Supplementary Material to: Multiple competition-based FDR control and its application to peptide detection

Kristen Emery[1], Syamand Hasam[1], William Stafford Noble[2 [0000−−0001−7283−4715]], and Uri Keich[1 [0000−−0002−3209−5011]]

[1] School of Mathematics and Statistics F07
University of Sydney
{keme6477@,shas8442@uni.,uri.keich@}sydney.edu.au
[2] Department of Genome Sciences and of Computer Science and Engineering
University of Washington
william−noble@uw.edu

6 Supplementary Material

6.1 Controlling the FDR in the multiple hypotheses testing (MHT) setup

In the multiple hypotheses testing (MHT) set up we simultaneously test m (null) hypotheses $H_1, \ldots, H_m$, looking to reject as many as possible subject to some statistical control of our error rate. The rejected hypotheses are typically referred to as "discoveries", and H_i corresponds to a false discovery if it is rejected but is in fact a true null hypothesis.

Pioneered by Benjamini and Hochberg, the common approach to controlling the error rate in the MHT context is through bounding the expected proportion of false discoveries at any desired level $\alpha \in (0, 1)$. More precisely, assume we have a selection procedure that produces a list of R discoveries of which, unknown to us, V are false. Let $Q = V/\max\{R, 1\}$ be the unobserved false discovery proportion (FDP). Benjamini and Hochberg showed that applying their selection procedure (BH) at level α controls the expected value of Q at that level: $E(Q) \leq \alpha$ [2]. They referred to $E(Q)$ as the false discovery rate (FDR), and their precise statement is that, provided the true null p-values are distributed as independent uniform $U(0, 1)$ random variables, $E(Q) \leq \alpha$ for any p-values the true alternative / false null hypotheses assume. Other, more powerful selection procedures that rely on estimating π_0, the fraction of true null hypotheses, are available. Generally referred to as "adaptive BH" procedures, with one particularly popular variant by Storey, these procedures are also predicated on our ability to assign a p-value to each of our tested hypotheses (e.g., [3, 4, 13, 14]). Hence, in particular they cannot be directly applied in our competition-based setup.

6.2 The problem with pooling the decoys

Two significant problems arise when pooling the decoys to compute the p-values. First, these p-values do not satisfy the assumption that the p-values of the true null hypotheses are independent: because all p-values are computed using the same batch of pooled decoy scores, it is clear that they are dependent to some extent. While this dependency

diminishes as $m \to \infty$, there is a second, more serious problem that in general cannot be alleviated by considering a large enough m.

Specifically, in pooling the decoys we make the implicit assumption that the score is calibrated, i.e., that all true null scores are generated according to the same distribution. If this assumption is violated, as is typically the case in the spectrum identification problem for one [7], then the p-values of the true null hypotheses are not identically distributed, and in particular they are also not (discrete) uniform in general. This means that even the more conservative BH procedure is no longer guaranteed to control the FDR, and the problem is much worse with Storey. Indeed, below we show that there are arbitrary large examples wherein Storey significantly fails to control the FDR, and similar ones where BH is essentially powerless. Those examples, demonstrate that, in general, applying BH or Storey's procedure to p-values that are estimated by pooling the competing null scores can be problematic both in terms of power and control of the FDR. Note that these issues have previously been discussed in the context of the spectrum identification problem, where the effect of pooling on power and on FDR control were demonstrated using simulated and real data [7, 6].

Examples of failings of the canonical procedures Consider BH applied to just $m = 2$ hypotheses with $d = 1$ decoy, and suppose that $P\left(\tilde{Z}_1^1 > \tilde{Z}_2^1\right) = 1$ (i.e., the support of the null distribution corresponding to H_1 is disjoint and to the right of the support of the null distribution corresponding to H_2). Suppose further that both H_i are true nulls, so that the FDR coincides with the FWER (family-wise error rate), which is the probability of at least one (false) discovery. It is easy to see that in this case using the FDR threshold $\alpha \coloneqq 2/3$ the event $\left\{Z_1 > \tilde{Z}_1^1, Z_2 < \tilde{Z}_2^1\right\}$ will produce one discovery ($p_1 \coloneqq \text{p-value}\,(Z_1) = 1/3, p_2 \coloneqq \text{p-value}\,(Z_2) = 1$), and the disjoint event $\left\{Z_2 > \tilde{Z}_2^1\right\}$ will produce two discoveries ($p_2 = 2/3$). However, these events have a total probability of $1/4 + 1/2 = 3/4$ so the FWER=FDR is $> \alpha$ in this case.

This effect can be much more pronounced in the case of Storey's method. Suppose that the null hypotheses split into two equal sized groups, A and B, where for every $i \in A$ and $j \in B$, $P\left(\tilde{Z}_i^1 < \tilde{Z}_j^1\right) = 1$. Suppose further that all the hypotheses in A are false nulls with scores Z_i satisfying $P\left(Z_i > \tilde{Z}_i^1\right) = 1$, and that all hypotheses in B are true nulls. The decoy-pooled p-values will be essentially no greater than 1/2. Hence, Storey's estimate of π_0, $\hat{\pi}_0(\lambda) = \frac{m - R(\lambda)}{(1-\lambda)m}$, where m is the number of hypotheses, and $R(\lambda)$ is the number of hypotheses whose p-value is $\leq \lambda$, will significantly underestimate π_0. For example, if $\lambda \geq 0.5$ then $\hat{\pi}_0 = 0$, which in turn implies that essentially all null hypotheses will be rejected at any FDR level α and particularly for $\alpha < 1/2$, while the actual FDP will clearly be $1/2$. Even if λ is chosen to better fit these p-values, e.g., $\lambda = 0.25$, or the set-up is changed slightly to allow some group A p-values to be null so $\hat{\pi}_0 \neq 0$, the procedure will still significantly underestimate π_0 and thus underestimate the actual FDR.

Example 1. As a specific example in the above vein we constructed an experiment with $m = 300$ and $d = 5$ decoys where group A's true null distribution is $N(0, 1)$**,** and

group B's true null distribution is $N(50,1)$. We set all 150 hypotheses in group B and 50 of the 150 hypotheses in group A to be true null, and we generated observed scores by sampling from the appropriate null distribution above. We next generated the observed scores for the 100 false null hypotheses in group A by sampling from the same, significantly shifted, $N(50,1)$ distribution that we used to generate all observed scores of group B. All competing null (decoy) scores were generated using the group's null distribution. In this setup we chose to leave a third of group A as true nulls so that approximately 50 of the p-values will exceed $1/2$ ensuring that $\hat{\pi}_0 > 0$.

We then computed the pooled p-values and applied Storey's FDR controlling procedure, as presented in the package `qvalue` [15] (with λ chosen using the `bootstrap` option). This experiment was repeated using 1,000 randomly drawn sets, noting each time the real FDP at FDR thresholds of $\alpha = 0.1$ and $\alpha = 0.2$. As expected in this setting, Storey's procedure clearly failed to control the FDR: at $\alpha = 0.1$, the empirical FDR (the FDP averaged over the 1K samples) was 0.24, or over 200% of what it should be, and for $\alpha = 0.2$ the empirical FDR was larger than 0.5 again indicating a significant violation.

We could not find such examples, with an essentially arbitrary large m and a substantial liberal bias, when using the BH procedure. However, we found a class of arbitrary large examples, similar to the above class (on which Storey fails to control the FDR), where due to pooling the conservative nature of BH was amplified to the point where it was essentially powerless. Consider four groups A, B, C and D and suppose that for every $i \in A$, $j \in B$, $k \in C$ and $l \in D$, $P\left(\tilde{Z}_i^1 < \tilde{Z}_j^1 < \tilde{Z}_k^1 < \tilde{Z}_l^1\right) = 1$. Suppose further that all the hypotheses in groups A and B are false null with scores that fall in the range of values of the subsequent group, so in particular $P\left(Z_i > \tilde{Z}_k^1\right) = 0$ and similarly $P\left(Z_j > \tilde{Z}_l^1\right) = 0$. It is easy to see that using pooling in this case the p-values for the (false) null hypotheses in groups A and B will be $\geq 1/2$ and $1/4$ respectively, and it follows that no discoveries can be made by BH with $\alpha < 1/4$, regardless of how large m and d are.

Example 2. Again, we construct a specific example according to the above general outline. We set $m = 300$, so that each of the four groups has 75 hypotheses, and we use $d = 5$ decoys. The null distribution of each group is set as $N(\mu,1)$, where μ increases from $\mu_A = 0$, by 50, to $\mu_D = 150$. The observed scores corresponding to the 150 false null hypotheses of groups A and B were drawn from the null distributions of group B and C respectively, whereas the 150 observed scores of groups C and D were drawn from their respective null distributions. Using pooled p-values BH does not yield any discovery for any $\alpha \leq 0.65$ amongst any of our 1000 samples, and it was not until using $\alpha = 0.7$ that we finally started seeing some samples on which BH had non-zero power. Incidentally, even using non-pooled p-values is slightly better here: the first samples with non-zero BH power appear for $\alpha = 0.3$.

6.3 Simulation setup

In order to analyze and compare the performance of the FDR-controlling procedures we simulated datasets with both calibrated (all true null scores are generated according

to the same distribution) and non-calibrated scores — a comparison that also allowed us to select our overall recommended procedure.

In the non-calibrated case we allow the distribution of the null scores to vary with the hypotheses so we sample from hypothesis-specific distributions. Specifically, for simulating using a non-calibrated score we associate with the null hypothesis H_i a normal $N(\mu_i, \sigma_i^2)$ distribution from which its competing null (decoy/knockoff) scores are sampled. If H_i is labeled a true null, this is also the distribution from which the observed score is sampled. Otherwise, H_i is a false null, so the observed score is sampled from a γ_i-shifted normal $N(\mu_i + \gamma_i, \sigma_i^2)$ distribution, where $\gamma_i > 0$. The parameters μ_i, σ_i^2, and γ_i are themselves sampled with each newly sampled set of scores:

- μ_i is sampled from a normal $N(\mu, \sigma^2)$ distribution with the hyper-parameters $\mu = 0$ and $\sigma^2 = 1$.
- σ_i^2 is sampled from $1 + \exp(\omega)$, where $\exp(\omega)$ is the exponential distribution with rate $\omega = 1$.
- γ_i is sampled from $1 + \exp(\nu)$, where ν is a hyper-parameter that determines the separation between the false and true null scores

When simulating using a calibrated score the parameters μ_i, σ_i and γ_i are kept constant.

In our non-calibrated score simulations we drew 10K random sets of observed and competing null scores (each with its own randomly drawn values of $\mu_i, \sigma_i, \gamma_i$) for each of the following 600 combinations of parameter (or hyper-parameter) values:

- The number of false null hypotheses, k, was set to each value in $\left\{1, 10, 10^2, 10^3, 10^4\right\}$.
- For each value of k, the total number of hypotheses, m, was set to $\min\{c \cdot k, 2 \cdot 10^4\}$ where c was set to each of the following factors $\{1.25, 2, 4, 10, 20, 100, 1000\}$ subject to the constraint that $m \geq 100$.
- For each values of k and m above, the hyper-parameter ν that determines the separation between the false and true null scores was set to each of the values in $\{0.01, 0.05, 0.1, 0.25, 0.5, 1.0\}$.
- For each values of k, m, and ν above, the number of decoys d was set to each of the values in $\{3, 5, 9, 19, 39\}$.

We then used the 10K sampled sets from each of the 600 experiments to find the empirical FDR as well as the power of each method for each selected FDR threshold $\alpha \in \Phi$. For a given threshold α, the power of a method is the average percentage of false nulls that are reported by the method at level α, and the empirical FDR is the average of the FDP in the discovery list (both averages are taken over the experiment's 10K runs). We used a fairly dense set of FDR thresholds Φ: from 0.001 to 0.009 by jumps of 0.001, from 0.01 to 0.29 by jumps of 0.01, and from 0.3 to 0.95 by jumps of 0.05.

Our calibrated score simulation also consisted of 600 experiments, or combinations of parameter values. Specifically, we used the same values of k, m, and d as in the above non-calibrated simulations and we let γ vary over the values in $\{0.8, 1, 1.4, 2, 3, 4\}$. In each experiment we again draw 10K random sets of observed and competing scores using $\mu_i \equiv 0, \sigma_i \equiv 1, \gamma_i \equiv \gamma$.

In both setups we examined the FDR control by looking at the ratio between the empirical FDR (the observed FDP averaged over 10K runs) and the selected threshold as well as at the power which is the average (over 10K runs) percentage of false nulls we discover.

6.4 The correspondence between the tuning parameters of Storey's procedure and AS's

Lei and Fithian pointed out the connection between the (c, λ) parameters of their AS procedure (they refer to c as s) and the corresponding parameters in Storey's procedure. Specifically, AS's λ is analogous to the parameter λ of [14] that determines the interval $(\lambda, 1]$ from which π_0, the fraction of true null hypotheses is estimated. Specifically, in the finite sample case, π_0 is estimated as:

$$\hat{\pi}_0^*(\lambda) = \frac{m - R(\lambda) + 1}{(1-\lambda)m}, \tag{7}$$

where m is again the number of hypotheses, and $R(\lambda)$ is the number of discoveries at threshold λ (number of hypotheses whose p-value is $\leq \lambda$). The c parameter is analogous to the threshold

$$t_\alpha\left(\widehat{\mathrm{FDR}}_\lambda^*\right) = \sup\left\{t \in [0,1] \,:\, \widehat{\mathrm{FDR}}_\lambda^* \leq \alpha\right\} \tag{8}$$

of [14], where

$$\widehat{\mathrm{FDR}}_\lambda^* = \begin{cases} \frac{m \cdot \hat{\pi}_0^*(\lambda) \cdot t}{R(t) \vee 1} & t \leq \lambda \\ 1 & t > \lambda \end{cases}. \tag{9}$$

6.5 Determining λ from the empirical p-values

Given an upper bound Λ on λ (we used 0.95), and a binomial test significance cutoff β (we used 0.1)

1. Initialize: $i \coloneqq 1$.
2. If $i \geq \Lambda \cdot d_1$ or $i = d$ then
   - set $i_\lambda \coloneqq i$ and stop.
3. If $i + d_1$ is even then
   - set $i_s \coloneqq (i + d_1)/2$,
   - $n_p^+ \coloneqq \#\left\{\tilde{p}_i \in [(i_s+1)/d_1, 1]\right\}$,
   - $n_p^- \coloneqq \#\left\{\tilde{p}_i \in [(i+1)/d_1, i_s/d_1]\right\}$.
4. Otherwise,
   - set $i_s \coloneqq (i + d_1 + 1)/2$,
   - $n_p^+ \coloneqq \#\left\{\tilde{p}_i \in [(i_s+1)/d_1, 1]\right\}$,
   - $n_p^- \coloneqq \#\left\{\tilde{p}_i \in [(i+1)/d_1, (i_s-1)/d_1]\right\}$.
5. Calculate $p_b = P(B \geq n_p^-)$ where $B \sim Binomial(n_p^+ + n_p^-, 0.5)$.
6. If $p_b > \beta$ then (the remaining tail of the p-value histogram "seems to have flattened")
   - set $i_\lambda \coloneqq i$ and stop.
7. Otherwise (we are yet to see the flattening of the tail of the p-value histogram),
   - set $i \coloneqq i + 1$,
   - return to step 2.

Note that the interval $(\lambda, 1]$ from which π_0 is estimated in (7) coincides with $[(i_\lambda + 1)/d_1, 1]$.

6.6 Finite-decoy Storey (FDS and FDS$_1$)

The procedure we call finite-decoy Storey (FDS) starts with determining λ as above (Supplementary Section 6.5). Then, given the FDR threshold $\alpha \in (0, 1)$, FDS proceeds along (7)-(9) using $R(\lambda) = |\{\tilde{p}_i : \tilde{p}_i \leq \lambda\}|$, to determine

$$t_\alpha\left(\widehat{\mathrm{FDR}}^*_\lambda\right) = \max\left\{i \in \{0, 1, \ldots, d_1 \cdot \lambda\} : \frac{m \cdot \hat{\pi}^*_0(\lambda) \cdot i/d_1}{R\left(i/d_1\right) \vee 1} \leq \alpha\right\}. \tag{10}$$

This in principle is our threshold c except that, especially when d is small, $t_\alpha\left(\widehat{\mathrm{FDR}}^*_\lambda\right)$ can often be zero which is not a valid value for c in our setup. Hence FDS defines

$$c \coloneqq \max\left\{1/d_1, t_\alpha\left(\widehat{\mathrm{FDR}}^*_\lambda\right)\right\}. \tag{11}$$

With (c, λ) determined, FDS continues by applying the mirandom procedure with the chosen parameter values.

FDS was defined as close as possible to Storey, Taylor and Siegmund's recommended procedure for guaranteed FDR control in the finite setting (albeit with a predetermined λ). We found that a variant of FDS that we denote as FDS$_1$, often yields better power. FDS$_1$ differs from FDS as follows:

- When computing $t_\alpha\left(\widehat{\mathrm{FDR}}^*_\lambda\right)$ (10) we use Storey's large sample formulation which does not include the $+1$ in the estimator $\hat{\pi}^*_0(\lambda)$ (7), and maximizes over $i \in \{0, 1, \ldots, d_1\}$.
- Instead of defining c as in (11), FDS$_1$ defines $c \coloneqq \min\left\{c_{\max}, 1/d_1 + t_\alpha\left(\widehat{\mathrm{FDR}}^*_\lambda\right)\right\}$, where $c_{\max}$ is some hard bound on c (we used $c_{\max} = 0.95$).
- With FDS$_1$'s tweaked definition of c the case $c > \lambda$ is now possible. However, mirandom does not allow this so in that case FDS$_1$ sets $\lambda \coloneqq c$ instead of the value of λ defined in Supplementary Section 6.5).

6.7 The limiting behavior of our FDR controlling methods

In its selection of the parameter c, FDS essentially applies Storey's procedure to the empirical p-values; however, there is a more intimate connection between FDS, and more generally between some of the methods described above and Storey's procedure that becomes clearer once we let $d \to \infty$. To elucidate that connection we further need to assume here that the score is calibrated, that is, that the distribution of the decoy scores is the same for all null hypotheses. In this case, we might as well assume our observed scores are already expressed as p-values: $Z_i = p_i$ (keeping in mind that this implies that small scores are better, not worse as they are elsewhere in this paper).

It is not difficult to see that for a given (c, λ), mirandom assigns, in the limit as $d \to \infty$, $W_i \coloneqq Z_i = p_i$ if $p_i \leq c$ ($L_i = 1$, or original win), and $W_i \coloneqq (1 - p_i) \cdot c/(1 - \lambda) \in [0, c)$ if $p_i > \lambda$ ($L_i = -1$, or decoy win). Sorting the scores W_i in increasing order (smaller scores are better here) $W_{(1)} < W_{(2)} < \cdots < W_{(m)}$, we note that for i with $W_{(i)} = p_{(i)} \leq c$ the denominator term $\#\left\{j \leq i : L_{(j)} = 1\right\}$ in (3) is the number of original scores, or p-values $p_j \leq p_{(i)}$. At the same time, for the same i

and $j \leq i$, $L_{(j)} = -1$ if and only if $p_{(j)} > \lambda$ and $W_{(j)} < W_{(i)} = p_{(i)} \leq c$ so we have for the numerator term

$$\#\left\{j \leq i \,:\, L_{(j)} = -1\right\} = \#\left\{j \,:\, p_j > \lambda, W_j < p_{(i)}\right\} = \#\left\{j \,:\, p_j > 1 - \frac{1-\lambda}{c} p_{(i)}\right\}.$$

Considering that $i_{\alpha c \lambda} < m$ in (3) must be attained at an i for which $W_i = p_i \leq c$ (original win), we can essentially rewrite (3) as

$$i_{\alpha c \lambda} = \max\left\{i \,:\, \frac{1 + \#\left\{j \,:\, p_j > 1 - \frac{1-\lambda}{c} p_{(i)}\right\}}{\#\left\{j \,:\, p_j \leq p_{(i)}\right\} \vee 1} \cdot \frac{c}{1-\lambda} \leq \alpha\right\}. \tag{12}$$

Consider now Storey's selection of the threshold t_α, which when using the more rigorous estimate (7) essentially amounts to

$$t_\alpha = \max\left\{t \in [0, \lambda^*] \,:\, \frac{1 + \#\left\{j \,:\, p_j > \lambda^*\right\}}{\#\left\{j \,:\, p_j \leq t\right\} \vee 1} \cdot \frac{t}{1-\lambda^*} \leq \alpha\right\},$$

where λ^* is a tuning parameter. Considering the cases where $t_\alpha \leq c$ and setting $\lambda^*(t) \coloneqq 1 - (1-\lambda)\, t/c$ Storey's threshold t_α becomes

$$t_\alpha = \max\left\{t \in [0, c] \,:\, \frac{1 + \#\left\{j \,:\, p_j > 1 - \frac{1-\lambda}{c} t\right\}}{\#\left\{j \,:\, p_j \leq t\right\} \vee 1} \cdot \frac{c}{1-\lambda} \leq \alpha\right\}.$$

Since in practice t_α can be taken as equal to one of the $p_{(i)}$ the equivalence with (12) becomes obvious by identifying t above with $p_{(i)}$ in (12).

Thus, for example, as $d \to \infty$ the mirror method ($\lambda = c = 1/2$) converges, in the context of a calibrated score, to Storey's procedure using $\lambda^*(t) \coloneqq 1 - t$, which coincides with the "mirroring method" of [16]. It is worth noting that the general setting $\lambda^*(t) \coloneqq 1 - (1-\lambda)\, t/c$ is not obviously supported by the finite sample theory of [14]; however, it can be justified by noting the above equivalence and our results here.

An even more direct connection with Storey's procedure is established by letting $d \to \infty$ in the context of FDS. Indeed, using the same λ determined by the progressive interval splitting procedure described in Section 6.5, Storey's finite-sample procedure (8) would amount to setting the rejection threshold to

$$t_\alpha = \max\left\{t \in [0, \lambda] \,:\, \frac{m \cdot \hat{\pi}_0^*(\lambda) \cdot t}{R(t) \vee 1} \leq \alpha\right\}.$$

Recalling that $d_1 \to \infty$ we note that the latter t_α coincides with the value FDS assigns to c via (10) and (11). Let i_c be such that the above $t_\alpha = c \in \left[p_{(i_c)}, p_{(i_c+1)}\right)$ (recall we assume no ties here), then we can assume without loss of generality that $t_\alpha = c = p_{(i_c)}$ and hence (compare with (12)) the mirandom part of FDS will find

$$i_{\alpha c \lambda} = \max\left\{i \leq i_c \,:\, \frac{1 + \#\left\{j \,:\, p_j > 1 - \frac{1-\lambda}{c} p_{(i)}\right\}}{\#\left\{j \,:\, p_j \leq p_{(i)}\right\} \vee 1} \cdot \frac{c}{1-\lambda} \leq \alpha\right\}.$$

But

$$\frac{1+\#\left\{j:\, p_j > 1-\frac{1-\lambda}{c}p_{(i_c)}\right\}}{\#\left\{j:\, p_j \le p_{(i_c)}\right\}\vee 1}\cdot\frac{c}{1-\lambda} = \frac{1+\#\left\{j:\, p_j>\lambda\right\}}{\#\left\{j:\, p_j\le c\right\}\vee 1}\cdot\frac{c}{1-\lambda} = \frac{m\cdot\hat{\pi}_0^*(\lambda)\cdot t_\alpha}{R(t_\alpha)\vee 1}\le\alpha.$$

Hence i_c satisfies the required inequality and $i_{\alpha c\lambda} = i_c$. It follows that the rejection lists of FDS and the above variant of Storey's procedure with the same λ coincide in the $d\to\infty$ limit.

6.8 Labeled resampling

For clarity of the exposition we break the description of our labeled resampling procedure into two parts with the first describing how we generate a sample of conjectured true/false null labels.

1. Determine λ as described in Supplementary Section 6.5
2. Using $c = \lambda$ from step 1 above, apply steps 1-2 of mirandom (Section 3.2 with $\varphi\equiv\varphi_{md}$ of Section 3.5) to define the assigned scores W_i and labels L_i, and order the hypotheses in a decreasing order of W_i
3. Initialize by setting:
   - $j\coloneqq 1$ (j is the index of the set of hypotheses we currently consider)
   - $i_1\coloneqq 1$, $i_0\coloneqq 0$ (i_j is the number of hypotheses in $\mathcal{H}_j$)
   - $l\coloneqq 0$ (l denotes the index of last drawn conjectured false null)
   - $\boldsymbol{f}\coloneqq(0,0,\ldots,0)$ ($\boldsymbol{f}_i$ is the indicator of whether or not we conjecture H_i is a false null)
4. Estimate a_j, the number of false null hypotheses in $\mathcal{H}_j=\{H_i\,:\,i\le i_j\}$, as

$$a_j\coloneqq\left(\#\left\{i\le i_j\,:\,L_i=1\right\}-\#\left\{i\le i_j\,:\,L_i=-1\right\}\cdot\frac{\lambda}{1-\lambda}\right)\vee 0.$$

   Note that the first term is the number of original wins among the hypotheses in $\mathcal{H}_j$ and the second is essentially the numerator of (3) (with $c=\lambda$), which uses the number of decoy wins to estimate the number of false discoveries among those original wins.
5. If $a_j > \|\boldsymbol{f}\|_1$ (the number of conjectured false nulls drawn so far) then draw $a_j - \|\boldsymbol{f}\|_1$ additional conjectured false nulls as follows:
   (a) for each $i\in\{l+1,l+2,\ldots,i_j\}$ let $w_i\coloneqq 1-\tilde{p}_i$, where $\tilde{p}_i$ are the empirical p-values
   (b) while $a_j-\|\boldsymbol{f}\|_1>0$:
      i. draw an index $i\in\{l+1,\ldots,i_j\}$ according to the categorical distribution with a probability mass function proportional to w_i
      ii. set $\boldsymbol{f}_i\coloneqq 1$ and $w_i\coloneqq 0$
6. If $i_j=m$ return the conjectured labels $\boldsymbol{f}$, else continue
7. Set $\delta_{j+1}\coloneqq i_j-i_{j-1}+1$ if no new conjectured false null were drawn in step 5, otherwise set $\delta_{j+1}\coloneqq i_j-i_{j-1}$
8. Set $i_{j+1}\coloneqq(i_j+\delta_{j+1})\wedge m$
9. Set $j\coloneqq j+1$ and go back to step 4

Note that step 7 lets the data determine the number of hypotheses in $\mathcal{H}_{j+1} \setminus \mathcal{H}_j$: this number grows if going from $\mathcal{H}_{j-1}$ to $\mathcal{H}_j$ we concluded we do not need to draw any additional conjectured false nulls. This scheme is well adapted to handle a fairly common scenario where most of the highest scoring hypotheses are false null, making sure they will be labeled as such in our resamples.

The second phase of the algorithm simply resamples the indices in the usual bootstrap manner and then randomly permutes the conjectured true null scores:

1. independently sample m indices $j_1, \ldots, j_m \in \{1, 2, \ldots, m\}$
2. for $i = 1, \ldots, m$:
   (a) if $\boldsymbol{f}_{j_i} = 0$ draw a permutation $\boldsymbol{\pi}_i \in \Pi_{d_1}$, else, $\boldsymbol{f}_{j_i} = 1$ so define $\pi_i \coloneqq Id \in \Pi_{d_1}$ (the identity permutation)
   (b) apply the permutation π_i to $\boldsymbol{V}_i \coloneqq \left(\tilde{Z}^0_{j_i} \coloneqq Z_{j_i}, \tilde{Z}^1_{j_i}, \ldots, \tilde{Z}^d_{j_i}\right)$: $\boldsymbol{V}_i \circ \pi_i \coloneqq \left(\tilde{Z}^{\pi_i(1)-1}_{j_i}, \ldots, \tilde{Z}^{\pi_i(d_1)-1}_{j_i}\right)$
3. return the resampled labeled data $\left\{\left(\boldsymbol{V}_i \circ \pi_i, \boldsymbol{f}_{j_i}\right) : i = 1, \ldots, m\right\}$

6.9 Labeled Bootstrap monitored Maximization (LBM)

Given the list of original and decoy scores, an ordered list of candidate methods $\mathcal{M}$, a fall-back method M_f, a set of considered FDR thresholds Φ, and the number of bootstrap samples n_b, LBM executes the following steps:

1. For each bootstrap/resample run $i = 1, \ldots n_b$:
   (a) Generate a labeled resample as describe in Supplementary Section 6.8 above.
   (b) Apply each method $M \in \mathcal{M}$ to the resample noting the number of discoveries $D^i_M(\alpha)$ for each $\alpha \in \Phi$, as well as the corresponding FDP, $F^i_M(\alpha)$ (computed based on the conjectured labels of the resample).
   (c) For each $\alpha \in \Phi$ sort the methods according to $D^i_M(\alpha)$ with ties broken according to the rank of the methods in the list $\mathcal{M}$, and
      i. record the rank $r^i_M(\alpha)$ of each method,
      ii. record $F^i_*(\alpha) \coloneqq F^i_M(\alpha)$ where M is the highest rank method (with the largest number of discoveries).
2. For each $\alpha \in \Phi$:
   (a) Estimate the FDR of the direct maximization approach as the simple average $\widehat{FDR}_*(\alpha) \coloneqq \frac{1}{n_b} \sum_{i=1}^{n_b} F^i_*(\alpha)$.
   (b) If $\widehat{FDR}_*(\alpha) > \alpha$ (see the comment below) then
      – (the FDR of direct maximization seems too high so) set the selected method for this α to the fall-back method: $S(\alpha) \coloneqq M_f$.

   Otherwise,
      – (direct maximization seems to work fine so) set the selected method to the one with the highest average rank: $S(\alpha) \coloneqq \operatorname{argmax}_M \sum_{i=1}^{n_b} r^i_M(\alpha)$ (ties are broken according to the rank of the methods in the list $\mathcal{M}$).

We added to LBM two options that in practice were used throughout our reported simulations. The first is that we allowed some slack when comparing $\widehat{FDR}_*(\alpha)$ with α to check whether the FDR of direct maximization seems too high (step 2b above). Specifically, particularly because the empirical mean is taken over a relatively small number of resamples (we used $n_b = 50$ in our applications), we instead checked whether

$$\widehat{FDR}_*(\alpha) > \alpha + 4\sigma(\alpha) \cdot (1 - \hat{\pi}_0^*(\lambda)) \,,$$

where $\hat{\pi}_0^*(\lambda)$ is the π_0 estimate used by FDS$_1$ described in Supplementary Sections 6.5 and 6.6, and $\sigma(\alpha)$ is the estimated standard error of $\widehat{FDR}_*(\alpha)$. In practice, this relaxation lead to some increase in power with no visible impact on the FDR control.

The second option is a post-processing step that aims to produce a monotone list of discoveries as a function of the FDR threshold α. Specifically, we check if the number of discoveries at α_{j+1} is smaller than the number we have when using $\alpha_j < \alpha_{j+1}$, and if that is the case, then we override our resampling-based selection of the optimal method for α_{j+1} and instead we use the same method that was previously selected for α_j, i.e., $S(\alpha_{j+1}) \coloneqq S(\alpha_j)$.

In terms of the list of candidate methods we consider, $\mathcal{M}$, we need to strike a balance between considering more methods, equivalently more choices of (c, λ), and the increasing likelihood that the fall-back would be triggered. In practice, we found that considering the methods of FDS$_1$, mirror, and FDS (and in that the tie-breaking order, so FDS has the highest priority) works well so the reported version of LBM uses this particular list of methods.

6.10 Revisiting the failings of the canonical procedures

Going back to the two examples of Supplementary Section 6.2 we note that all our methods essentially control the FDR with the empirical FDR (FDP averaged over the 1K sets samples sets) below the selected FDR threshold for all $\alpha \in \Phi$ with a single exception in Example 1 at $\alpha = 0.2$, where FDS, FDS$_1$, and LBM have an empirical FDR in the range of 0.203-0.206 or up to 3% over the threshold (for reference, the empirical FDR of the guaranteed max method here is even higher: 20.9%), compared with the $> 250\%$ violation of Storey with pooled p-values.

Interestingly, when comparing the power of our methods in Example 2, where BH applied to the pooled p-values made no true discoveries even at $\alpha = 0.65$, we find that both the mirror and FDS$_1$ are significantly weaker than FDS, LF and LBM, again demonstrating the utility of LBM. Specifically, at $\alpha = 0.15$ both FDS's and LBM's power stand at 79.6% and 79.5% respectively, and LF's at 62.8% compared with 0% power for both the mirror and FDS$_1$. At $\alpha = 0.2$ FDS, LBM, and LF boast 100% power while the mirror power stands at 0.1% and FDS$_1$'s power is 0.6%.

6.11 Analysis of real data

We applied our analysis to three datasets.

The human data set consists of a single control run (`CTL_R1_1` from the data set with MassIVE identifier `MSV000079437` [17]. The data was generated on an LTQ-Orbitrap

Velos Pro on proteins extracted from human SH-SY5Y cells treated with 200 μM H_2O_2. The human reference proteome was downloaded from Uniprot on 28 Nov 2016.

The yeast data set consists of a single run (`Yeast_In-gel_digest_2`) selected at random from the data set with PRIDE identifier `PXD002726` [12]. The data was generated on an LTQ Orbitrap Velos on proteins extracted from an in-gel digest of *S. cerevisiae* lysate. The yeast reference proteome was downloaded from Uniprot on 28 Nov 2016.

The ISB18 data set is derived from a series of experiments using an 18-protein standard protein mixture (`https://regis-web.systemsbiology.net/PublicDatasets` https://regis-web.systemsbiology.net/PublicDatasets, [8]). We use 10 runs carried out on an Orbitrap (`Mix_7`). The database consists of the 18 proteins from the standard mixture, augmented with the full *H. influenzae* proteome, as provided by Klimek et al.

Searches were carried out using the Tide search engine [5] as implemented in Crux [11]. The peptide database included fully tryptic peptides, with a static modification for cysteine carbamidomethylation (`C+57.0214`) and a variable modification allowing up to six oxidized methionines (`6M+15.9949`). Precursor window size was selected automatically with Param-Medic [9]. The XCorr score function was employed for uncalibrated searches, using a fragment bin size selected by Param-Medic.

Clearly, the competition-based control of the FDR is subject to the variability of the drawn competing scores. To ameliorate this variability here, we initially searched the spectra against 100 randomly shuffled decoy databases, and then for each $d \in \{1, 3, 5, 7, 9\}$ we repeated our analysis drawing 100 sets, each with d of those decoy databases, while making sure that the 100 drawn sets are distinct. We can then compare the number of discoveries reported by each considered method at the selected FDR threshold α (here $\alpha \in \{0.01, 0.05, 0.1\}$). More precisely, for each number of decoys d we average the number of discoveries over the 100 randomly drawn sets of d decoys.

The ISB18 is a fairly unusual dataset in that it was generated using a controlled experiment, so the peptides that generated the spectra could have essentially only come from the 18 purified proteins used in the experiment. We used this dataset to get some feedback on how well our methods control the FDR, as explained next.

The spectra set was scanned against a target database that included, in addition to the 463 peptides of the 18 purified proteins, 29,379 peptides of 1,709 *H. influenzae* proteins (with ID's beginning with `gi|`). The latter foreign peptides were added in order to help us identify false positives: any foreign peptide reported is clearly a false discovery. Moreover, because the foreign peptides represent the overwhelming majority of the peptides in the target database (a ratio of 63.5 : 1), a native ISB18 peptide reported is most likely a true discovery (a randomly discovered peptide is much more likely to belong to the foreign majority). Taken together, this allows us to gauge the actual FDP for each set of d drawn decoys, FDR threshold α, and the FDR controlling procedure that generated the discovery list. More precisely, we average the FDP over the 100 drawn sets of d decoys.

The 87,549 spectra of the ISB18 dataset were assembled from 10 different aliquots, so in practice we essentially have 10 independent replicates of the experiment. However, the last aliquot had only 325 spectra that registered any match against the combined target database, compared with an average of over 3,800 spectra for the other 9 aliquots,

so we left it out when we independently applied our analysis to each of the replicates. By averaging the above decoy-drawn averaged FDP over the 9 aliquots we obtain an estimate of the FDR that we can compare to the selected FDR threshold.

Similarly, when gauging the power of a method on the ISB18 dataset our analysis was separately applied to each aliquot and then averaged over the aliquots.

Finally note that for purely technical reasons the peptide detection analysis was run using a slightly different version of LBM than the one described in Section 6.9 above. Specifically, instead of randomly breaking the ties in the ranks in step 1(c), the tied ranks were averaged at that point.

6.12 Gene Ontology enrichment Analysis

Gene Ontology term	TDC q-value	LBM q-value
cellular response to unfolded protein (GO:0034620)	9.35E-05	1.41E-04
response to unfolded protein (GO:0006986)	7.79E-05	1.18E-04
chaperone-mediated protein folding (GO:0061077)	1.37E-04	2.07E-04
cellular response to heat (GO:0034605)	1.83E-04	2.77E-04
response to heat (GO:0009408)	1.99E-04	3.00E-04
response to topologically incorrect protein (GO:0035966)	3.11E-04	4.68E-04
tRNA metabolic process (GO:0006399)	2.42E-02	3.31E-02
response to stress (GO:0006950)	7.73E-03	1.15E-02
cellular amino acid metabolic process (GO:0006520)	1.14E-02	1.68E-02
protein folding (GO:0006457)	1.31E-02	1.93E-02
translation (GO:0006412)	9.74E-07	2.94E-06
formation of translation initiation ternary complex (GO:0001677)	1.37E-05	3.38E-05
translational termination (GO:0006415)	9.10E-06	2.26E-05
translational elongation (GO:0006414)	6.83E-06	1.69E-05
cellular protein localization (GO:0034613)	—	2.93E-02
cellular macromolecule localization (GO:0070727)	—	3.13E-02
gene expression (GO:0010467)	2.47E-02	4.77E-02

Table 1: **Statistical overrepresentation of Gene Ontology terms in the yeast data set.** Each row is a Gene Ontology biological process term that is deemed significant at FDR < 0.05. Enrichments were tested twice, with respect to peptides identified using TDC and LBM.

The goal of many proteomics experiments is to understand what biological pathways are active in a given sample. We note that, as reported above, when using a single run of the yeast data for 33 of the 100 decoy databases we drew, TDC found 0 discoveries at this 1% FDR threshold meaning that in 1/3 of similarly conducted experiments we would not be able to draw any conclusion using this threshold. As noted above, LBM always reports some discoveries when analyzing the same spectra set at 1% FDR.

We next added two more spectra runs to the yeast dataset (Supplementary Section 6.11) representing a higher budget experiment. In this case at 1% FDR the average number of TDC discoveries was 275.9 and for LBM using $d = 5$ decoys it was 294. Accordingly, we focused on two sets of reported peptides, one of 294 peptides detected

by LBM (with $d = 5$) and another of 276 peptides detected via TDC. We eliminated from each group peptides that occur in more than one protein, and then subjected the remaining 54 and 50 proteins, respectively, to analysis via the PANTHER Classification System (`http://pantherdb.org`) [10]. Specifically, we performed an overrepresentation test for Gene Ontology biological process terms relative to the whole genome background. We used the "slim" term set and Fisher's exact test, controlling FDR via BH at 5%. This process yields 15 significantly overrepresented biological process terms from the TDC list and 17 from the LBM list (Table 1). The two missing terms—"cellular protein localization" and "cellular macromolecule localization"—are closely related and imply that the sample under investigation is enriched for proteins responsible in shuttling or maintaining other proteins in their proper cellular compartments. Critically, an analysis based solely on the traditional TDC approach would entirely miss this property of the sample being analyzed.

6.13 Conditional null exchangeability

The following condition which is a generalization of the conditional exchangeability property of [1] is weaker than the iid decoys condition. Nevertheless, it can be shown that it is sufficient for our statements to hold.

Definition 3. *Let* $\boldsymbol{V}_i \coloneqq \left(\tilde{Z}_i^0, \tilde{Z}_i^1, \ldots, \tilde{Z}_i^d\right)$*, where* $\tilde{Z}_i^0 \coloneqq Z_i$ *is the* i*th original score and* $\tilde{Z}_i^1, \ldots, \tilde{Z}_i^d$ *are the corresponding* d *decoy scores, and let* Π_{d_1} *denote the set of all permutations on* $\{1, \ldots, d, d+1 =: d_1\}$*. With* $\pi \in \Pi_{d_1}$ *we define* $\boldsymbol{V}_i \circ \pi \coloneqq \left(\tilde{Z}_i^{\pi(1)-1}, \ldots, \tilde{Z}_i^{\pi(d_1)-1}\right)$*, i.e., the permutation* π *is applied to the indices of the vector* $\boldsymbol{V}_i$ *rearranging the order of its entries. Let* $N \subset \{1, 2, \ldots, m\}$ *be the indices of the true null hypotheses and call a sequence of permutations* $\pi_1, \ldots, \pi_m$ *with* $\pi_i \in \Pi_{d_1}$ *a* null-only sequence *if* $\pi_i = Id$ *(the identity permutation) for all* $i \notin N$*. We say the data satisfies the* conditional null exchangeability property *if for any null-only sequence of permutations* $\pi_1, \ldots, \pi_m$*, the joint distribution of* $\boldsymbol{V}_1 \circ \pi_1, \ldots, \boldsymbol{V}_m \circ \pi_m$ *is invariant of* $\pi_1, \ldots, \pi_m$*.*

6.14 Figures

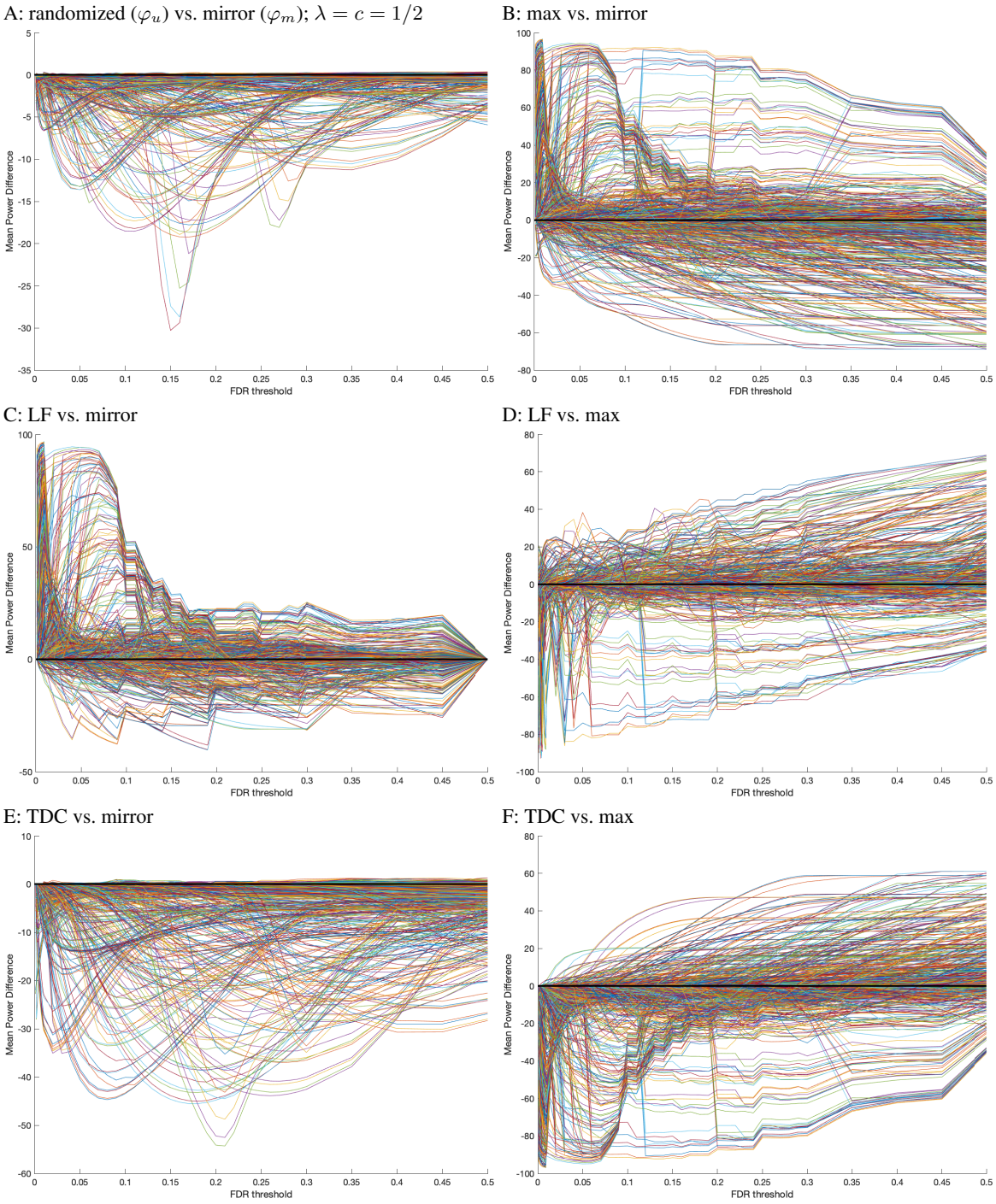

Fig. 1: **Power difference.** Each of the panels show the difference in the average power of the two compared methods (positive values indicate the first method is more powerful). Each panel is made of 1200 curves, each of which shows the difference in power averaged over the 10K (100K for panel A) sets. The sets were drawn simulating both calibrated and non-calibrated scores using the experiment-specific parameter combination as described in Supplementary Section 6.3. The power of each method is the 10K-average (100K for panel A) percentage of false nulls that are discovered at the given FDR threshold. Note that figures' y-axes are on different scales. **(A:)** with $c = \lambda = 1/2$ the mirror map (φ_m) is consistently better than the randomized uniform map (φ_u); 100K draws for each of the 1200 parameter combinations. **(B-D:)** for each of the mirror, max, and LF procedures there are numerous cases where its power is significantly below one of the other methods. **(F:)** the mirror is consistently better than TDC. **(E:)** the max is not consistently better than TDC.

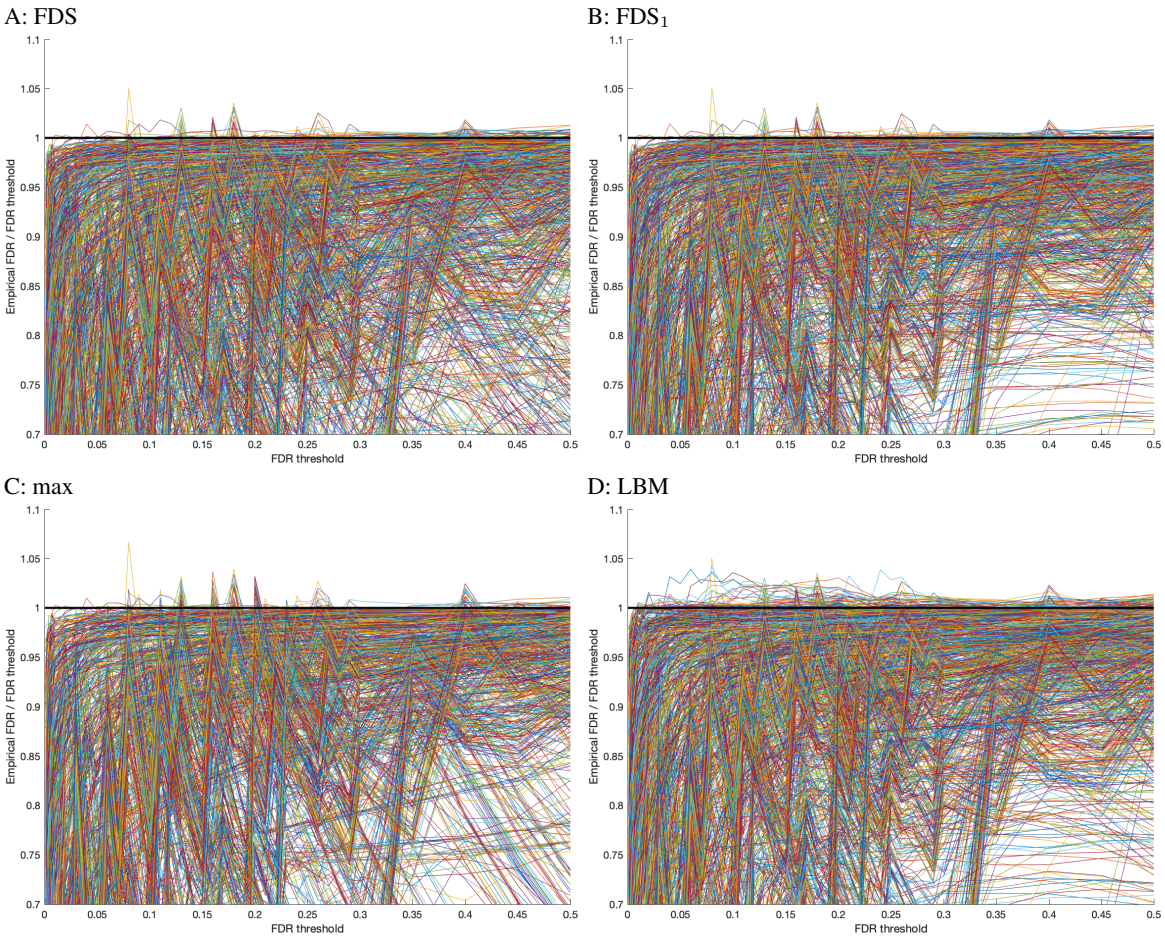

Fig. 2: **FDR control.** The panels show the ratio of the empirical FDR to the selected FDR threshold, and each is made of 1200 curves, each of which corresponds to one experiment involving 10K randomly drawn sets. The empirical FDR is the 10K-sample average of the FDP of each method's discovery list at the selected FDR threshold. The 10K sets were drawn simulating both calibrated and non-calibrated scores using the experiment-specific parameter combination as described in Supplementary Section 6.3. Notably there are relatively few cases where the empirical FDR is above the threshold (ratio > 1), and it is instructive to compare the ones observed in FDS, FDS$_1$ and LBM with those we note in the max method. Specifically, the overall maximal observed violation is 5.0% for FDS, FDS$_1$ and LBM while it is 6.7% for max. Similarly, the number of curves (out of 1200) in which the maximal violation exceeds 2% is 7 for FDS and FDS$_1$, 21 for LBM, and 24 for the max. Given that the max provably controls the FDR these simulations suggest that FDS, FDS$_1$ and LBM essentially control the FDR as well.

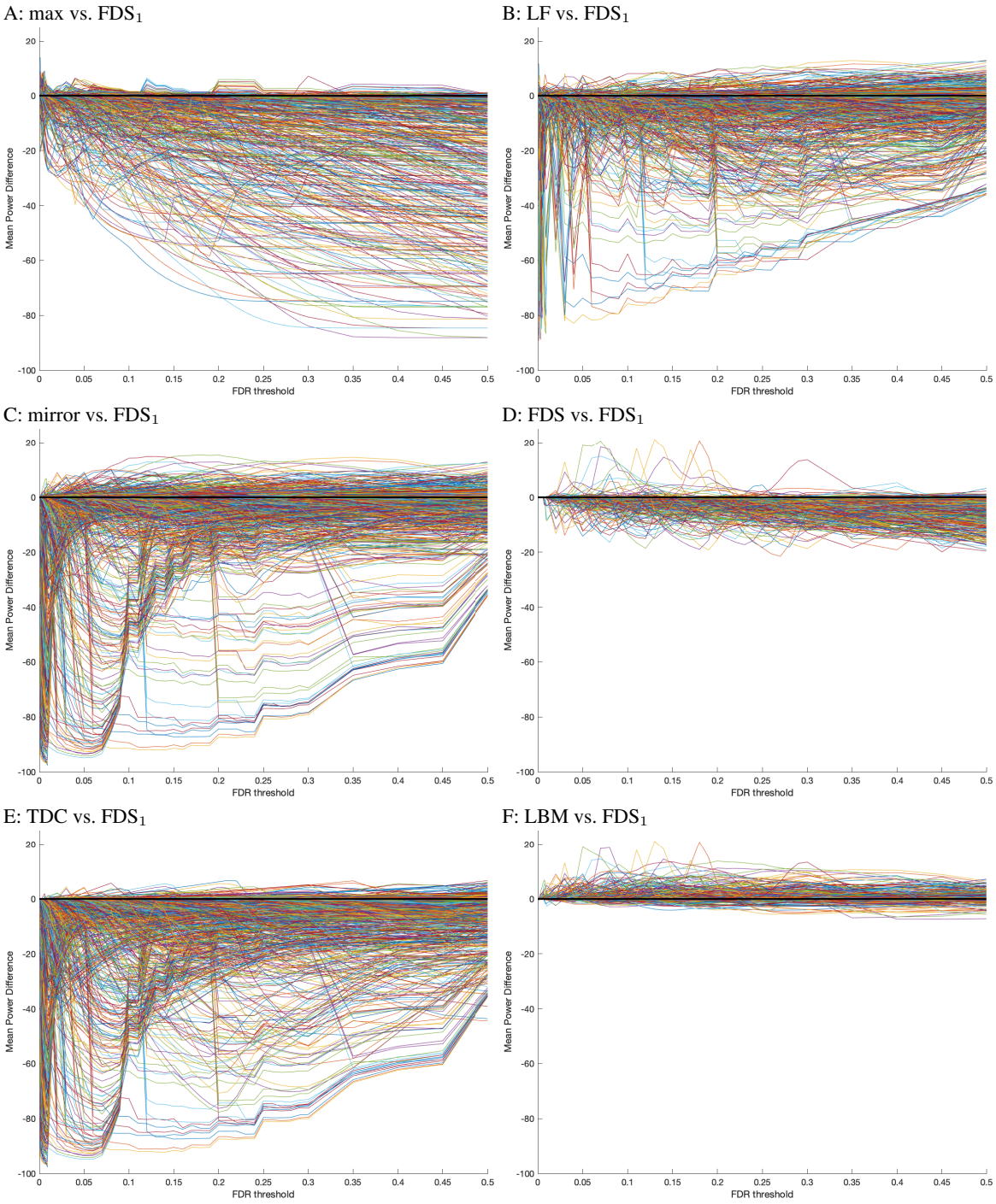

Fig. 3: **Power relative to FDS$_1$.** Each of the panels show the difference between the average power of the noted method and FDS$_1$ (negative values indicate FDS$_1$ is more powerful). Each panel is made of 1200 curves, each of which shows the difference in power averaged over the 10K sets. The sets were drawn simulating both calibrated and non-calibrated scores using the experiment-specific parameter combination as described in Supplementary Section 6.3. The power of each method is the 10K-average percentage of false nulls that are discovered at the given FDR threshold. **(A-E:)** FDS$_1$ arguably offers the best compromise among the multi-decoy procedures of mirror, max, LF, and FDS, as well as the single decoy TDC. **(F:)** LBM seems to offer an overall more powerful procedure.

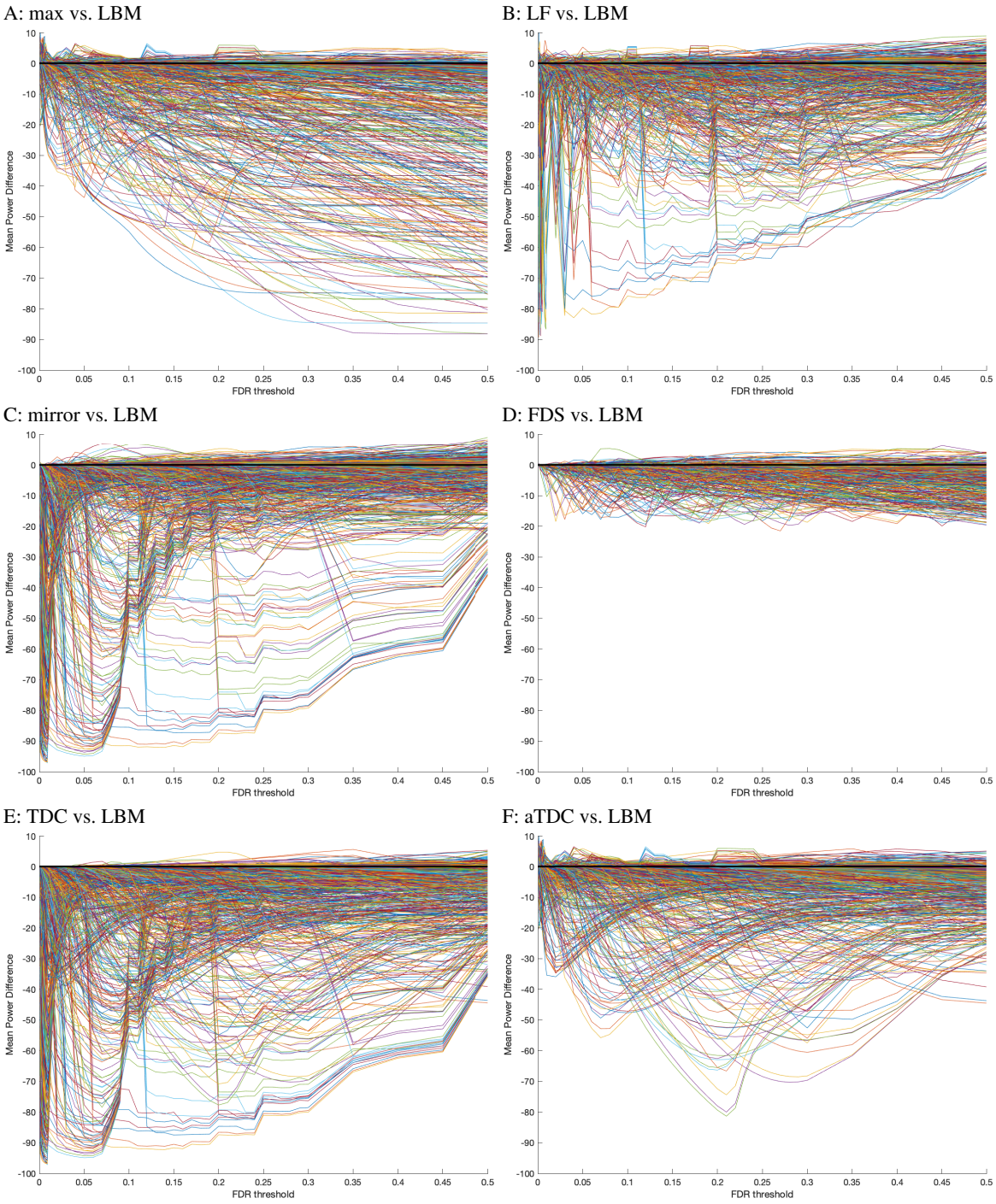

Fig. 4: **Power relative to LBM.** Each of the panels shows the difference between the average power of the noted method and LBM (negative values indicate LBM is more powerful). For LBM vs. FDS_1 see panel F of Supplementary Figure 3. Each panel is made of 1200 curves, each of which shows the difference in power averaged over the 10K sets. The sets were drawn simulating both calibrated and non-calibrated scores using the experiment-specific parameter combination as described in Supplementary Section 6.3. The power of each method is the 10K-average percentage of false nulls that are discovered at the given FDR threshold. LBM seems to offer an overall optimal procedure among the competition-based procedures considered here.

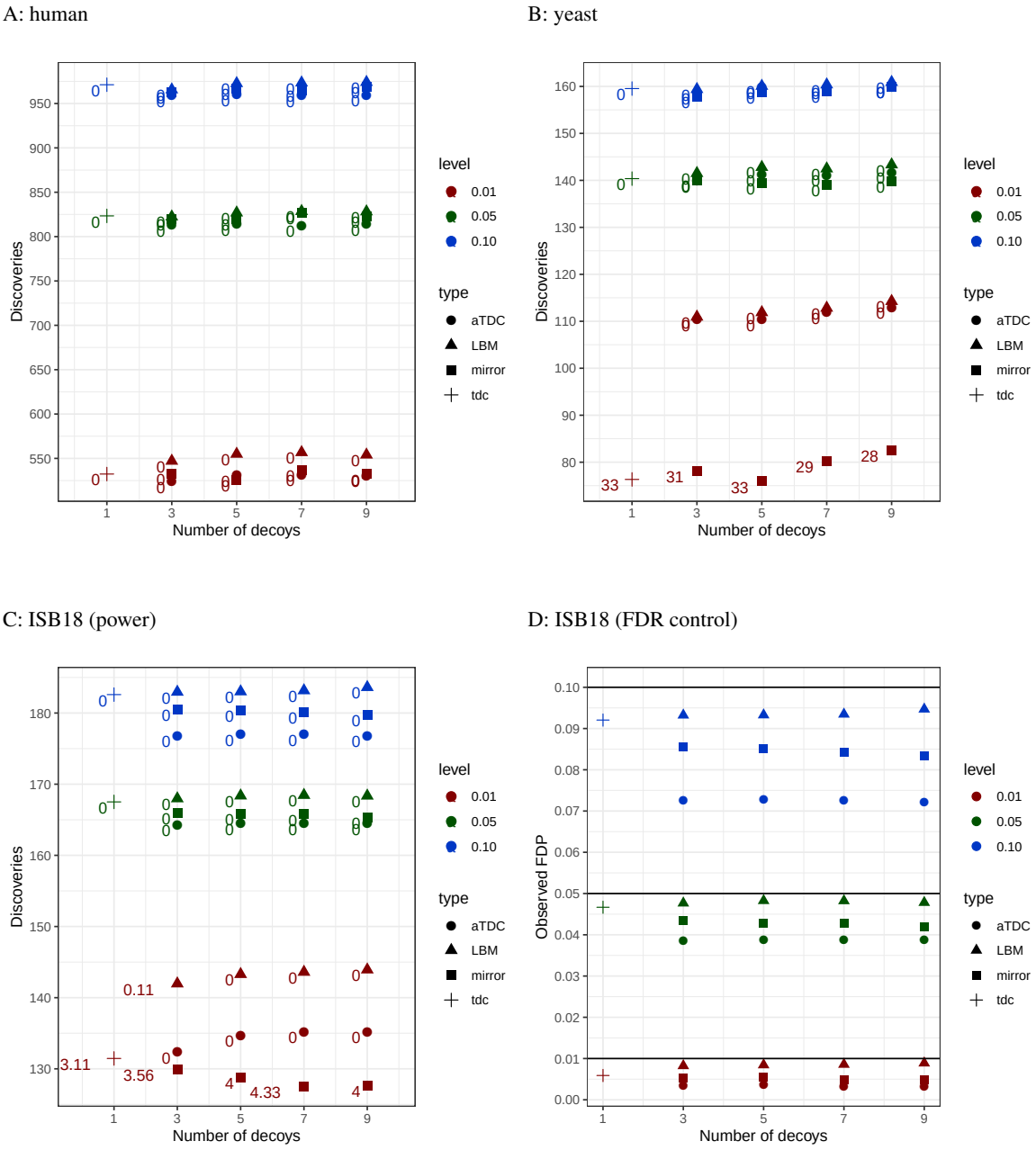

Fig. 5: **Peptide detection. (A: human)** The number of discoveries at the given FDR threshold is the average over the 100 randomly drawn decoys sets. Specifically, for each $d \in \{1, 3, 5, 7, 9\}$ we randomly drew 100 decoys sets, each with exactly d decoys, and applied TDC ($d = 1$), the mirror, LBM and aTDC ($d \in \{3, 5, 7, 9\}$) to the target and the drawn set of d decoy scores. We then noted the number of target discoveries at FDR thresholds of 1%, 5% and 10%, and finally we averaged those numbers over the 100 drawn decoy sets. The numbers to the left of the markers indicate the number of runs (out of 100) in which no discovery was reported. **(B: yeast)** Same as (A) but for the yeast dataset. **(C: ISB18, power)** Similar to (A) and (B) except for the ISB18 where the data consists of 9 aliquots or replicates. Therefore, the number of discoveries is averaged over the 9 aliquots, where for each aliquot we averaged the number of discoveries over 100 randomly drawn sets of d decoys as explained above. The numbers to the left of the marker indicate the aliquots-average number of runs (out of 100) in which no discovery was reported. **(D: ISB18, FDR)** Similar to (C) only here we noted the putative FDP of each run (as explained in Supplementary Section 6.11), then we averaged the FDP across the 100 runs to get an empirical FDR for each aliquot that we then averaged over the aliquots. Notably, in all cases the empirical FDR was lower than the selected threshold.

References

1. Barber, R.F., Candès, E.J.: Controlling the false discovery rate via knockoffs. The Annals of Statistics **43**(5), 2055–2085 (2015)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B **57**, 289–300 (1995)
3. Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational Behavioral Statististics **25**(60–83) (2000)
4. Benjamini, Y., Krieger, A.M., Yekutieli, D.: Adaptive linear step-up procedures that control the false discovery rate. Biometrika **93**(3), 491–507 (2006)
5. Diament, B., Noble, W.S.: Faster SEQUEST searching for peptide identification from tandem mass spectra. Journal of Proteome Research **10**(9), 3871–3879 (2011)
6. Keich, U., Kertesz-Farkas, A., Noble, W.S.: Improved false discovery rate estimation procedure for shotgun proteomics. Journal of Proteome Research **14**(8), 3148–3161 (2015)
7. Keich, U., Noble, W.S.: On the importance of well calibrated scores for identifying shotgun proteomics spectra. Journal of Proteome Research **14**(2), 1147–1160 (2015)
8. Klimek, J., Eddes, J.S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P.R., Katz, J.E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J.K., Aebersold, R., Martin, D.B.: The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. Journal of Proteome Research **7**(1), 96–1003 (2008)
9. May, D.H., Tamura, K., Noble, W.S.: Param-Medic: A tool for improving MS/MS database search yield by optimizing parameter settings. Journal of Proteome Research **16**(4), 1817–1824 (2017)
10. Mi, H., Muruganujan, A., Thomas, P.T.: PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Research **41**(Database issue), D377–D386 (2013)
11. Park, C.Y., Klammer, A.A., Käll, L., MacCoss, M.P., Noble, W.S.: Rapid and accurate peptide identification from tandem mass spectra. Journal of Proteome Research **7**(7), 3022–3027 (2008)
12. Schittmayer, M., Fritz, K., Liesinger, L., Griss, J., Birner-Gruenberger, R.: Cleaning out the litterbox of proteomic scientists' favorite pet: Optimized data analysis avoiding trypsin artifacts. Journal of Proteome Research **15**(4), 1222–1229 (2016)
13. Storey, J.D.: A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B **64**, 479–498 (2002)
14. Storey, J.D., Taylor, J.E., Siegmund, D.: Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society, Series B **66**, 187–205 (2004)
15. Storey, J.D., Bass, A.J., Dabney, A., Robinson, D.: qvalue: Q-value estimation for false discovery rate control (2019), `http://github.com/jdstorey/qvalue`, r package version 2.14.1
16. Xia, F., Zhang, M.J., Zou, J.Y., Tse, D.: NeuralFDR: Learning discovery thresholds from hypothesis features. In: Advances in Neural Information Processing Systems. pp. 1541–1550 (2017)
17. Zhong, L., Zhou, J., Chen, X., Lou, Y., Liu, D., Zou, X., Yang, B., Yin, Y., Pan, Y.: Quantitative proteomics study of the neuroprotective effects of B12 on hydrogen peroxide-induced apoptosis in SH-SY5Y cells. Scientific Reports **6**, 22635 (2016)