\mathbb{P}(\{P_k \le \alpha E_k / K\} \cap A_\delta^c)$ for null k by $\mathbb{P}(P_k \le \alpha \tilde{E}_k / K)$ and,

$$\mathbb{P}\left(P_k \le \frac{\alpha \tilde{E}_k}{K}\right) \overset{(*)}{=} \mathbb{E}\left[\mathbb{P}\left\{P_k \le \frac{\alpha S_k^2}{(1-\delta)\sum_{\ell \in \mathcal{N}} \sigma_\ell^2} \,\middle|\, S_k^2\right\}\right] \overset{(**)}{\le} \mathbb{E}\left\{\frac{\alpha S_k^2}{(1-\delta)\sum_{\ell \in \mathcal{N}} \sigma_\ell^2}\right\}.$$

In $(*)$ we applied iterated expectation, and in $(**)$ we used the independence in (7). To conclude, we recall that $\mathbb{E}(S_k^2) = \sigma_k^2$ for null k, and then we sum the above inequalities over all $k \in \mathcal{N}$. □

One may seek to recalibrate α via Theorem 5.1 to achieve finite-sample error control at the desired level. We do not advocate for such recalibration, since the type-I error inflation due to δ, η will typically be negligible, as we demonstrate next using standard concentration arguments [Boucheron et al., 2013, Chapter 2.4]. We postpone proof details to Supplement S1.4.

Proposition 5.2. *In Theorem 5.1, we may choose $\delta = \delta(\eta)$ for any $\eta \in (0,1)$ with $\delta(\eta) < 1$, where $\delta(\eta) := 2\{\log(1/\eta)\sum_{\ell\in\mathcal{N}}\sigma_\ell^4/(n-1)\}^{1/2}/\sum_{\ell\in\mathcal{N}}\sigma_\ell^2$. If $\sigma_k \in [\underline{\sigma}, \bar{\sigma}]$ for all $k \in \mathcal{N}$, where $0 < \underline{\sigma} \le \bar{\sigma}$, then $\delta(\eta) \le 2[\log(1/\eta)\bar{\sigma}^4/\{(n-1)K_0\underline{\sigma}^4\}]^{1/2}$, i.e., $\delta(\eta) = O\{\sqrt{\log(1/\eta)}/\sqrt{(nK_0)}\}$.*

As an example, suppose we apply ep-BH at target level $\alpha = 0.1$ and that model (5) holds with $\sigma_k^2 = 1$ for all $k \in \mathcal{K}$, $\pi_0 = 0.95$, $n = 10$, and $K = 20,000$. Then we can choose $\eta = 0.001$, $\delta < 0.0128$, so that the false discovery rate will be provably controlled at level at most 0.1023.

5.2 On the choice of weighting function

Above we explained our results for the compound e-values $E_k \propto S_k^2$. This choice renders the results most transparent. In the case of normalized weighting, instead of taking $w_k \propto S_k^2$, we could have proceeded with $w_k \propto \psi(S_k^2)$ for a fixed function $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$. Due to normalization, any (fixed) choice of $\psi(\cdot)$ leads to type-I error control, but the power of the resulting procedures depends on $\psi(\cdot)$. For example, Westfall et al. [2004] also considered the choice $\psi(s) = s^\nu$ for fixed $\nu > 0$, while Bourgon et al. [2010], Guo and Romano [2017] considered $\psi(s) = \mathbb{1}(s > c)$ for fixed $c > 0$.Ignatiadis et al. [2016], Ignatiadis and Huber [2021] proposed independent hypothesis weighting (IHW) which (in the present setting) uses normalized weights $w_k \propto \widehat{\psi}_{-k}(S_k^2)$, where for any $k \in \mathcal{K}$, $\widehat{\psi}_{-k}(\cdot)$ is learned based on $(S_\ell^2)_{\ell\in\mathcal{K}}$ as well as a subset of the p-values that excludes P_k. This subset is based on a cross-fold construction ("cross-weighting") which avoids overfitting and ensures type-I error control.

In the case of unnormalized weights, as developed in this work, it is also possible to consider $E_k \propto \psi(S_k^2)$ for more general choices of $\psi(\cdot)$. In Supplement S6, we motivate the choice

$$E_k := K\sum_{d=0}^{6}\frac{n^d\Gamma(n/2)}{4^d d!\Gamma(n/2+d)}(nS_k^2)^d \Bigg/ \sum_{\ell=1}^{K}\sum_{d=0}^{6}\frac{n^d\Gamma\{(n-1)/2\}}{4^d d!\Gamma\{(n-1)/2+d\}}\{(n-1)\hat{\sigma}_\ell^2\}^d. \tag{10}$$

By using (10) instead of (9), it is possible to increase the weight budget even further, i.e., $\sum_{k=1}^K E_k$ can be substantially larger than K. In the simulations of Section 7.2, (10) is powerful, however, we leave an investigation of the optimal (potentially data-driven) choice of e-value weights to future work.

5.3 A hierarchy of assumptions on side-information

Our discussion so far indicates a hierarchy of potential assumptions on side-information in multiple testing. Side-information in multiple testing refers to additional contextual knowledge, often in the form of covariates, that goes beyond the p-values P_k and can enhance the power of multiple testing procedures. (i) A practical and widely used assumption [Ignatiadis et al., 2016, Lei and Fithian, 2018, Li and Barber, 2019, Ignatiadis and Huber, 2021] is that the side-information (in this section, S_k^2) is independent of the p-value P_k for null k (this holds by (7) in the present setting). (ii) In other cases, the side-information can be used to compute a second p-value P_k' that is independent of P_k. This is true in the meta-analysis setting, however, it is also possible in the setting of this section. Du and Zhang [2014, Section 6.4] posit model (5) with $\sigma_k^2 = 1$ for all k, and assume that the common value of all σ_k^2 is known to the analyst. They let P_k be the t-test p-value as in (6) and $P_k' := 1-\chi_n(nS_k^2)$, where χ_n is the chi-square distribution with n degrees of freedom. Then, P_k' is a p-value that is independent of P_k for $k \in \mathcal{N}$. (iii) Our methodological developments (e.g., Theorem 5.1) indicate that there are assumptions for side-information intermediate in strength between (i) and (ii). In addition to (i), it suffices to control certain moments of the side-information averaged over all null coordinates (4).

One of the key messages of this paper is that it can be valuable to pursue (iii) through compound e-values as weights. In the model of this section, existing methods for multiple testing with side-information, e.g., independent hypothesis weighting, forfeit potential power gains by ignoring the distributional knowledge available for S_k^2 and instead using normalized weights (the upshot being that independent hypothesis weighting is applicable to more general forms of side-information). On the other hand, the assumption in Du and Zhang [2014, Section 6.4] that $\sigma_k^2 = 1$ for all k (and that their

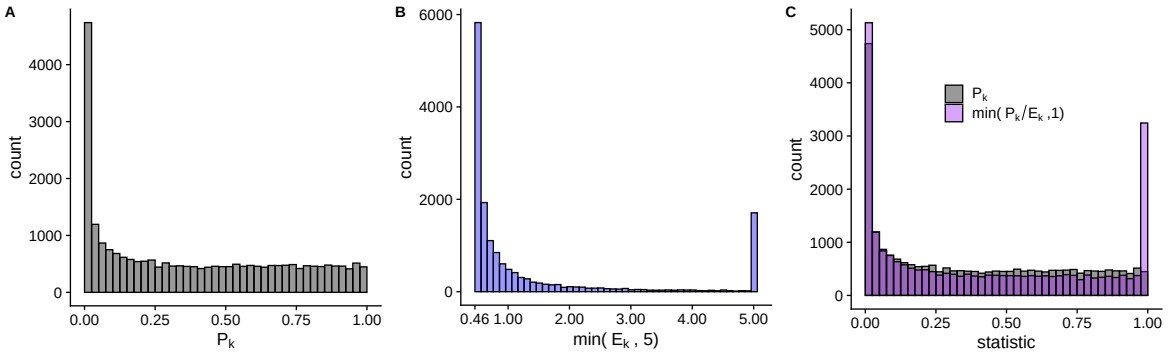

Figure 2: Differential gene expression analysis: We plot histograms of A) RNA-Seq p-values, B) microarray e-values (capped at 5), and C) Q-combiner p-values $Q(P_k, E_k) = (P_k/E_k) \wedge 1$ overlaid with the histogram of p-values from panel A). The smallest e-value (panel B) is equal to 0.46. There is an enrichment of $Q(P_k, E_k)$ compared to P_k in the bin closest to 0 (panel C).

common value is known to the analyst) is strong: if we are willing to impose it, then it would be advisable to conduct a z-test instead of the t-test in (6).

6 Differential gene expression based on RNA-Seq and Microarray data

As a demonstration of the practicality, applicability, and power of ep-BH, we seek to detect genes that are differentially expressed in the striatum of two mice strains (C57BL/6J and DBA/2J). We use two sources of information: RNA-Seq p-values and microarray e-values.

In more detail, we use the first two experimental batches of the RNA-Seq data collected by Bottomly et al. [2011] which comprise 7 adult male mice of each of the two strains (14 mice in total). We compute p-values for differential gene expression across the two strains using `DESeq2` [Love, Huber, and Anders, 2014] adjusting for the experimental batches. Furthermore, we use microarray (Affymetrix) expression measurements collected by Bottomly et al. [2011] for 10 adult male mice (5 of each strain). p-values for such comparisons are routinely computed based on the empirical Bayes model of Lönnstedt and Speed [2002] with the `limma` software package [Smyth, 2004], see e.g., the workflow of Klaus and Reisenauer [2018]. Here we demonstrate how one would proceed if e-values had been computed instead. In Supplement S7, we provide a construction of e-values under the replicated microarray model assumptions of Lönnstedt and Speed [2002], Smyth [2004]. The construction may be of independent interest for applications in high-throughput biology, e.g., for microarray or RNA-Seq data [Ritchie et al., 2015].

Our analysis leads to RNA-Seq p-values for 24,906 genes (Fig. 2A), and microarray e-values for a subset of 15,875 genes (Fig. 2B), where we map the microarray probe identifiers to Ensembl gene identifiers. We set the e-value for the remaining 9,031 genes to 1 (which is a valid e-value). The p-values are approximately independent of all e-values since Bottomly et al. [2011] used distinct mice for the RNA-Seq and Affymetrix microarray data collection.

We consider the following variants of weighted p-BH: (i) Unweighted p-BH that only uses the p-values P_k (and weights $w_k = 1$, thus ignoring the information in the e-values). (ii) E-value weighted p-BH procedure (ep-BH) with e-values as unnormalized weights $w_k = E_k$. Fig. 2C shows a histogram of the combined p-values $(P_k/E_k) \wedge 1$ that are used as input for ep-BH. (iii) Weighted p-BH (wBH) with normalized e-value weights, i.e., $w_k = K \cdot E_k / \sum_{\ell=1}^{K} E_\ell$. (iv) Independent hypothesis weighting (IHW) BH [Ignatiadis and Huber, 2021, Ignatiadis et al., 2016] with p-values P_k and the e-values E_k as side-information (see Section 5.2 for a brief description of the method). We use the imple-

	Avg. Weight	90% Weight	Discoveries
Non-adaptive			
Benjamini-Hochberg (BH, unweighted)	1.00	1.00	1973
E-value Weighted BH (ep-BH, our proposal)	18.11	2.70	2387
Weighted BH (wBH; normalized e-value weights)	1.00	0.15	1310
Indep. Hypothesis Weighted BH (IHW)	1.00	2.11	2016
Fisher BH	—	—	2354
Single Index Modulated BH (SIM)	—	—	2282
Adaptive			
Storey-BH (unweighted)	1.34	1.34	2147
E-value Weighted Storey-BH (our proposal)	24.35	3.63	2540
Weighted Storey-BH (normalized e-value weights)	2.63	0.39	1536
Indep. Hypothesis Weighted Storey-BH	1.54	3.49	2274
Fisher Storey-BH	—	—	2556
Single Index Modulated Storey-BH	—	—	2479

Table 1: Multiple testing for differential gene expression based on RNA-Seq p-values and microarray e-values: The last column shows the number of discoveries of each method. The first column shows the weight budget $\sum_{k=1}^{K} w_k/K$. For BH-type procedures with normalized weights, this quantity is always equal to 1. For Storey-type procedures, this budget is inflated due to accounting for the null proportion. The second column shows the upper 90% quantile among weights w_i used by each method.

mentation of independent hypothesis weighting [Ignatiadis and Huber, 2021, "IHW Grenander"] in the R/Bioconductor package "IHW" that stratifies hypotheses according to the side-information into $\lfloor K/1500 \rfloor$ groups of equal size.

We also consider two approaches that take two p-values P_k, P'_k per hypothesis as input (instead of a p-value and an e-value). P_k is the RNA-Seq p-value as above and P'_k is the p-value returned from the microarray analysis using `limma` (Supplement S7). These approaches are the (v) p-BH procedure applied to the Fisher combination p-values (Fisher), and (vi) single index modulated (SIM) p-BH [Du and Zhang, 2014], which applies p-BH to $P_k(\hat{\theta})$, where $P_k(\theta) := \Phi\{\cos(\theta)\Phi^{-1}(P_k) + \sin(\theta)\Phi^{-1}(P'_k)\}$ and Φ is the standard normal distribution function. $\hat{\theta}$ is selected as the index $\theta \in [0, \pi/2]$ that maximizes the number of discoveries of p-BH applied to $(P_k(\theta))_{k\in\mathcal{K}}$. (Single index modulated p-BH controls the false discovery rate asymptotically, but there is no finite-sample guarantee due to the data-driven choice of $\hat{\theta}$.)

We also consider null proportion adaptive versions of the above procedures using variants of Storey's adjustment (with $\tau = 0.5$): for the unweighted method, Fisher, and single index modulation we use Storey's procedure (Storey et al., 2004 and (3)), for e-value weights we use ep-Storey, for normalized e-value weights we use weighted Storey as described in Ramdas et al. [2019] and for independent hypothesis weighting we follow Ignatiadis and Huber [2021, Theorem 2]. All procedures are applied to control the false discovery rate at the target level $\alpha = 0.01$.

The results of the analysis are shown in Table 1. Among the non-adaptive procedures, ep-BH makes the most discoveries, even compared to Fisher BH and single index modulated BH which have access to two independent p-values per hypothesis. The procedure that normalizes the e-value weights makes by far the least discoveries. The reason is that with the exception of the genes with the highest e-values, all other genes receive very small weights. Independent hypothesis weighting makes more discoveries than unweighted BH demonstrating that the ordering of e-values can be used to increase power, even among procedures that use normalized weights. The findings for the adaptive procedures are analogous, although in this case, Fisher Storey-BH makes the most discoveries (with a small margin compared to our ep-Storey method).

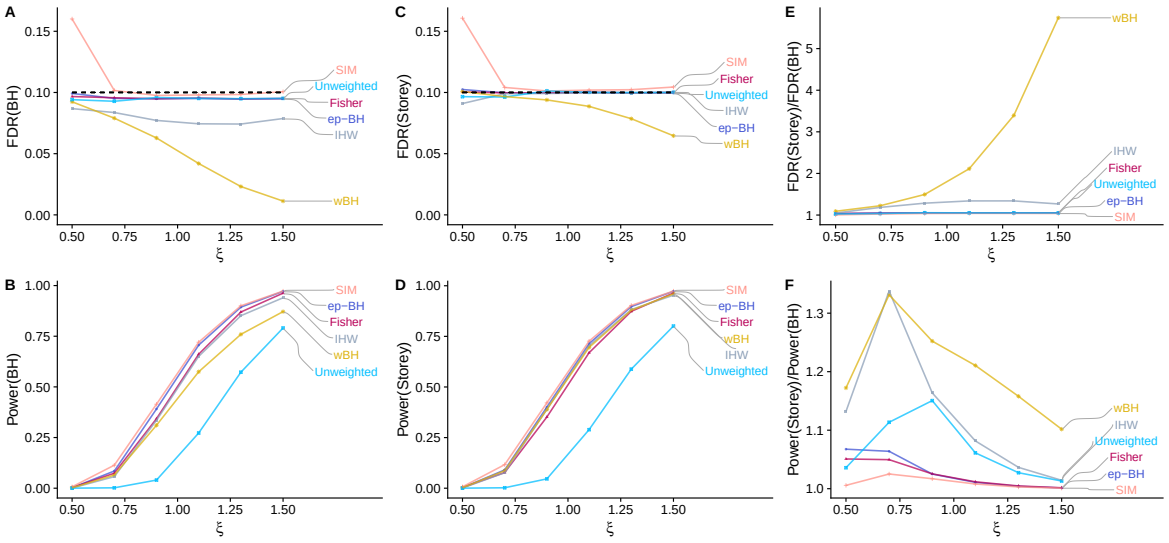

Figure 3: t-test simulations: We compare non-adaptive (BH-based) methods plotting A) false discovery rate (FDR) and B) power against the effect size parameter ξ. All methods control the false discovery rate except single index modulated BH (SIM) at small ξ. Single index modulated BH is most powerful, followed by ep-BH, Fisher BH, and independent hypothesis weighting (IHW). We also evaluate the adaptive (Storey-based) counterparts of the same methods plotting their C) false discovery rate and D) power, as well the ratio of E) false discovery rate and F) power between the Storey-based and BH-based methods. Weighted p-BH (wBH) and independent hypothesis weighting benefit the most from null proportion adaptivity, cf. Remark 4.11.

7 Simulation study

7.1 Evaluation

For the simulation study we compare the same methods as in Section 6. We apply these methods at a target false discovery rate of $\alpha = 0.1$. We evaluate methods in terms of their false discovery rate, and their power, which we define as $\mathbb{E}\{(R_{\mathcal{D}} - F_{\mathcal{D}})/(K - K_0)\}$.

7.2 One sample t-test

We conduct a simulation study in the setting of Section 5 and generate data from model (5). We let $n = 10$, $K = 20,000$, $\pi_0 = 0.95$ and set $\mu_k = \xi$ for the alternative hypotheses, where the effect size $\xi \in [0.5,\ 1.5]$ is a varying simulation parameter. We fix $\sigma_k = 1$ for all k. We use the e-values (10) for ep-BH and weighted p-BH, and the secondary p-values P'_k defined in (ii) of Section 5.3 for Fisher and single index modulated BH. We average results over $4,000$ Monte Carlo replicates of each simulation setting.

Among the weighted BH methods, all procedures control the false discovery rate (Fig. 3A), and ep-BH has the most power (Fig. 3B), followed by independent hypothesis weighting. Weighted p-BH (with normalized e-value weights) and unweighted p-BH have low power. Single index modulated BH has the most power, but the difference to ep-BH is small, especially considering the requirement of an additional p-value (instead of e-value) per hypothesis and that it exceeds the target false discovery rate at small ξ. The power of ep-BH and Fisher BH is similar. The false discovery rate, resp. power of the adaptive procedures is shown in Fig. 3C, resp. 3D. Figs. 3E,F show the ratio of the false discovery rate and power of the adaptive methods compared to their non-adaptive counterparts. As explained

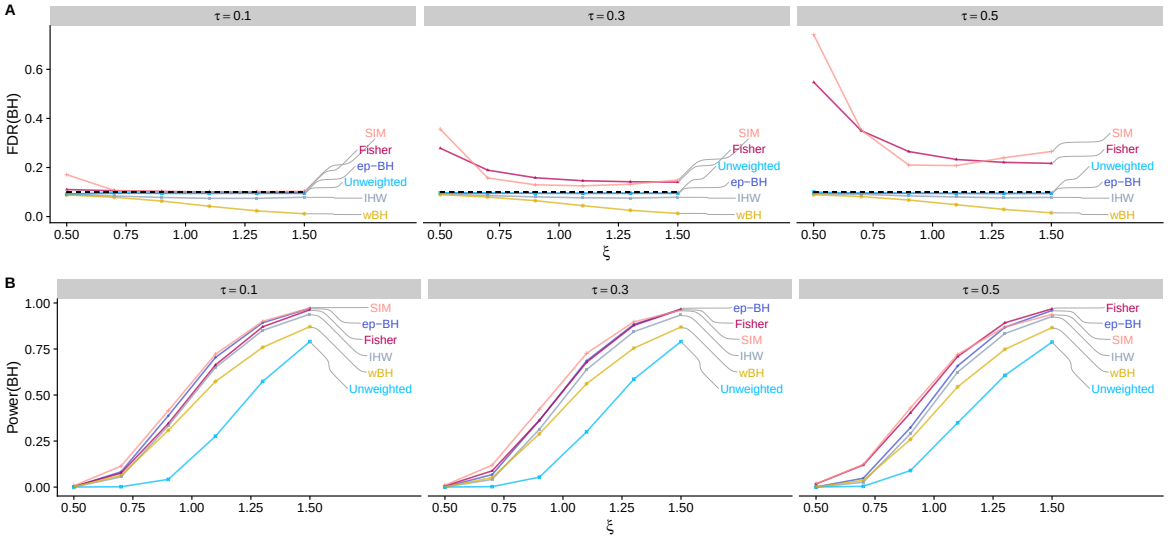

Figure 4: t-test simulations with heteroscedasticity: Panels A) and B) are analogous to panels A and B of Fig. 3. We compare non-adaptive methods in terms of their false discovery rate (FDR) and power. While in Fig. 3, $\sigma_k^2 = 1$ for all $k \in \mathcal{K}$, here we enforce increasingly strong heteroscedasticity by drawing $\sigma_k^2 \sim \mathrm{U}(1-\tau, 1+\tau)$, where $\tau \in \{0.1, 0.3, 0.5\}$ corresponds to the different facets. Fisher and single index modulated BH do not control the false discovery rate under strong heteroscedasticity ($\tau \in \{0.3, 0.5\}$), while the other methods do (cf. Section 5.3). ep-BH has the most power among methods controlling the false discovery rate.

in Remark 4.11, the procedures with normalized weights derive most benefit from null proportion adaptivity. ep-BH achieves strong power gains even without null proportion adaptivity.

We next tweak the simulation and introduce heteroscedasticity by drawing $\sigma_k^2 \sim \mathrm{U}(1-\tau, 1+\tau)$, where $\tau \in \{0.1, 0.3, 0.5\}$ is a simulation parameter. In Fig. 4 we show results for the non-adaptive methods only. The weighted BH methods perform similarly as in the homoscedastic case of Fig. 3. Methods that take two p-values as input (Fisher and single index modulated BH) strongly violate the target false discovery rate for $\tau \in \{0.3, 0.5\}$. The reason is that the secondary p-value is no longer a valid p-value when the assumption $\sigma_k^2 = 1$ is violated (cf. Section 5.3).

7.3 Combining RNA-Seq and microarray data

We next consider a simulation that mimics the RNA-Seq/microarray application of Section 6. We simulate datasets with $K = 10,000$ genes of two-sample comparisons with 20 samples for each combination of condition (control/treatment) and technology (RNA-Seq/microarray). For the synthetic RNA-Seq datasets, inspired by the simulation setup in Love et al. [2014], we generate negative binomial count data with mean and dispersion parameters chosen to approximate realistic moments by resampling from the joint distribution of mean/dispersion parameters of the simulations in Love et al. [2014], truncated to mean values of at least 1. We let $\pi_0 = 0.8$, and sample the alternative genes uniformly among all genes and then set the (binary) logarithmic fold changes of the treated samples to $+\xi$, resp. $-\xi$ with probability 1/2, where $\xi \in [0.3, 0.9]$ is a varying simulation parameter. To generate synthetic microarray data, we follow the model described in Supplement S7. We first sample variances σ_k^2, $k \in \mathcal{K}$ from (S16) with $\nu_0 = 3.64$ and $s_0^2 = 0.0144$ (which are the estimates of ν_0 and s_0^2 in the microarray data described in Section 6). We then order the variances according to the order of the mean counts from the RNA-Seq simulation (i.e., the gene with largest mean count in the RNA-Seq simulation also has the largest variance in the microarray experiment). For all null genes, the effect

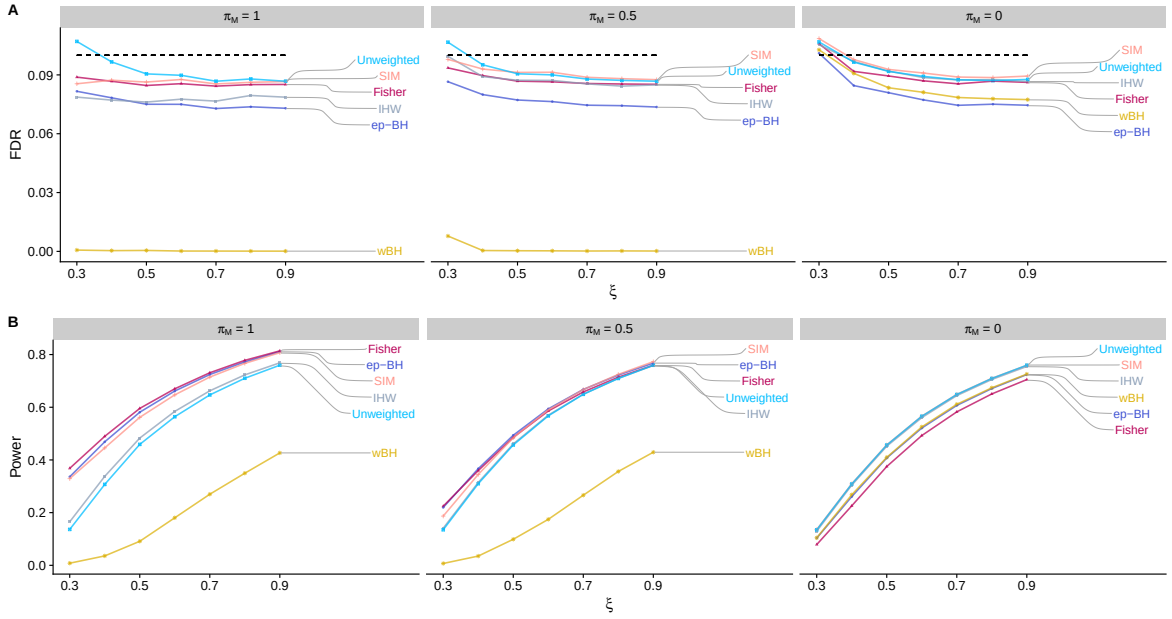

Figure 5: RNA-Seq and microarray meta-analysis simulation: We compare non-adaptive (BH-based) methods plotting A) false discovery rate (FDR) and B) power against the effect size ξ and against the informativeness of the microarray data (parameter π_M in the facets). When $\pi_M = 1$, the microarray data are fully informative and Fisher BH has the most power followed by ep-BH. weighted p-BH (wBH) has less power than even unweighted p-BH. When the microarray data are completely uninformative ($\pi_M = 0$), then unweighted p-BH, independent hypothesis weighting (IHW), and single index modulated BH (SIM) have the most power, and Fisher BH has the least power. At intermediate informativeness ($\pi_M = 0.5$), weighted p-BH has the least power.

size is $\beta_k = 0$, while for differentially expressed genes we let $\beta_k \sim (1 - \pi_M)\delta_0 + \pi_M \mathrm{N}(0,\, 0.5\sigma_k^2)$, where δ_0 is a point mass at 0 and $\pi_M \in \{1, 0.5, 0\}$ is a simulation parameter. In words, when $\pi_M = 0$, the microarray dataset is completely uninformative, while when $\pi_M = 1$ all differentially expressed genes in the RNA-Seq dataset are also differentially expressed in the microarray dataset. We generate summary statistics for the 20 vs. 20 comparisons for each gene as in (S15). Finally, we compute p-values and e-values as in Section 6. Results are averaged over 100 Monte Carlo replicates.

All non-adaptive methods control the false discovery rate (Fig. 5A). When $\pi_M = 1$, Fisher BH has the most power (Fig. 5B), followed by ep-BH, single index modulated BH, and independent hypothesis weighting. ep-BH more than doubles the power compared to unweighted p-BH at the smallest values of the logarithmic fold change ξ. Weighted p-BH has less power than unweighted p-BH and its false discovery rate is almost 0. The case $\pi_M = 0$ is chosen as a challenging setting for ep-BH with completely uninformative e-values. In this case, ep-BH and weighted p-BH have similar power, which is only slightly lower than the power of unweighted p-BH, single index modulated BH, and independent hypothesis weighting. Single index modulated BH, and independent hypothesis weighting essentially collapse to unweighted p-BH when the second p-value (resp. e-value) is uninformative. Fisher BH has the lowest power in this case. For the intermediate choice of $\pi_M = 0.5$, single index modulated BH, ep-BH, and Fisher BH have the most power. Supplementary Fig. S1 shows the false discovery rate and power of the adaptive procedures and the overall take-home message remains the same: when e-values are informative, ep-BH can lead to substantial and practical power gains while maintaining type-I error guarantees.

Supplementary material

Supplementary material includes omitted proofs, methodological details (e.g., the first (minimally) adaptive e-BH procedure, inspired by Solari and Goeman [2017] in Supplement S3.2), and additional simulation results. All numerical results of this paper are fully third-party reproducible, and we provide the code on Github: `https://github.com/nignatiadis/evalues-as-weights-paper`.

References

Timothy B. Armstrong. False discovery rate adjustments for average significance level controlling tests. *arXiv preprint arXiv:2209.13686*, 2022.

Rina Foygel Barber and Aaditya Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (4):1247–1268, 2017.

Debabrata Basu. On statistics independent of a complete sufficient statistic. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(4):377–380, 1955.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57:289–300, 1995.

Yoav Benjamini and Yosef Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

Carlo E Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste*. Studi in Onore del Professore Salvatore Ortu Carboni, Rome, Italy, 1935.

Daniel Bottomly, Nicole AR Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J Buck, Robert P Searles, Michael Mooney, Shannon K McWeeney, and Robert Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PloS One*, 6(3):e17820, 2011.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, United Kingdom, 2013.

Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21): 9546–9551, 2010.

Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York, New York, NY, 2008.

Lilun Du and Chunming Zhang. Single-index modulated multiple testing. *The Annals of Statistics*, 42(4):1262–1311, 2014.

Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *arXiv preprint arXiv:2111.09254*, 2021.

Guillermo Durand. Adaptive p-value weighting with power optimality. *Electronic Journal of Statistics*, 13(2):3336–3385, 2019.

Helmut Finner, Thorsten Dickhaus, and Markus Roters. On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, 37(2):596–618, 2009.

Livio Finos and Luigi Salmaso. FDR- and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference*, 137(12):3859–3870, 2007.

Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *arXiv preprint arXiv:1906.07801v3*, 2021.

Wenge Guo and Joseph P Romano. Analysis of error control in large scale two-stage multiple hypothesis testing. *arXiv preprint arXiv:1703.06336*, 2017.

Joshua D Habiger. Adaptive false discovery rate control for heterogeneous data. *Statistica Sinica*, pages 1731–1756, 2017.

N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018.

Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4): 800–802, 1988.

Peter Hoff. Smaller p-values via indirect information. *Journal of the American Statistical Association*, 117(539):1254–1269, 2022.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

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

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Nikolaos Ignatiadis and Wolfgang Huber. Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):720–751, 2021.

Nikolaos Ignatiadis and Bodhisattva Sen. Empirical partially Bayes multiple testing and compound χ^2 decisions. *arXiv preprint arXiv:2303.02887*, 2023.

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580, 2016.

Bernd Klaus and Stefanie Reisenauer. An end to end workflow for differential gene expression using Affymetrix microarrays. *F1000Research*, 5:1384, 2018.

Lihua Lei and William Fithian. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4): 649–679, 2018.

Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74, 2019.

Kun Liang and Dan Nettleton. Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):163–182, 2012.

Ingrid Lönnstedt and Terry Speed. Replicated microarray data. *Statistica Sinica*, pages 31–46, 2002.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

Mengyin Lu and Matthew Stephens. Variance adaptive shrinkage (vash): flexible empirical Bayes estimation of variances. *Bioinformatics*, 32(22):3428–3434, 2016.

Ruth Marcus, Peritz Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Aaditya Ramdas, Rina F. Barber, Martin J. Wainwright, and Michael I. Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47:2790–2821, 2019.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*, 2022.

Zhimei Ren and Rina Foygel Barber. Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*, 2022.

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

Kathryn Roeder and Larry Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24(4):398, 2009.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754, 1986.

Gordon K Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.

Aldo Solari and Jelle J Goeman. Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biometrical Journal*, 59(4):776–780, 2017.

John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Weijie J. Su. The FDR-linking theorem. *arXiv preprint arXiv:1812.08965*, 2018.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):822–852, 2022.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Peter H. Westfall, Siegfried Kropf, and Livio Finos. Weighted FWE-controlling methods in high-dimensional situations. In *Recent Developments in Multiple Comparison Procedures*, volume 47 of *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 143–154. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2004.

Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. A unified framework for bandit multiple testing. In *Advances in Neural Information Processing Systems*, volume 34, pages 16833–16845, 2021.

S1 Omitted proofs

S1.1 Proof of Theorem 3.1

Let P be a p-value and E be an e-value. They are assumed independent in (i) and (ii) below. For a fixed $e \in [0,\infty)$, we will frequently rely on a specific distribution F_e of e-values given by, for $X \sim F_e$, $\mathbb{P}(X = e) = 1/e = 1 - \mathbb{P}(X = 0)$ if $e \geq 1$ and $\mathbb{P}(X = e) = 1/2 = \mathbb{P}(X = 2 - e)$ if $e < 1$. It is clear that $\mathbb{E}(X) = 1$.

(i) We have $\mathbb{E}\{h(P)E\} \leq 1$ since $h(P)$ is an e-value independent of E. Hence, Π_h is an i-pe/e combiner. To show its admissibility, suppose for the purpose of contradiction that an i-pe/e combiner f satisfies $f \geq \Pi_h$ and $f(p,e) > \Pi_h(p,e)$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. Clearly, $e \in [0,\infty)$ and $p \in (0,1]$. Since h is upper semicontinuous and $q \mapsto f(q,e)$ is decreasing, there exists $p' < p$ such that $f(q,e) \geq f(p,e) > \Pi_h(p',e) \geq \Pi_h(q,e)$ for all $q \in [p',p]$. Take $P \sim \mathrm{U}(0,1)$ and $E \sim F_e$. Since $\mathbb{E}\{\Pi_h(P,E)\} = 1$, $f \geq \Pi_h$, and $f(q,e) > \Pi_h(q,e)$ for all $q \in [p',p]$, we have $\mathbb{E}\{f(P,E)\} > \mathbb{E}\{\Pi_h(P,E)\} = 1$ which means that $f(P,E)$ is not an e-value, contradicting the fact that f is an i-pe/e combiner. This contradiction shows that Π_h is admissible.

(ii) For $\alpha \in (0,1)$, we have $\mathbb{P}\{Q(P,E) \leq \alpha\} = \mathbb{P}(P \leq \alpha E) = \mathbb{E}\{\mathbb{P}(P \leq \alpha E|E)\} \leq \mathbb{E}(\alpha E) \leq \alpha$. Therefore, Q is an i-pe/p combiner. To show its admissibility, suppose for the purpose of contradiction that an i-pe/p combiner f satisfies $f \leq Q$ and $f(p,e) < (p/e) \wedge 1$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. Since $a \mapsto f(p,a)$ is decreasing, we can assume $e \in [p,\infty)$ by replacing e with p if $e < p$. Take $P \sim \mathrm{U}(0,1)$. Since $q \mapsto f(q,e)$ is increasing, there exists $p' < p$ such that $f(q,e) \leq f(p,e) < p'/e$ for all $q \in [0,p]$. This gives $\mathbb{P}\{f(P,e) \leq p'/e\} \geq \mathbb{P}(P \leq p) = p$. For $\alpha = p'/e \in (0,1)$, if $e \geq 1$, then take $E \sim F_e$, so that:

$$\mathbb{P}\{f(P,E) \leq \alpha\} = \mathbb{P}\{f(P,e) \leq p'/e\}e^{-1} + \mathbb{P}\{f(P,0) \leq \alpha\}(1 - e^{-1}) \geq p/e > \alpha.$$

If $e < 1$, then take E distributed such that $\mathbb{P}(E = e) = \lambda$ and $\mathbb{P}\{E = (1-\lambda e)/(1-\lambda)\} = 1 - \lambda$ with $\lambda \in (0,1)$ chosen sufficiently small, so that $\alpha(1-\lambda e)/(1-\lambda) < 1$. Then:

$$\begin{aligned}
\mathbb{P}\{f(P,E) \leq \alpha\} &= \lambda \mathbb{P}\{f(P,e) \leq p'/e\} + (1-\lambda)\mathbb{P}[f\{P,(1-\lambda e)/(1-\lambda)\} \leq \alpha] \\
&\geq \lambda p + (1-\lambda)\mathbb{P}\{P \leq \alpha(1-\lambda e)/(1-\lambda)\} \\
&= \lambda p + (1-\lambda)\alpha(1-\lambda e)/(1-\lambda) \\
&= \lambda p + (1-\lambda e)p'/e \\
&= \lambda(p-p') + p'/e > \alpha.
\end{aligned}$$

Hence, $f(P,E)$ is not a p-value, and this contradicts the fact that f is an i-pe/p combiner. This contradiction shows that Q is admissible.

(iii) The weighted average of two arbitrary e-values is an e-value; hence M_h^λ is a pe/e combiner. Its admissibility follows essentially the same proof as part (i), which we do not repeat.

(iv) Since $1/E$ is a p-value, the Bonferroni combination of P and $1/E$, $2\min(P,1/E)$, is a p-value, and hence B is a pe/p combiner. To show its admissibility, suppose for the purpose of contradiction that a pe/p combiner f satisfies $f \leq B$ and $f(p,e) < \{2(p \wedge e^{-1}) \wedge 1\}$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. By monotonicity of f, we can increase e to $1/p$ or decrease p to $e^{-1} \wedge (1/2)$, and this does not change the value of $\{2(p \wedge e^{-1}) \wedge 1\}$. Hence, we can assume that $f(p,1/p) < 2p$ for some $p \in (0,1/2]$ by noting that $f(p,1/p) < 2p$ automatically holds for $p > 1/2$ because $B \leq 1$. Since $q \mapsto f(q,e)$ is increasing, there exists $p' < p$ such that $f(q,1/p) \leq f(p,1/p) < 2p'$ for all $q \in [0,p]$. Take $P \sim \mathrm{U}(0,1)$ and define $E = (1/p)\mathbb{1}(P \in [p',p)) + (1/p')\mathbb{1}(P \in [p,p+p'p'/p])$. Then $\mathbb{E}(E) = (p-p')/p + p'/p = 1$. Next, consider the following three cases: 1) when $P \leq p'$, then $f(P,E) \leq 2p'$, 2) when $P \in [p,p+p'p'/p]$, then $E \leq 1/p'$ and so $f(P,E) \leq 2p'$, 3) when

$P \in (p', p)$, then $f(P, E) \leq f(p, 1/p) \leq 2p'$. Hence $f(P, E) \leq 2p'$ on the event $\{P \leq p + p'p'/p\}$, and so:

$$\mathbb{P}\{f(P, E) \leq 2p'\} \geq \mathbb{P}(P \leq p + p'p'/p) = p + p'p'/p = p'(p/p' + p'/p) > 2p'.$$

Hence, $f(P, E)$ is not a p-value, and this contradicts the fact that f is a pe/p combiner. This contradiction shows that B is admissible.

S1.2 Proof of Theorem 4.10

As in the proof of Theorem 4.7 we start by arguing conditionally on $(E_k)_{k\in\mathcal{K}}$. The proof of Ramdas et al. [2019, Theorem 1a] yields that:

$$\mathbb{E}\left\{\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mid (E_k)_{k\in\mathcal{K}}\right\} \leq \frac{\alpha}{K}\sum_{k\in\mathcal{N}} E_k \mathbb{E}\left\{\frac{1}{\widehat{\pi}_0^{-k}} \mid (E_k)_{k\in\mathcal{K}}\right\}, \text{ where } \widehat{\pi}_0^{-k} = \frac{1 + \sum_{j\neq k} \mathbb{1}_{\{P_j > \tau\}}}{K(1-\tau)}.$$

By Ramdas et al. [2019, Lemma 3], it follows that $\mathbb{E}\{1/\widehat{\pi}_0^{-k} \mid (E_k)_{k\in\mathcal{K}}\} \leq 1/\pi_0$, and hence,

$$\mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right) = \mathbb{E}\left[\mathbb{E}\left\{\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mid (E_k)_{k\in\mathcal{K}}\right\}\right] \leq \frac{\alpha}{K}\mathbb{E}\left(\sum_{k\in\mathcal{N}} E_k \frac{K}{K_0}\right) \leq \alpha,$$

as claimed.

S1.3 Proof for ep-BH with data-driven weights (Theorem 5.1)

Proof. We now prove the FDR control guarantee for ep-BH with data-driven e-values. We call the procedure $\mathcal{D}$. Let $\hat{c} := \sum_{\ell=1}^{K} \hat{\sigma}_\ell^2$, so that $E_k = KS_k^2/\hat{c}$.

It will be helpful to note the following standard decomposition:

$$\sum_{j=1}^{n} Y_{kj}^2 = \sum_{j=1}^{n} (Y_{kj} - \hat{\mu}_k)^2 + n\hat{\mu}_k^2 \implies nS_k^2 = (n-1)\hat{\sigma}_k^2 + n\hat{\mu}_k^2.$$

Dividing by $\hat{\sigma}_k^2$ yields that $nS_k^2/\hat{\sigma}_k^2 = (n-1) + T_k^2$, i.e.,

$$\hat{\sigma}_k^2 = \frac{nS_k^2}{(n-1) + T_k^2}.$$

Let us write $\hat{c}$ as a function of S_k^2, T_k^2, $k \in \mathcal{K}$:

$$\hat{c} = \frac{1}{K}\sum_{\ell=1}^{K} \frac{nS_\ell^2}{(n-1) + T_\ell^2}.$$

Finally, noting that P_k is a strictly decreasing function of T_k^2, we see that

$$\hat{c} = h((P_k)_{k\in\mathcal{K}}, (S_k^2)_{k\in\mathcal{K}}),$$

where $h(\cdot)$ is a function $[0,1]^K \times \mathbb{R}_{>0}^K \to \mathbb{R}_{>0}$. For fixed $(S_k^2)_{k\in\mathcal{K}}$, $h(\cdot, (S_k^2)_{k\in\mathcal{K}})$ is increasing in $(P_k)_{k\in\mathcal{K}}$.

By the preceding arguments, we see that in fact we may interpret the whole multiple testing procedure $\mathcal{D}$ as a function of $(P_k)_{k\in\mathcal{K}}, (S_k^2)_{k\in\mathcal{K}}$, i.e., $\mathcal{D} = \mathcal{D}((P_k)_{k\in\mathcal{K}}, (S_k^2)_{k\in\mathcal{K}})$. Furthermore, for fixed $(S_k^2)_{k\in\mathcal{K}}$, the number of rejections of $\mathcal{D}$ are decreasing in $(P_k)_{k\in\mathcal{K}}$—this follows by standard arguments for (weighted) p-BH along with the monotonicity established for $\hat{c}$.

Let us also define $c^* := \sum_{\ell \in \mathcal{N}} \sigma_\ell^2$ and also $\tilde{c} := \max\{\hat{c}, (1-\delta)c^*\}$. Furthermore let $\mathcal{D}'$ be the ep-BH procedure with (approximate) e-values $\tilde{E}_k = KS_k^2/\tilde{c}$ (instead of $KS_k^2/\hat{c}$).[1] Notice that $\mathcal{D}$ and $\mathcal{D}'$ are identical on the event A_δ^c, i.e., on the complement of the event A_δ. Furthermore, $\mathcal{D}'$ and $\tilde{c}$ inherit the monotonicity properties that we established for $\mathcal{D}$ and $\hat{c}$ above. We argue that it suffices to study $\mathcal{D}'$:

$$\mathrm{FDR}_{\mathcal{D}} = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right) = \mathbb{E}\left\{\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\mathbb{1}(A_\delta)\right\} + \mathbb{E}\left\{\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\mathbb{1}(A_\delta^c)\right\} \leq \mathbb{P}(A_\delta) + \mathbb{E}\left(\frac{F_{\mathcal{D}'}}{R_{\mathcal{D}'}}\right).$$

By assumption, it holds that $\mathbb{P}(A_\delta) \leq \eta$, and so it suffices to bound the second term.

Let us call $\mathcal{O} = \{(S_k^2)_{k\in\mathcal{K}}, (P_k)_{k\in\mathcal{K}\setminus\mathcal{N}}\}$. By (7) and our assumption of joint independence of all the Y_{kj}, it follows that $(P_k)_{k\in\mathcal{N}}$ is distributed as $U[0,1]^{K_0}$ conditionally on $\mathcal{O}$. For $k \in \mathcal{N}$:

$$\begin{aligned}\mathbb{E}\left\{\frac{\mathbb{1}(H_k \text{ rejected})}{R_{\mathcal{D}'}}\right\} &\leq \mathbb{E}\left\{\frac{\mathbb{1}\left(P_k \leq \frac{\alpha R_{\mathcal{D}'} S_k^2}{\tilde{c}}\right)}{R_{\mathcal{D}'}}\right\} = \mathbb{E}\left[\mathbb{E}\left\{\frac{\mathbb{1}\left(P_k \leq \frac{\alpha R_{\mathcal{D}'} S_k^2}{\tilde{c}}\right)}{R_{\mathcal{D}'}}\middle| \mathcal{O}\right\}\right] \\ &\overset{(*)}{\leq} \mathbb{E}\left(\frac{\alpha S_k^2}{\tilde{c}}\right) \leq \frac{\alpha}{(1-\delta)c^*}\mathbb{E}(S_k^2) = \frac{\alpha}{(1-\delta)} \cdot \frac{\sigma_k^2}{\sum_{\ell\in\mathcal{N}}\sigma_\ell^2}.\end{aligned}$$

The crucial argument is $(*)$. Herein we applied the superuniformy lemma of Ramdas et al. [2019, Lemma 1(b)] conditionally on $\mathcal{O}$. Summing over all $k \in \mathcal{N}$, we conclude. □

S1.4 Proof of Proposition 5.2

Proof. The crux of the argument is that the left tail of a gamma random variable is sub-Gaussian. In particular, let $X \sim \Gamma(a, b)$ for $a, b > 0$ (where a is the shape and b is the scale). Then, e.g., by Boucheron et al. [2013, Chapter 2.4]:

$$\mathbb{E}(\exp[\lambda\{X - \mathbb{E}(X)\}]) \leq \exp(\lambda^2 ab^2/2) \quad \text{for any} \quad \lambda < 0.$$

Next notice that for any $k \in \mathcal{K}$, it holds that $\hat{\sigma}_k^2 \sim \Gamma\{(n-1)/2,\, 2\sigma_k^2/(n-1)\}$ and $\mathbb{E}(\hat{\sigma}_k^2) = \sigma_k^2$. Hence, using independence across k, it follows that:

$$\mathbb{E}\left(\exp\left[\lambda\left\{\sum_{\ell=1}^K(\hat{\sigma}_\ell^2 - \sigma_\ell^2)\right\}\right]\right) \leq \exp\left(\frac{\lambda^2}{2}\frac{2}{n-1}\sum_{\ell=1}^K\sigma_\ell^4\right) \quad \text{for any} \quad \lambda < 0.$$

By a standard Chernoff argument (applied to the left tail), this implies that for any $\varepsilon > 0$:

$$\mathbb{P}\left\{\sum_{\ell=1}^K(\hat{\sigma}_\ell^2 - \sigma_\ell^2) < -\varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon^2(n-1)}{4\sum_{\ell=1}^K\sigma_\ell^4}\right\}. \tag{S1}$$

We may upper bound the probability of the "bad event" A_δ as follows:

$$\mathbb{P}(A_\delta) = \mathbb{P}\left\{\sum_{\ell=1}^K\hat{\sigma}_\ell^2 < (1-\delta)\sum_{\ell\in\mathcal{N}}\sigma_\ell^2\right\} \leq \mathbb{P}\left\{\sum_{\ell\in\mathcal{N}}(\hat{\sigma}_\ell^2 - \sigma_\ell^2) < -\delta\sum_{\ell\in\mathcal{N}}\sigma_\ell^2\right\}.$$

Now let $\varepsilon = \delta\sum_{\ell\in\mathcal{N}}\sigma_\ell^2$. Then, so that $\mathbb{P}(A_\delta) < \eta$ for $\eta \in (0, 1)$, it suffices to choose ε such that the right hand side in (S1) (applied only to the nulls) is less than η, i.e., we may choose any

$$\varepsilon \geq 2\left\{\log(1/\eta)\frac{\sum_{\ell\in\mathcal{N}}\sigma_\ell^4}{n-1}\right\}^{1/2}.$$

[1] An analogous proof strategy is pursued in Blanchard and Roquain [2008, Lemma 4.3].

Rearranging in terms of δ, it thus suffices that:

$$\delta \geq 2\left\{\log(1/\eta)\frac{\sum_{\ell\in\mathcal{N}}\sigma_\ell^4}{n-1}\right\}^{1/2}\Big/\sum_{\ell\in\mathcal{N}}\sigma_\ell^2 =: \delta(\eta). \tag{S2}$$

The specific form of $\delta(\eta)$ announced in the statement of the proposition follows from the requirement that $\delta \in [0,1)$.

Now suppose that $\sigma_k \in [\underline{\sigma}, \bar{\sigma}]$ for all $k \in \mathcal{N}$. Then:

$$\delta(\eta) \leq 2\left\{\log(1/\eta)\frac{K_0\bar{\sigma}^4}{n-1}\right\}^{1/2}\Big/(K_0\underline{\sigma}^2) = 2\left\{\log(1/\eta)\frac{\bar{\sigma}^4}{(n-1)K_0\underline{\sigma}^4}\right\}^{1/2}.$$

Let us also highlight at this point that the above bound does not depend at all on the configuration of the variances of the alternative hypotheses. □

S2 Results for additional multiple testing procedures

S2.1 Definition of additional procedures

Definition S2.1 (p-Holm procedure (Holm, 1979))**.** For $k \in \mathcal{K}$, let $P_{(k)}$ be the k-th order statistic of the p-values $P_1, \ldots, P_K$, from the smallest to the largest. The p-Holm procedure rejects the k^*_{hm} hypotheses with the smallest p-values, where

$$k^*_{\text{hm}} := \max\left\{k \in \mathcal{K} \,:\, P_{(j)} \leq \frac{\alpha}{K-j+1} \text{ for all } j = 1, \ldots, k\right\},$$

with the convention $\max(\varnothing) = 0$.

The p-Holm procedure controls the family-wise error rate under arbitrary dependence between the p-values. The p-Holm procedure may be derived as the closed testing procedure based on the p-Bonferroni test.

Definition S2.2 (p-Hommel procedure (Hommel, 1988))**.** For $k \in \mathcal{K}$, let $P_{(k)}$ be the k-th order statistic of the p-values $P_1, \ldots, P_K$, from the smallest to the largest. The p-Hommel procedure computes

$$k^*_{\text{hl}} := \max\left\{k \in \mathcal{K} \,:\, P_{(K-k+i)} > \frac{i\alpha}{k} \text{ for all } i = 1, \ldots, k\right\},$$

with the convention $\max(\varnothing) = 0$. If $k^*_{\text{hl}} = 0$, then the p-Hommel procedure rejects all hypotheses, otherwise it rejects all hypotheses with $P_k \leq \alpha/k^*_{\text{hl}}$.

The p-Hommel procedure controls the family-wise error rate when $(P_k)_{k\in\mathcal{K}}$ is positive regression dependent within nulls. Furthermore, the p-Hommel procedure is the exact closed testing procedure based on the p-Simes test (while p-Hochberg is a shortcut).

S2.2 Results for ep-$\mathcal{D}$ procedures

By plugging in the p-Hommel procedure (Definition S2.2) into Definition 4.2 we get the ep-Hommel procedure. Analogously, by plugging in the p-Holm procedure (Definition S2.1) into Definition 4.2, we get the ep-Holm procedure.

Suppose the assumptions of Theorem 4.3 hold. Then ep-Holm controls the family-wise error rate at level α (since p-Holm controls the family-wise error rate under arbitrary p-value dependence).

Suppose the assumptions of Theorem 4.9 hold. Then, the ep-Hommel procedure controls the family-wise error rate at level α; the proof is entirely analogous to the proof for ep-Hochberg.

S3 Multiple testing with the i-pe/e combiner Π_h

In this supplement we provide further results on multiple testing with e-values and p-values that go beyond the Q-combiner. Throughout we assume that Assumption 4.1 holds, that is, P_k is independent of E_k for $k \in \mathcal{N}$. Hence, under this assumption we can combine P_k and E_k with the admissible i-pe/e combiner Π_h to get e-values $E_k^* = h(P_k)E_k$.

S3.1 The pe-BH procedure for false discovery rate control

The i-pe/p combiner Π_h and the e-BH procedure motivate the following procedure as an alternative to ep-BH:

Definition S3.1 (p-weighted e-BH procedure (pe-BH))**.** Choose $h \in \mathcal{C}^{\mathrm{p/e}}$. For $k \in \mathcal{K}$, compute $E_k^* = h(P_k)E_k$ by applying the i-pe/e merger Π_h, and then supply $(E_1^*, \dots, E_K^*)$ to the e-BH procedure at level α.

We immediately have the following result:

Theorem S3.2. *Suppose that Assumption 4.1 holds. Then, the pe-BH procedure has false discovery rate at most $\alpha K_0/K$.*

Proof. The theorem follows by combining the fact that $E_1^*, \dots, E_K^*$ are e-values for $H_1, \dots, H_K$ due to Theorem 3.1 and the false discovery rate guarantee $\alpha K_0/K$ of the e-BH procedure in Wang and Ramdas [2022, Theorem 2]. We emphasize that no dependence assumption on either $(P_k)_{k\in\mathcal{K}}$ or $(E_k)_{k\in\mathcal{K}}$ is required. □

We now contrast the pe-BH procedure to the ep-BH procedure. In the ep-BH procedure, e-values are used as weights for the p-values. Intuitively, if $E_k > 1$, then there is some evidence against H_k being a null, and we have $P_k/E_k < P_k$ (assuming $P_k \neq 0$); that is, the weight strengthens the signal of P_k. Conversely, if $E_k < 1$, then there is no evidence against H_k being a null, and we have $P_k/E_k > P_k$. The above interpretation of e-values as weights is quite natural. The situation for the pe-BH procedure, where p-values are used as weights for the e-values, is somewhat different. For simplicity, suppose that we use the calibrator $h \in \mathcal{C}^{\mathrm{p/e}}$ given by $h(p) = p^{-1/2} - 1$. It is clear that $h(p) > 1$ if and only if $p < 1/4$. Hence, the signal of the e-value E_k will be strengthened in case $P_k < 1/4$. This is not surprising as observing a p-value in $(0.25, 1)$ generally does not indicate evidence against the null. Other choices of $h \in \mathcal{C}^{\mathrm{p/e}}$ lead to different thresholds, and this is consistent with the fact that there is no universal agreement on which moderate values of a p-value should be considered as carrying some (weak) evidence against the null.

In terms of power, the ep-BH procedure dominates the pe-BH procedure when both are valid (that is, any hypothesis rejected by pe-BH will also be rejected by ep-BH). To show this, we proceed as follows: the e-BH procedure with input $(e_1, \dots, e_K)$ is equivalent to the p-BH procedure with input $(1/e_1, \dots, 1/e_K)$. Hence, the pe-BH procedure can be seen as applying the p-BH procedure to $(1/E_1^*, \dots, 1/E_K^*)$. If $h(p) > 1/p$ for even a single $p \in (0, 1)$, then h is not a p/e calibrator. Indeed, for $P \sim \mathrm{U}(0, 1)$, by decreasing monotonicity of h, we get $\mathbb{E}\{h(P)\} \geq \mathbb{E}\{h(P)\mathbb{1}(P < p)\} > \mathbb{E}\{\mathbb{1}(P < p)/p\} = \mathbb{P}(P < p)/p = 1$, which contradicts the fact that $h(P)$ is an e-value. Therefore, $p \mapsto 1/p$ is an upper bound for all p/e calibrators h, and this implies, for each $k \in \mathcal{K}$,

$$\frac{1}{E_k^*} = \frac{1}{h(P_k)E_k} \geq \frac{P_k}{E_k} = P_k^*.$$

Hence, the pe-BH procedure is dominated by the ep-BH procedure.

On the other hand, the ep-BH procedure requires some dependence assumption, such as positive regression dependence on a subset (Definition 2.2). By Theorem 4.7, if $(P_k)_{k\in\mathcal{K}}$ is positive regression dependent on a subset, then the ep-BH procedure is valid, and it should be the better choice than

the pe-BH procedure which is dominated. However, if there is no dependence information of $(P_k)_{k\in\mathcal{K}}$, then $(P_k^*)_{k\in\mathcal{K}}$ is arbitrarily dependent, and one may need to apply the p-BH procedure with the BY correction in Benjamini and Yekutieli [2001]. In this case, the pe-BH and the ep-BH procedures do not dominate each other. In particular, one needs to compare the inputs

$$\frac{1}{h(P_k)E_k} \quad \text{and} \quad \frac{\ell_K P_k}{E_k}, \quad \text{where } \ell_K := \sum_{k=1}^{K} \frac{1}{k} \approx \log K.$$

This is analogous to the trade-off between the p-BH procedure with BY correction and the e-BH procedure, where one compares $h(P_k)$ and $(\ell_K P_k)^{-1}$ [Wang and Ramdas, 2022, Section 6.5].

S3.2 A minimally adaptive e-BH procedure

The discussion of the pe-BH procedure (and its comparison to ep-BH) raises the following question: can we use the i-pe/e combiner Π_h within a procedure that controls the false discovery rate and is null-proportion adaptive (i.e., an analogous procedure to ep-Storey)? The challenge here is that null proportion adaptive procedures analogous to Storey's are not known for the e-BH procedure (and consequently for the pe-BH procedure).

In the remainder of this supplement we describe the first (minimally) adaptive procedure by proposing a tiny but uniform improvement of the e-BH procedure, inspired by Solari and Goeman [2017]. We remark that, similarly to the situation of Solari and Goeman [2017], this improvement is negligible for large values of K and it may only be practically interesting for small K such as $K \leq 10$. We mainly focus on the case without boosting; see Wang and Ramdas [2022] for e-value boosting.

First, choose an e-merging function $F : [0,\infty]^K \to [0,\infty]$ in the sense of Vovk and Wang [2021], i.e., F satisfies that $F(E_1,\dots,E_K)$ is an e-value for any e-values $E_1,\dots,E_K$. By Proposition 3.1 of Vovk and Wang [2021], the arithmetic average

$$M : (e_1,\dots,e_K) \mapsto \frac{1}{K}\sum_{k=1}^{K} e_k$$

is the "best" symmetric e-merging function, in the sense that it is uniformly more powerful than any other symmetric e-merging functions. We allow for a general choice of F other than M as it will be useful for the discussion later on boosted e-values.

With a chosen e-merging function F and a level $\alpha \in (0,1)$, the improved e-BH procedure, denoted by $\mathcal{D}^F(\alpha)$, is designed as follows. We first test the global null $\bigcap_{k=1}^{K} H_k$ via the rejection condition $F(e_1,\dots,e_K) \geq 1/\alpha$, which has a type-I error of at most α, and if the global null is rejected, we then apply the e-BH procedure at level $\alpha' = K\alpha/(K-1)$. In other words,

1. if $F(e_1,\dots,e_K) < 1/\alpha$, then $\mathcal{D}^F(\alpha) = \varnothing$;
2. if $F(e_1,\dots,e_K) \geq 1/\alpha$, then $\mathcal{D}^F(\alpha) = \mathcal{D}(\alpha')$ where $\alpha' = K\alpha/(K-1)$ and $\mathcal{D}(\alpha')$ is the e-BH procedure at level α'.

The next proposition shows that by choosing $F = M$, the resulting improved BH procedure dominates the base BH procedure.

Proposition S3.3. *The improved e-BH procedure $\mathcal{D}^F(\alpha)$ applied to arbitrary e-values has false discovery rate at most α. In case $F = M$, $\mathcal{D}^M(\alpha)$ dominates the e-BH procedure $\mathcal{D}(\alpha)$, that is, $\mathcal{D}(\alpha) \subseteq \mathcal{D}^M(\alpha)$.*

Proof. The first statement on false discovery rate can be shown in a similar way as Solari and Goeman [2017]. Let A be the the event that $F(e_1,\dots,e_K) \geq 1/\alpha$, treated as random. If $K_0 < K$, then, by

using Theorem 5.1 of Wang and Ramdas [2022],

$$\mathbb{E}\left\{\frac{F_{\mathcal{D}^F(\alpha)}}{R_{\mathcal{D}^F(\alpha)}}\right\} = \mathbb{E}\left\{\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\mathbb{1}(A)\right\} + \mathbb{E}\left[\frac{F_{\varnothing}}{R_{\varnothing}}\{1-\mathbb{1}(A)\}\right]$$
$$= \mathbb{E}\left\{\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\mathbb{1}(A)\right\} \le \mathbb{E}\left\{\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\right\} \le \frac{K_0}{K}\alpha' \le \alpha.$$

If $K_0 = K$, then the false discovery rate of $\mathcal{D}^F(\alpha)$ is at most the probability $\mathbb{P}(A)$ of rejecting the global null via $F(e_1, \dots, e_K) \ge 1/\alpha$. In this case, $\mathbb{P}(A) \le \alpha$ by Markov's inequality and the fact that F is an e-merging function. Hence, in either case, the FDR of $\mathcal{D}^F(\alpha)$ is at most α.

To show the second statement on dominance, let

$$S : (e_1, \dots, e_K) \mapsto \max_{k=1,\dots,K} \frac{k e_{[k]}}{K}. \tag{S3}$$

The function S is an e-merging function and it is dominated by M on $[0,\infty]^K$; see Section 6 of Vovk and Wang [2021]. Note that by definition, $S(e_1, \dots, e_K) < 1/\alpha$ implies $\mathcal{D}(\alpha) = \varnothing$. Therefore, if $M(e_1, \dots, e_K) < 1/\alpha$, then $\mathcal{D}(\alpha) = \varnothing = \mathcal{D}^M(\alpha)$. Moreover, since $\alpha < \alpha'$, we always have $\mathcal{D}(\alpha) \subseteq \mathcal{D}(\alpha')$. Hence, $\mathcal{D}(\alpha) \subseteq \mathcal{D}^M(\alpha)$. □

Next, we briefly discuss the case of boosted e-values. The arithmetic average of boosted e-values is not necessarily a valid e-value, so one must be a bit more careful. Nevertheless, it turns out that we can use the function S in (S3) on the boosted e-values. The new procedure can be described as the following steps.

1. Boost the raw e-values with level α.
2. If $S(e'_1, \dots, e'_K) < 1/\alpha$ where $e'_1, \dots, e'_K$ are the boosted e-values in step 1, then return $\varnothing$.
3. Else: boost the raw e-values with level $\alpha' = K\alpha/(K-1)$.
4. Return the discoveries by applying the base e-BH procedure to the boosted e-values in step 3.

This new procedure dominates the e-BH procedure, and it has FDR at most α. To show these two statements, it suffices to note that the probability of rejecting the global null test $S(e'_1, \dots, e'_K) \ge 1/\alpha$ is at most α since the e-BH procedure has FDR at most α by Theorem 5.1 of Wang and Ramdas [2022]; the rest of the proof is similar to that of Proposition S3.3.

S4 Using two samples for the one-sided z-test

S4.1 Setup

We first describe the setup in more detail and more generality compared to our treatment in Section 3.3. Suppose that we have two samples of iid data points, $X = (X_1, \dots, X_m)$ and $Y = (Y_1, \dots, Y_n)$, both from a distribution $\mathbb{P}$, where $m \ge 0$ and $n \ge 1$. Here, if $m = 0$ then X has no data. We would like to test $H_0 : \mathbb{P} = P$ against $H_1 : \mathbb{P} = Q$ where P and Q are distinct distributions. For illustration, we will focus on the simple case $P = \mathrm{N}(0,1)$ and $Q = \mathrm{N}(\delta, 1)$ where $\delta > 0$ is known.

For an observation $x \in \mathbb{R}$, the likelihood ratio of $\mathrm{N}(\delta, 1)$ over $\mathrm{N}(0,1)$ is

$$L_\delta(x) := \frac{\exp\{-(x-\delta)^2/2\}}{\exp(-x^2/2)} = \exp(\delta x - \delta^2/2).$$

The likelihood ratio based on the sample X is the e-value E_X for Q given by

$$E_X := \prod_{i=1}^m L_\delta(X_i) = \exp\left(\delta \sum_{i=1}^m X_i - \frac{m\delta^2}{2}\right),$$

which has a log-normal distribution under P with parameters $(\mu, \sigma^2) = (-m\delta^2/2, m\delta^2)$. Our convention is $E_X = 1$ if $m = 0$. In particular, the transformed log-likelihood statistic T_X defined by

$$T_X := \frac{1}{\delta\sqrt{m}}\left(\log E_X + \frac{m\delta^2}{2}\right) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m} X_i$$

has a standard normal distribution under P. Based on the statistic T_X, we can compute the likelihood ratio p-value

$$P_X := 1 - \Phi(T_X),$$

where Φ is the standard normal distribution function. Quantities like P_Y and E_Y are defined similarly with $(X_1, \ldots, X_m)$ replaced by $(Y_1, \ldots, Y_n)$.

We consider three possible approaches to test the hypothesis using the two samples.

(a) Combine two samples, that is, use $Z = (X, Y)$ and then compute the p-value $P_{\mathrm{LR}} := 1 - \Phi(T_Z)$ based on the likelihood ratio of Z, where

$$T_Z := \frac{1}{\sqrt{m+n}}\left(\sum_{i=1}^{m} X_i + \sum_{i=1}^{n} Y_i\right).$$

(b) Compute an e-value E_X from X and a p-value P_Y from Y, and use the Q-combiner to compute a p-value $P_{\mathrm{E}} := Q(P_Y, E_X) = (P_Y/E_X) \wedge 1$.[2]

(c) Compute a p-value P_X from X and another p-value P_Y from Y, and use the Fisher statistic $-2\log(P_Y P_X)$. The Fisher p-value is $P_{\mathrm{F}} := 1 - \chi_4(-2\log(P_Y P_X))$, where χ_4 is the the chi-square distribution with 4 degrees of freedom.

Among the three methods, our intuition is that P_{LR} should be the most powerful since it uses the full likelihood ratio of the sample. The other two methods, by combining two p-values or a p-value and an e-value, should lose some power.

S4.2 Pitman's asymptotic relative efficiency

We study Pitman's asymptotic relative efficiency (ARE; see Van der Vaart, 1998, Section 14.3) between the full likelihood ratio method and the P/E method. Note that if $n = 0$ then both methods are equivalent. We will consider that $m, n \to \infty$ and $m = \theta^2 n$ with the signal ratio $\theta \geq 0$ fixed. (More precisely, one can use $m = \lfloor \theta^2 n \rfloor$ which does not make a difference to the asymptotic analysis.)

Fix two levels α and β such that $1 > \beta > \alpha > 0$. Let $\mathbb{P}$ be the probability that generates the data from the alternative hypothesis, and Z' be a standard normal random variable under $\mathbb{P}$ independent of X. Define N_{LR} the number of sample points needed for a level-α test to reach power $\beta > \alpha$ under the alternative; that is, N_{LR} is the smallest number $n \in \mathbb{N}$ such that

$$\mathbb{P}(P_{\mathrm{LR}} \leq \alpha) \geq \beta; \quad \text{and equivalently, } \mathbb{P}\left\{Z' + \delta\sqrt{n\theta^2 + n} \geq \Phi^{-1}(1-\alpha)\right\} \geq \beta. \tag{S4}$$

Similarly, N_{E} is the smallest $n \in \mathbb{N}$ such that

$$\mathbb{P}(P_{\mathrm{E}} \leq \alpha) \geq \beta; \quad \text{and equivalently, } \mathbb{P}\left[Z' + \delta\sqrt{n} \geq \Phi^{-1}\{(1 - \alpha E_X)_+\}\right] \geq \beta. \tag{S5}$$

Pitman's asymptotic relative efficiency between P_{LR} and P_{E} is defined as

$$\mathrm{ARE}_\theta(\alpha, \beta) := \lim_{\delta \downarrow 0} \frac{N_{\mathrm{LR}}}{N_{\mathrm{E}}},$$

where we emphasize its reliance on θ. The asymptotic relative efficiency intuitively means the ratio of the needed sample size for P_{LR} to that for P_{E} when the signal is very small.

[2] In what follows we slightly abuse notation and often ignore the truncation of P_Y/E_X to 1. This does not make any difference with respect to the rejection of the null hypothesis.

Proposition S4.1. *For $1 > \beta > \alpha > 0$ and $\theta \geq 0$, we have*

$$\mathrm{ARE}_\theta(\alpha, \beta) = \frac{z^2}{k^2}, \tag{S6}$$

where $k > 0$ is the smallest solution to

$$\int_{-\infty}^{\infty} \Phi\left(\Phi^{-1}\left[\{1 - \alpha \exp(k^2\theta^2/2 + k\theta w)\}_+\right] - k\right) \mathrm{d}\Phi(w) = 1 - \beta, \tag{S7}$$

and $z > 0$ is given by

$$z = \frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta)}{\sqrt{1+\theta^2}}. \tag{S8}$$

In particular, $\mathrm{ARE}_0(\alpha, \beta) = 1$. Moreover, for fixed $\beta \in (0, 1)$ and $\theta \geq 0$, we have

$$\lim_{\alpha \downarrow 0} \mathrm{ARE}_\theta(\alpha, \beta) = 1. \tag{S9}$$

Proof. First, let β and α be fixed with $1 > \beta > \alpha > 0$. It is easy to see that both N_{LR} and N_{E} tend to infinity as $\delta \downarrow 0$. In this part of the proof, all convergence and asymptotic equivalence statements are with respect to $\delta \downarrow 0$.

The case of N_{LR} is easy to compute. By (S4), we have

$$\frac{N_{\mathrm{LR}}\delta^2}{z^2} \to 1 \quad \text{as } \delta \downarrow 0, \tag{S10}$$

where z is in (S8).

Next, we analyze N_{E}. Let $W_i = X_i - \delta$ for $i \in \mathbb{N}$ which is standard normally distributed under Q. Write $E_X(m)$ as E_X with sample size m. Note that

$$\log E_X(m) = \delta \sum_{i=1}^{m} W_i + \frac{m\delta^2}{2},$$

and therefore,

$$\log E_X(\theta^2 n) - \frac{n\theta^2\delta^2}{2} = \delta \sum_{i=1}^{\theta^2 n} W_i \sim \mathrm{N}(0, n\theta^2\delta^2). \tag{S11}$$

Let W be a standard normal random variable independent of Z'. We have $\log E_X(\theta^2 n)$ is identically distributed as $\delta\theta\sqrt{n}W + n\theta^2\delta^2/2$. Using (S11) and (S5), $N_{\mathrm{E}} \sim k^2/\delta^2$ where $k > 0$ is the smallest such that

$$\mathbb{P}\left(Z' + k \geq \Phi^{-1}\left[\{1 - \alpha \exp(k^2\theta^2/2 + k\theta W)\}_+\right]\right) = \beta.$$

By independence of Z' and W,

$$\mathbb{E}\left\{\Phi\left(\Phi^{-1}\left[\{1 - \alpha \exp(k^2\theta^2/2 + k\theta W)\}_+\right] - k\right)\right\} = 1 - \beta.$$

Equivalently, k is the smallest such that

$$\int_{-\infty}^{\infty} \Phi\left(\Phi^{-1}\left[\{1 - \alpha \exp(k^2\theta^2/2 + k\theta w)\}_+\right] - k\right) \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \,\mathrm{d}w = 1 - \beta. \tag{S12}$$

The equation (S12) is precisely (S7). Using $N_{\mathrm{E}} \sim k^2/\delta^2$ and $N_{\mathrm{LR}} \sim z^2/\delta^2$, we get $\mathrm{ARE}_\theta(\alpha, \beta) \sim z^2/k^2$.

Next, we prove (S9). In what follows, all convergence and asymptotic equivalence statements are with respect to $\alpha \downarrow 0$. We know that $k \to \infty$ since $k \geq z$ and $z \to \infty$ as $\alpha \downarrow 0$. For (S12) to hold, the term

$$f(\alpha) := \Phi^{-1}\left[\{1 - \alpha \exp(k^2\theta^2/2 + k\theta w)\}_+\right] - k$$

needs to be of the order $O(1)$ for some $w \in \mathbb{R}$. We claim that for this to happen, we need

$$R_\alpha := \frac{k^2(1+\theta^2)}{-2\log\alpha} \to 1, \quad \text{as } \alpha \downarrow 0.$$

Note that

$$f(\alpha) = \Phi^{-1}\left[\{1 - \alpha^{1-R_\alpha\theta^2/(1+\theta^2)} e^{k\theta w}\}_+\right] - k. \tag{S13}$$

Using the approximation $\Phi^{-1}(1-\epsilon)/\sqrt{-2\log\epsilon} \to 1$ as $\epsilon \downarrow 0$ (see Example 8.13 of DasGupta, 2008), we have

$$\Phi^{-1}\left[\{1 - \alpha^{1-R_\alpha\theta^2/(1+\theta^2)} e^{k\theta w}\}_+\right] \sim \sqrt{-2\left[\{1 - R_\alpha\theta^2/(1+\theta^2)\}\log\alpha + k\theta w\right]},$$

and by definition

$$k = \sqrt{\frac{R_\alpha(-2\log\alpha)}{1+\theta^2}}.$$

Putting the above two equations together, we get

$$\begin{aligned}\frac{\Phi^{-1}\left[\{1 - \alpha^{1-R_\alpha\theta^2/(1+\theta^2)} e^{k\theta w}\}_+\right]}{k} &\sim \sqrt{\frac{-2\left[\{1 - R_\alpha\theta^2/(1+\theta^2)\}\log\alpha\right](1+\theta^2)}{R_\alpha(-2\log\alpha)} - \frac{2\theta w}{k}} \\ &\sim \sqrt{\frac{1+\theta^2 - R_\alpha\theta^2}{R_\alpha}},\end{aligned}$$

and the above term is asymptotically equivalent to 1 if and only if $R_\alpha \to 1$. Since $k \to \infty$, if the above ratio is not 1 then $f(\alpha)$ tends to ∞ or $-\infty$ for every w, violating (S12). From this, we conclude that $R_\alpha \to 1$, and hence, $k^2 \sim -2\log\alpha/(1+\theta^2)$. Using the approximation $\Phi^{-1}(1-\epsilon)/\sqrt{-2\log\epsilon} \to 1$ again and (S8), we have $z^2 \sim -2\log\alpha/(1+\theta^2)$. Therefore, we obtain (S9). □

In Table S1 we report numerical values for $\mathrm{ARE}_\theta(\alpha,\beta)$ for some choices of $\theta \in [0,1]$ and α, β. For instance, for $\theta = 1$ and $(\alpha,\beta) = (0.05, 0.9)$, by using the P/E method compared to the full likelihood ratio, one at most loses 1 data point in every 8 data points. This remains true (as a conservative statement) for any value of $\theta \in [0,1]$, since $\mathrm{ARE}_\theta(\alpha,\beta) \geq \mathrm{ARE}_1(\alpha,\beta)$ for $\theta \in [0,1]$. If X has less signal than Y, then the ARE is even closer to 1. For instance, with the same $(\alpha,\beta) = (0.05, 0.9)$, $\mathrm{ARE}_{0.5}(0.05, 0.9) = 0.956$, meaning that one loses 1 data point in every 23 data points.

S5 Extensions to the setting of Section 5

S5.1 Simultaneous two-sample t-tests

In place of (5) , consider a two-sample situation in which we observe (independent) Y_{kj}, V_{kj} for $k = 1, \ldots, K$ and $j = 1, \ldots, n$, drawn as follows:[3]

$$Y_{kj} \sim \mathrm{N}(\mu_{Y,k},\, \sigma_k^2),\ V_{kj} \sim \mathrm{N}(\mu_{V,k},\, \sigma_k^2),\ \ \mu_{Y,k}, \mu_{V,k} \in \mathbb{R},\ \ \sigma_k > 0, \tag{S14}$$

[3]The assumption that the we observe the same number of observations, n, for each sample, is merely for notational convenience.

(α, β)	$(0.05, 0.5)$	$(0.01, 0.5)$	$(0.05, 0.9)$	$(0.01, 0.9)$
$\mathrm{ARE}_1(\alpha, \beta)$	0.956	0.974	0.874	0.914
$\mathrm{ARE}_{0.5}(\alpha, \beta)$	0.989	0.995	0.956	0.970

Table S1: Asymptotic relative efficiency: We numerically compute the values of Pitman's asymptotic relative efficiency (following Proposition S4.1) between the full likelihood ratio method and the P/E method for different choices of the size α, power β, and signal ratio θ. For instance, with $\theta = 0.5$, and $(\alpha, \beta) = (0.05, 0.9)$, $\mathrm{ARE}_{0.5}(0.05, 0.9) = 0.956$, meaning that one loses 1 data point in every 23 data points. We observe that the asymptotic relative efficiency is quite close to 1 and increases as α and θ decrease.

We seek to test $H_k : \mu_{Y,k} = \mu_{V,k}$. Let us compute the following,

$$\hat{\mu}_{Y,k} := \frac{1}{n}\sum_{j=1}^{n} Y_{kj}, \quad \hat{\mu}_{V,k} := \frac{1}{n}\sum_{j=1}^{n} V_{kj},$$

$$\hat{\sigma}^2_{Y,k} := \frac{1}{n-1}\sum_{j=1}^{n}(Y_{kj} - \hat{\mu}_{Y,k})^2, \quad \hat{\sigma}^2_{V,k} := \frac{1}{n-1}\sum_{j=1}^{n}(V_{kj} - \hat{\mu}_{V,k})^2,$$

$$T_k := \frac{\sqrt{n}(\hat{\mu}_{Y,k} - \hat{\mu}_{V,k})}{\sqrt{\hat{\sigma}^2_{Y,k} + \hat{\sigma}^2_{V,k}}}, \quad P_k := 2\{1 - F_{t,2n-2}(|T_k|)\},$$

where $F_{t,2n-2}$ is the cumulative distribution function of a random variable following the t-distribution with $2n-2$ degrees of freedom. This is the p-value of the standard equal variance two-sample t-test. Finally, let:

$$\hat{\mu}_k := \frac{1}{2}(\hat{\mu}_{Y,k} + \hat{\mu}_{V,k}), \quad S^2_k := \frac{1}{2n-1}\sum_{j=1}^{n}\{(Y_{kj} - \hat{\mu}_k)^2 + (V_{kj} - \hat{\mu}_k)^2\}.$$

Notice that $\hat{\mu}_k$ and S^2_k are the sample mean and sample variance after pooling all observations $Y_{k,1}, \ldots, Y_{k,n}, V_{k,1}, \ldots, V_{k,n}$ and ignoring their group assignment.

In analogy to (7) in the manuscript, we can then show that under the null (i.e., when $\mu_{Y,k} = \mu_{V,k}$), then P_k and S^2_k are independent. For further context and references, see, e.g., Ignatiadis and Huber [2021, Supplement S6.2.2].

The subsequent argumentation and methodological development could proceed analogously to the one-sample t-test problem that we study in the main text. We focus on the one-sample t-test instead of the two-sample t-test for the sake of simplicity, and notational compactness.

S5.2 Further extensions

Analogous constructions that lead to an independence statement under the null of the form (7) are available beyond the Gaussian models (5) and (S14). For example, suppose that for the k-th hypothesis we observe two samples, $Y_{k1}, \ldots, Y_{kn}$, and $V_{k1}, \ldots, V_{kn}$. We seek to conduct a nonparametric two-sample test. If $\{Y_{k1}, \ldots, Y_{kn}, V_{k1}, \ldots, V_{kn}\}$ are assumed to be exchangeable for $k \in \mathcal{N}$, then any permutation-invariant statistic is independent of the Wilcoxon rank sum statistic. See Bourgon et al. [2010] for further details.

S6 More general weighting functions in the setting of Section 5

Recall that our goal is to construct $\psi(\cdot)$ with the following two properties: it leads to more powerful e-values compared to $\psi(s^2) = s^2$, and second, a data-driven scaling analogous to (9) is practical and stable.

Our starting point is the likelihood ratio L_k of nS_k^2 under the noncentral chi-square distribution with n degrees of freedom and noncentrality parameter (ncp) λ and under the (central) chi-square distribution with n degrees of freedom.[4] We have the following expansion of L_k in terms of powers of S_k^2:

$$L_k = \sum_{d=0}^{\infty} \frac{\exp(-\lambda/2)\lambda^d \Gamma(n/2)}{4^d d! \Gamma(n/2+d)} (nS_k^2)^d.$$

Our proposal is to fix $D \in \mathbb{N}$ and to truncate the above power series to the first $D+1$ terms, i.e.,

$$\psi(S_k^2) \equiv L_k^D := \sum_{d=0}^{D} \frac{\exp(-\lambda/2)\lambda^d \Gamma(n/2)}{4^d d! \Gamma(n/2+d)} (nS_k^2)^d.$$

In our implementation, we take $D = 6$ and $\lambda = n$. Furthermore, let

$$\tilde{L}_k^D := \sum_{d=0}^{D} \frac{\exp(-\lambda/2)\lambda^d \Gamma\{(n-1)/2\}}{4^d d! \Gamma\{(n-1)/2+d\}} \{(n-1)\hat{\sigma}_k^2\}^d.$$

We may verify that for $k \in \mathcal{N}$, $\mathbb{E}(\tilde{L}_k^D) = \mathbb{E}(L_k^D)$. Hence this motivates the following choice of e-value rescaling analogous to (9):

$$E_k := K L_k^D \Big/ \sum_{\ell=1}^{K} \tilde{L}_k^D.$$

This is precisely (10).

The conclusions of Theorem 5.1 hold verbatim after replacing the event A_δ by the event:

$$A_\delta := \left\{ \sum_{k=1}^{K} \tilde{L}_k < (1-\delta) \sum_{k \in \mathcal{N}} \mathbb{E}(\tilde{L}_k) \right\}.$$

S7 E-values for replicated microarray data

We first provide a quick summary of the distributional assumptions and p-value constructions in Lönnstedt and Speed [2002] and Smyth [2004] and then derive analogous e-values. The starting point is that we seek to test K hypotheses $H_k : \beta_k = 0$ wherein for the k-th hypothesis we have summarized our data as $\widehat{\beta}_k, S_k^2$, where

$$\widehat{\beta}_k \mid \beta_k, \sigma_k^2 \sim \mathrm{N}(\beta_k,\, v_k \sigma_k^2), \quad S_k^2 \mid \sigma_k^2 \sim \frac{\sigma_k^2}{\nu_k} \chi_{\nu_k}^2. \tag{S15}$$

Above, v_k and ν_k are known fixed numbers and $\chi_{\nu_k}^2$ is the chi-square distribution with ν_k degrees of freedom. To be concrete, in case we conduct an equal variance two-sample t-test for each gene based on n_k^C control samples and n_k^T treated samples, then under standard normality assumptions we may take $v_k = (1/n_k^C + 1/n_k^T)$ and $\nu_k = n_k^T + n_k^C - 2$ in (S15).

[4] In fact, this likelihood ratio was the e-value we used in the simulation study of an earlier working paper of this work. The disadvantage of L_k is that it is unclear how to scale it as in (9) when $\sigma_k^2 > 0$ is unknown.

To share information across genes, Lönnstedt and Speed [2002], Smyth [2004] further posit the following distributional assumption on the residual variances σ_k^2:

$$\frac{1}{\sigma_k^2} \sim \frac{1}{\nu_0 s_0^2}\chi^2_{\nu_0}, \tag{S16}$$

where s_0^2, ν_0 are fixed numbers that determine the location and concentration of the distribution of the σ_k^2. Under (S15) and (S16), it also follows that,

$$\widetilde{T}_k \mid \beta_k = 0 \sim t_{\nu_0+\nu_k}, \quad \text{where } \widetilde{T}_k := \frac{\widehat{\beta}_k}{\tilde{S}_k\sqrt{v_k}}, \quad \tilde{S}_k^2 := \frac{\nu_0 s_0^2 + \nu_k S_k^2}{\nu_0+\nu_k}, \tag{S17}$$

where $t_{\nu_0+\nu_k}$ is the t-distribution with $\nu_0+\nu_k$ degrees of freedom with cumulative distribution function $F_{t,\nu_0+\nu_k}$. Hence, $P_k = 2\{1-F_{t,\nu_0+\nu_k}(|\widetilde{T}_k|)\}$ is a p-value for the null hypothesis $H_k : \beta_k = 0$. The upshot of positing (S16) is that we may studentize $\widehat{\beta}_k$ with sample variances that are shrunk toward s_0^2, and increase the degrees of freedom of the t-statistic from ν_k to $\nu_0+\nu_k$.[5] Model (S16) and the computation of the p-values entails knowledge of the two unknown parameters s_0^2, ν_0. The predominant approach in practice, as advocated by Smyth [2004], is to plug-in parametric empirical Bayes estimates $\widehat{\nu}_0$ and $\widehat{s}_0^2$ of these parameters (based on the data for all genes, $k = 1, \ldots, K$).

Taking inspiration from the above p-value construction, we propose the following e-value.

Proposition S7.1. *Suppose* (S15) *and* (S16) *hold. Then, for any* $\gamma > 0$,

$$E_k := \frac{1}{\sqrt{\gamma_k+1}}\left\{1 - \frac{\gamma_k \widetilde{T}_k^2}{(1+\gamma_k)(\nu_k+\nu_0+\widetilde{T}_k^2)}\right\}^{-\frac{\nu_0+\nu_k+1}{2}}, \ \gamma_k = \gamma/v_k, \tag{S18}$$

is an e-value for $H_k : \beta_k = 0$, *in particular,* $\mathbb{E}(E_k \mid \beta_k = 0) = 1$.

Proof. Let $p_{k,0}(\cdot)$ be the (marginal) density of the moderated t-statistic $\widetilde{T}_k$ (S17) under (S15), (S16), and $\beta_k = 0$. For $\gamma > 0$, let $p_{k,\gamma}(\cdot)$ be the (marginal) density of the moderated t-statistic $\widetilde{T}_k$ when $\beta_k \mid \sigma_k^2 \sim \mathrm{N}(0, \gamma\sigma_k^2)$ and (S15), (S16) hold. Then E_k in (S18) is equal to the likelihood ratio $p_{k,\gamma}(\widetilde{T}_k)/p_{k,0}(\widetilde{T}_k)$. Hence:

$$\mathbb{E}(E_k \mid \beta_k = 0) = \int \{p_{k,\gamma}(t)/p_{k,0}(t)\}p_{k,0}(t)dt = \int p_{k,\gamma}(t)dt = 1.$$

□

The e-value construction above requires a choice of a tuning parameter $\gamma > 0$. The proof above hints at a way of choosing γ in a data-driven way. We make the additional working model assumption:

$$\beta_k \mid \sigma_k^2 \sim \tilde{\pi}_0\delta_0 + (1-\tilde{\pi}_0)\mathrm{N}(0, \gamma\sigma_k^2), \quad \tilde{\pi}_0 = 1/2, \tag{S19}$$

where δ_0 is a point mass at 0. We then estimate γ by empirical Bayes as described in Smyth [2004, Section 6.3] by positing that (S19) holds for all genes $k = 1 \ldots, K$ in addition to (S15) and (S16). Analogously to the computation of p-values in `limma` [Smyth, 2004], we ignore uncertainty introduced due to the estimation of γ.

The choice $\tilde{\pi}_0 = 1/2$ in (S19) is a conservative choice. If we were to further increase the proportion assigned to the null component ($\tilde{\pi}_0$), then the estimated γ would typically be larger, and this would lead to more extreme e-values. Instead, we make the safe choice $\tilde{\pi}_0 = 1/2$ in anticipation of the

[5] One may wonder if the additional assumption (S16) is justified. For microarray data and RNA-Seq data analyzed via `limma` [Ritchie et al., 2015], (S16) often provides an adequate fit with respect to downstream inferences [Lu and Stephens, 2016]. Lu and Stephens [2016] and Ignatiadis and Sen [2023] replace (S16) by a nonparametric generalization.

downstream task of combining e-values with p-values. In particular, we emphasize, that inferences will be valid even if the true null proportion π_0 is different than our posited $\tilde{\pi}_0$.

The e-value E_k has the following elegant interpretation for $\tilde{\pi}_0 = 1/2$: it is equal to the posterior odds statistic proposed by Lönnstedt and Speed [2002, Equation 3]. The posterior odds statistic relies on the validity of (S19), and this led Lönnstedt and Speed [2002] to write that "we cannot rely on any standard cutoff value [...] for the selection of differentially expressed genes." However, since the posterior odds statistic is an e-value, we no longer need to rely on (S19) (it is merely a working assumption), and we can rigorously proceed with the multiple testing task.

S8 Additional simulation figures

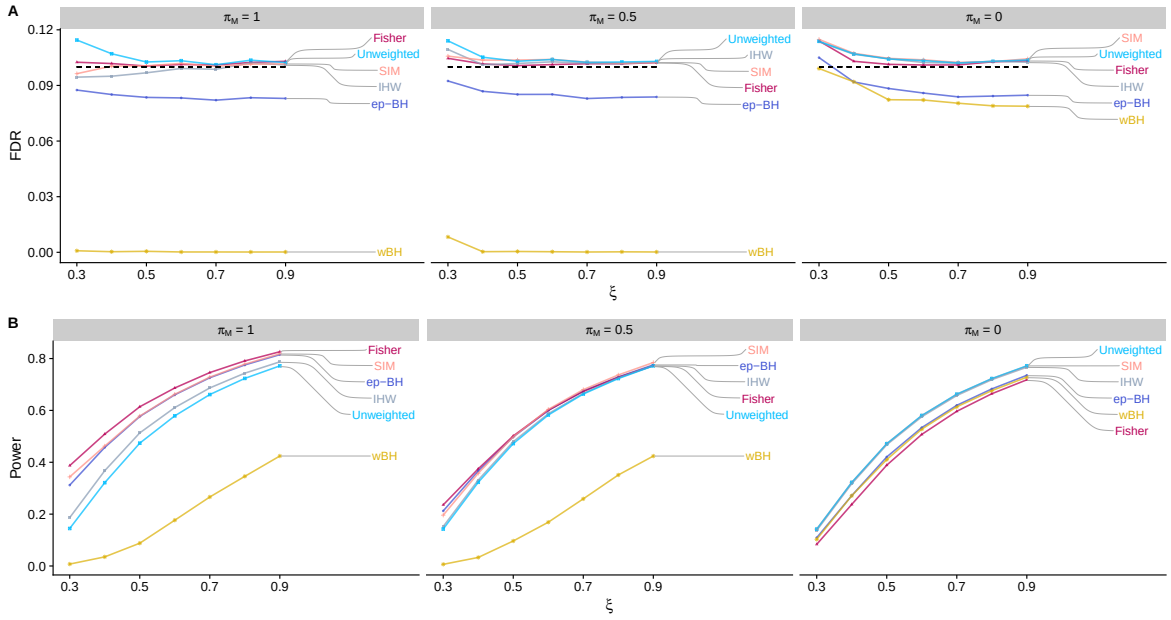

Figure S1: RNA-Seq and microarray meta-analysis simulation with null-proportion adaptive methods: This figure is analogous to Fig. 5 with the difference that we compare null-proportion adaptive variants of the same methods. We plot A) the false discovery rate (FDR) and B) power against the effect size parameter ξ and against the informativeness of the microarray data (parameter π_M in the facets). We note that in this case there is slight exceedance of FDR control for several methods (including unweighted p-BH) at small values of ξ (which is also slightly visible—but less so—in Fig. 5). The reason may be that `DESeq2` [Love et al., 2014] p-values are computed based on asymptotic approximations, and so may not be exactly super-uniform in finite samples. The remaining takeaways are similar to those of Fig. 5: Fisher Storey-BH has the most power when microarray data are fully informative ($\pi_M = 1$), but has the least power when the microarray data are fully uninformative ($\pi_M = 0$).