Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(523): 973–982. doi:10.1080/01621459.2017.1375931.

Controlling the FDR in imperfect matches to an incomplete database

Uri Keich and

School of Mathematics and Statistics F07, University of Sydney

William Stafford Noble

Departments of Genome Sciences and of Computer Science and Engineering, University of Washington

Abstract

We consider the problem of controlling the FDR among discoveries from searching an incomplete database. This problem differs from the classical multiple testing setting because there are two different types of false discoveries: those arising from objects that have no match in the database and those that are incorrectly matched. We show that commonly used FDR controlling procedures are inadequate for this setup, a special case of which is tandem mass spectrum identification. We then derive a novel FDR controlling approach which extensive simulations suggest is unbiased. We also compare its performance with problem-specific as well as general FDR controlling procedures using both simulated and real mass spectrometry data.

Keywords

multiple hypothesis testing; tandem mass spectrometry; false discovery rate

1 Introduction

In the problem of multiple testing we simultaneously test m (null) hypotheses $H_1, \dots, H_m$, 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 correct or a true null hypothesis.

When analyzing a large number of hypotheses one typically ends up controlling the false discovery rate (FDR), which is defined as the expected value of the false discovery proportion (FDP). Benjamini and Hochberg were the first to introduce a simple procedure (BH) that allows us to control the FDR at any desired level $\alpha \in (0, 1)$ [2].

Follow-up analysis by Benjamini and co-authors [3, 5] and Storey and co-authors [18, 21, 20], concentrated on further improving the discovery power of BH by incorporating into the procedure an estimate of $\pi_0 = m_0/m$, the fraction of true null hypotheses.¹ These methods,

¹Appendix Table 1 summarizes our notations for this multiple testing context.

which are generally referred to as “adaptive BH,” became very popular even though some of them only control the FDR in some asymptotic setting [21, 20].

Consider next a more specific setup where through some process a set of unobserved “sources” generate observable “products”. Given the set of products, our goal is to try to identify as many of the unobserved sources as possible using a database matching procedure. Specifically, we have a database of potential sources, and each of the observed products is scanned against it with the goal of finding a source that maximizes our match scoring function. The hope is that this scan will indeed correctly match the product with its source but the matches are imperfect and at the same time unverifiable: we have no way to confirm that we have the correct match.

Borrowing the multiple hypothesis testing terminology and accepting all matches whose score exceeds a threshold ρ we declare those as discoveries. The FDP in this case is the proportion of incorrect matches among our discoveries, and our goal is to address the problem of controlling the FDR in this context. This goal is further complicated by our assumption that our database of sources is incomplete so our products fall into two categories: “native” products whose matching source is in the database and “foreign” products whose source is missing. Thus, any pairing of a foreign product is of course incorrect, but due to the imperfect nature of the match scoring function some of the pairings involving native products are incorrect as well. While from a practical perspective whether a false discovery is due to a native or foreign product is irrelevant, as we will see, this information is pertinent for correctly controlling the FDR.

An important special case of the general problem described above is identification of tandem mass spectra. In tandem mass spectrometry an input stream of peptides (sources) generates a very large number of spectra (products), each of which consists of a list of observed “peaks”. The location of each peak corresponds to the mass-to-charge ratio of an isolated fragment of the peptide and its height is related to the fragment abundance. The primary downstream analysis task is to try and identify, based on this pattern of peaks, the amino acid sequence of the peptide that was responsible for generating each observed spectrum. This is typically done using a peptide database search procedure that scans each spectrum against the database to find the peptide that optimizes the score assigned to each considered peptide-spectrum match (PSM) ([17] reviews these search tools).

As in the general problem, the spectra are divided into the set of native spectra, which were generated by peptides in the database, and the complementary set of foreign spectra, which were generated by molecules that are not in the database. In practice, foreign spectra are generated by contaminant peptides, peptide variants that are not in the given database, peptides with unexpected post-translational modifications and non-peptide species, as well as spectra for which the charge state was not correctly identified.

Practical solutions to the problem of estimating the FDR in spectra identification commonly rely on comparing the spectrum identifications obtained from searching a real (*target*) peptide database with those obtained from searching against a decoy database that is drawn

according to some null model, typically of shuffled peptides.² This is most commonly done using *target-decoy competition* (TDC, Section 3.3.1) [7].

One reason for the popularity of TDC is its wide applicability. Specifically, the procedure does not require computing any p-values because it only uses the PSM score to rank the matches. Moreover, the score function need not be well calibrated with respect to the spectra.³ Still, we recently showed that if the score is calibrated then it is preferable to control the rate of false discoveries using our newly introduced decoy-dependent estimator of the FDR, mix-max [13]. Specifically, mix-max typically yields slightly more discoveries and exhibits less variability compared with TDC.

The general problem we study here was formulated from the observation that, in practice, mix-max calibrates the original PSM score by converting it to a p-value using an empirical cumulative distribution function (CDF) (Appendix Section B.1). Specifically, having carried out this conversion, we can use the knowledge that we have p-values, and not just calibrated scores, to cast the problem in the familiar language of hypothesis testing, albeit with an extra twist: the p-values are computed relative to the null hypothesis that the database is randomly drawn. This has the unfortunate effect of implicitly modifying the null hypothesis to be “the product/spectrum is foreign” rather than the more desirable “the match is incorrect”.⁴ Hence, if we are not careful what we end up controlling is only the proportion of foreign products rather than of incorrect matches.

We propose here a novel procedure, “expected mix-max” (e-mix-max), specifically designed to control the FDR in this context of searching an incomplete database with an imperfect matching procedure. We then provide substantial empirical evidence that e-mix-max is currently the optimal method for controlling the FDR in our context: in contrast with the conservative bias of BH and the liberal bias of a widely used variant of adaptive BH due to Storey and colleagues, e-mix-max’s estimation of the FDR is apparently unbiased. Moreover, our simulations show that e-mix-max exhibits less variation than the decoy-dependent mix-max, which we previously showed was better than TDC in this aspect.

For concreteness, in the remainder of the paper we primarily employ mass spectrometry terminology; however, our analysis applies to the more general problem, which arises in multiple domains. For example, the problem arises in metagenomics sequence homology search. Suppose that we have just sequenced DNA from an environmental sample (dirt, seawater, etc.). From the DNA sequences, we can infer amino acid sequences, and we would like to search these amino acid sequences against the non-redundant protein database (or some other appropriate database) to ascertain whether a given sequence has a protein homolog or not. The problem is that our inference of protein sequences from the metagenomic data is imperfect. In some cases, the inferred amino acid sequence actually

²Non-randomly generated decoys such as reversing the peptides are also commonly used. We consider such decoys as pseudo-random draws from the null database distribution. Note that, once shuffled, the peptide will generally no longer provide a good match to a spectrum it generated.

³A PSM score function is *calibrated* with respect to spectra (and a null peptide database model) if the distribution of the optimal PSM score when searching a null drawn peptide database is invariant of the spectrum.

⁴In the Discussion we elaborate on why one cannot simply compute the p-value relative to the intuitively preferable null that the match is incorrect.

corresponds to a real peptide sequence (native product in our setup); in other cases, it is simply a stretch of DNA that happens not to contain a stop codon but does not actually correspond to a portion of a translated protein (foreign product in our setup). Since we do not know which of our query peptides falls into which category this problem naturally fits our setup. Our theoretical framework also applies to forensic search, whether searching for a match to a crime scene fingerprint in a fingerprint database, or searching for a match to a pattern of DNA markers against a database such as the FBI's Combined DNA Index System database.

2 Our model

Our setup is essentially the same one we introduced in [13]. Briefly, we have a target database db of peptides and a set Σ of n spectra. The spectra are divided into the set Σ_1 of native spectra, which were generated by peptides in db , and the complementary set $\Sigma_0 = \Sigma \setminus \Sigma_1$ of foreign spectra, which were generated by molecules not in db . Let $n_i = |\Sigma_i|$ and let $\pi_i = n_i/n$ be the proportion of foreign ($i = 0$) and native ($i = 1$) spectra in our input data set Σ . For reference, our notation in this (peptide) database search context is summarized in Appendix Table 2.

The quality of the match between a spectrum σ and a peptide $\ell \in db$ is determined by a score function $S_{db}(\sigma, \ell)$. Some scores can be invariant with respect to the database db , e.g., Xcorr [9], whereas others can depend on db , for example, through the number of candidate peptides, e.g., E-value [15].

We pair every spectrum $\sigma_i \in \Sigma$ with $\ell(\sigma_i, db)$, its best candidate peptide in db ,⁵ and we refer to the pair as an optimal PSM and denote its score by $w_i := S(\sigma_i, \ell(\sigma_i, db))$. We loosely refer to w_i as the score of the best match to σ_i in db .

We assume that for each native spectrum $\sigma_i \in \Sigma_1$ there exists a unique peptide $G(\sigma_i) \in db$ that generated it, and we let $x_i = S_{db}(\sigma_i, G(\sigma_i))$, be the score of the match between that spectrum σ_i and its generating peptide. In our model each x_i is a particular realization of a random variable X_i , and we assume that the X_i 's are drawn independently from one another from an unknown but common distribution.

For each foreign spectrum $\sigma_i \in \Sigma_0$ we define $y_i = w_i$ the score of the best match to this foreign spectrum in db . Similarly, for $\sigma_i \in \Sigma_1$ we define $y_i = \max_{\ell \in db \setminus \{G(\sigma_i)\}} S_{db}(\sigma_i, \ell)$ as the score of the best match to σ_i in db minus the generating peptide, that is, the score of the best match against the irrelevant part of db . Note that defining $x_i := -\infty$ for $\sigma_i \in \Sigma_0$ we have $w_i = \max\{x_i, y_i\}$ for all i .

We assume that each value y_i is a realization of a random variable Y_i whose distribution is the same as the distribution of the score of the optimal match between σ_i and a peptide database DB that is randomly drawn from our null database model. In our PSM context DB is generated by essentially randomly shuffling the order of the amino acid sequence of each

⁵We assume there are no ties, or equivalently that ties are broken randomly.

of the peptides in the target database. Consistent with assuming the score is calibrated, we assume the Y_i 's are iid (independent and identically distributed).⁶

We further assume that the Y_i are also independent of the X_i and finally, that in the context of this theoretical model, no ties occur: $x_i \neq y_i$ for all i . This assumption can be justified either by assuming the scores are sampled from a jointly continuous distribution or by breaking ties using a (fair) coin flip.

To model searching a randomly drawn decoy database we denote by z_i the observed score of the optimal match to σ_i in this decoy, and we assume each z_i is a realization of a corresponding random variable Z_i . The Z_i 's, which are distributed the same as the Y_i 's, are independent from one another as well as from all Y_i 's and X_i 's. Again, we assume no ties: $w_i \neq z_i$ for all i .

We define the *p-value* of an optimal target PSM w_i as $P(Y_i \leq w_i)$, which is the same as the probability that we will observe a score at least as good as w_i assuming the spectrum is foreign. Note that the probability here is taken with respect to a null drawn database. This p-value can be computed using dynamic programming [16, 11] for some score functions but can also be estimated using Monte Carlo methods (Appendix Section B.1).

The pairing of σ_i with its optimal match, (σ_i, db) , is incorrect, or is a false discovery at score level ρ , if $y_i > \max(x_i, \rho)$, where by definition $x_i := -\infty$ for $\sigma_i \in \Sigma_0$.

Note that for any spectrum, w_i is observed but x_i and y_i are not, hence, the number of “discoveries” $D = D(\rho) := |\{i : w_i > \rho\}|$ is also observable, but the number $F = F(\rho)$ of false discoveries, is not, nor is the FDP $Q = F/\max\{D, 1\}$.

We are interested in a procedure that, given a threshold α , will allow us to determine $\rho = \rho(\alpha)$ so that, if we declare all PSMs scoring above ρ as discoveries, then our FDR, which is $E(Q)$, is $\leq \alpha$.

3 Controlling the false discovery rate

3.1 General protocol for FDR control

Storey, Taylor and Siegmund (STS) outlined the following general heuristic for FDR control [20] (similar ideas were explored in [10]).

Let $V(t)$ and $R(t)$, be the empirical processes denoting the numbers of false discoveries and discoveries, respectively, when rejecting all null hypotheses whose p-values are $< t$.

Denoting the FDP among these rejected nulls by $Q(t) = V(t)/(R(t) \vee 1)$, the corresponding FDR is $FDR(t) = E(Q)$, and recall the q-value is defined as [18, 21]

$$q\text{-value}(t) = \min_{p \geq t} FDR(p). \quad (1)$$

⁶We assume the database is large so that the slight difference in the database size due to removal of $G(\sigma_i)$ for native spectra can be glossed over.

The STS approach assumes we can readily compute estimators $\widehat{\text{FDR}}(t)$ of $\text{FDR}(t)$ for each $t \in (0, 1)$. A typical such estimator has a downward jump at every observed p-value p_i and increases in between adjacent order statistics $p_{(i)}$ and $p_{(i+1)}$. Hence for those estimators we can readily derive estimators of q-value(p_i) based on (1) above:

$$\widehat{\text{q-value}}(p_i) = \min_{p_j \geq p_i} \widehat{\text{FDR}}(p_j). \quad (2)$$

In a generalized form, the selection procedure of STS essentially amounts to computing $\widehat{\text{q-value}}(p_i)$ based on the estimates $\widehat{\text{FDR}}(p_j)$ and then rejecting all null hypotheses H_i for which $\widehat{\text{q-value}}(p_i) \leq \alpha$.

One advantage of tackling the FDR control problem through this combination of constructing point estimators $\widehat{\text{FDR}}(t)$ and thresholding the resulting $\widehat{\text{q-value}}(p_i)$ is that it is often intuitively clear how to construct $\widehat{\text{FDR}}(t)$. We stress that even if $\widehat{\text{FDR}}(t)$ is consistent for all t it does not automatically follow that rejecting all hypotheses for which $\widehat{\text{q-value}}(p_i) \leq \alpha$ results in a procedure that controls the FDR at level α . However, aside from demonstrating that this approach works for a couple of specific estimators, STS also argue that, if you can show that in some reasonable asymptotic setting the point estimators $\widehat{\text{FDR}}(t)$ are almost surely asymptotically simultaneously conservative (that is, asymptotically uniformly conservative over all rejection regions), then the above procedure does guarantee FDR control in your asymptotic setting [20].

In what follows we adopt this q-value based meta procedure for controlling the FDR and apply it to different estimators $\widehat{\text{FDR}}(t)$. We start with the estimators that are commonly used in the multiple testing context followed by estimators, including our newly proposed one, which are specialized to our context. We will comment on the validity of the procedure on a case by case basis.

3.2 Estimators of $\text{FDR}(t)$ in the multiple testing context

To simplify our notations below we assume the null hypotheses $H_1, \dots, H_m$ are sorted according to their p-values so that $p_i = p_{(i)}$.

3.2.1 Benjamini-Hochberg (BH)—The BH procedure is equivalent to using the above general protocol with the estimator

$$\widehat{\text{FDR}}_{\text{BH}}(t) = \frac{mt}{R(t) \vee 1}. \quad (3)$$

Since $R(p_i) = i$ this specializes to $\widehat{\text{FDR}}_{\text{BH}}(p_i) = mp_i/i$, which in turn implies that $\widehat{\text{q-value}}_{\text{BH}}(p_i) = \min_{j \geq i} mp_j/j$. The proof that it indeed controls the FDR at level α can be found in [2] (or, more in line with our presentation, in [20]).

The rationale for the estimator (3) is clear: the numerator is $E[V(t)]$ when considering a worst case scenario where all the hypotheses are true null, whereas the denominator essentially represents our best guess of $E[R(t)]$. The ratio of the two is therefore a natural estimator of the marginal FDR, $E[V(t)]/E[R(t)]$, which in turn is closely related to the FDR [19, 21].

The conservative nature of BH is suggested by our derivation of $\widehat{\text{FDR}}_{\text{BH}}(t)$: we estimate $E[V(t)] = mt$, which assumes all hypotheses are true null. Indeed, using this estimator to control the FDR among the list of rejected hypotheses in the classical multiple testing setup results in a conservative bias: the actual FDR will be lower than the nominal one.

By the same token, applying BH to our database search context implicitly assumes that all spectra are foreign and estimates the number of false discoveries at level p_i as $n \cdot p_i$. It is intuitively clear that this estimate is again conservative since this estimate coincides with the expected number of Y_j whose corresponding p-value is more significant than p_i . However, many of these Y_j , which correspond to native spectra $\sigma_j \in \Sigma_1$, will lose to the corresponding native match score ($X_j > Y_j$) and hence will not be false discoveries. We will later see that this is borne out in our simulations.

3.2.2 Adaptive BH (Storey)—The conservative nature of $\widehat{\text{FDR}}_{\text{BH}}(t)$ was addressed in several works, notably including [3, 5, 18, 21, 20], which proposed an adaptive BH procedure that essentially includes an estimate of $\pi_0 = m_0/m$, the overall fraction of true null hypotheses. Here we chose to concentrate on one highly popular such method, which relies on two estimators of $\text{FDR}(t)$ that are implemented in Storey’s `qvalue` R package.

The `qvalue` package provides two options for estimating $\text{FDR}(t)$ that we collectively refer to as $\widehat{\text{FDR}}_S(t)$. The estimators differ only in how they estimate π_0 (Appendix Section B.2): one is a bootstrap derived estimate [20] and the other is a spline-based estimate [21]. Given the selected estimator $\hat{\pi}_0$ the adaptive BH estimator that we refer to as “Storey” is defined as

$$\widehat{\text{FDR}}_S(t) = \frac{\hat{\pi}_0 \cdot m \cdot t}{R(t) \vee 1}. \quad (4)$$

Again, using $R(p_i) = i$ this specializes to $\widehat{\text{FDR}}_S(p_i) = \hat{\pi}_0 \cdot m \cdot p_i / i$, which in turn implies that $\widehat{\text{q-value}}_S(p_i) = \min_{j \geq i} \hat{\pi}_0 \cdot m \cdot p_j / j$.

Again, the rationale for this estimator is obvious once we realize that $\hat{\pi}_0 \cdot m$ is an estimator of m_0 . Storey, Taylor and Siegmund claim that in the asymptotic setting they consider, it can be shown that, when using the bootstrap estimator of π_0 , $\widehat{\text{FDR}}_S(t)$ and $\widehat{\text{q-value}}_S(t)$ are almost surely asymptotically uniformly conservative and hence the above procedure will control the FDR at least in that asymptotic sense [20]. A similar claim is made by Storey and Tibshirani regarding the spline-estimator of π_0 [21].

In our context, the Storey estimator (4) is expected to be liberally biased since $\hat{\pi}_0$ estimates the overall fraction of foreign spectra rather than that of false discoveries. As such, the estimator fails to account for the false discoveries due to native spectra. Again this is borne out in our experiments, and while we did not test other adaptive BH procedure we expect them to be liberally biased for the same reason.

3.3 Estimators of $\text{FDR}(\rho)$ specialized to tandem mass spectrum identification

Controlling the FDR in tandem mass spectrum identification has been practically done for some time. In this section we recast a couple of existing practices using the standard statistical terminology where possible.

Overwhelmingly, FDR control in tandem mass spectrum identification is accomplished using target-decoy competition (TDC, Section 3.3.1 below) using the PSM score function, $S_{DB}(\sigma, \ell)$. Indeed, methods for calculating p-values were only recently introduced [16, 11]. Correspondingly, we need to derive point estimators $\widehat{\text{FDR}}(\rho)$, and $\widehat{\text{q-value}}(\rho)$ where our parameter ρ represents a score threshold rather than a p-value threshold. Note that unlike for p-values, higher scores are better so in particular $\widehat{\text{q-value}}(\rho) = \min_{\tau \leq \rho} \widehat{\text{FDR}}(\tau)$.

To emphasize the distinct setup in this problem, we use slightly different notation here. As introduced in Section 2, let $D(\rho)$ denote the number of discoveries or PSMs scoring $> \rho$ and $F(\rho)$ denote the number of false discoveries, or incorrect PSMs scoring $> \rho$. We define $\text{FDR}(\rho) = E[F(\rho)/(D(\rho) \vee 1)]$.

3.3.1 Target-decoy competition (TDC)—TDC relies on a competition between target and decoy PSMs to determine the list of reported PSMs. In the variant we refer to as T-TDC each optimal target PSM score w_i is compared with its corresponding decoy PSM score z_i . At a given score threshold ρ , T-TDC reports as discoveries the $D_T(\rho)$ target PSMs for which $w_i > \max\{\rho, z_i\}$. The corresponding number of decoy discoveries, $F_D = |\{i: z_i > \max\{\rho, w_i\}\}|$, is then used to estimate the FDR among the $D_T(\rho)$ target discoveries by [6, 12, 8]:

$$\widehat{\text{FDR}}_{\text{TT}}(\rho) = \frac{F_D(\rho)}{D_T(\rho)}. \quad (5)$$

Note that any target PSM that loses its decoy competition ($z_i > w_i$) is lost, or filtered out, so in particular $D(\rho) = D_T(\rho)$.

Recently, Barber and Candés proved, in an altogether different context, that when replacing F_D in the numerator of (5) with $1 + F_D$ the general q-value procedure (Section 3.1) achieves rigorous FDR control for any pre-selected threshold [1]. In particular, it follows that using T-TDC, as it is, that is declaring the non-filtered target PSM i as a discovery if $\widehat{\text{q-value}}_{\text{TT}}(w_i) \leq \alpha$, where $\widehat{\text{q-value}}_{\text{TT}}(\rho) = \min_{\tau \leq \rho} \widehat{\text{FDR}}_{\text{TT}}(\tau)$, controls the FDR at level α for any reasonable asymptotic setting (see also [13]).

3.3.2 Mixture-maximum (mix-max)—As mentioned, the T-TDC procedure occasionally sacrifices a target discovery that loses to a decoy peptide offering a better match to the same spectrum. Ideally we would like to control the FDR among all sufficiently high scoring target PSMs rather than in some randomly filtered version of this list. We recently introduced the mixture-maximum (mix-max) procedure, which relies on a randomly drawn decoy database to accomplish this goal [13].

Designed in the vein of $\widehat{\text{FDR}}_S$, (4), mix-max estimates $\text{FDR}(\rho)$ among all $D(\rho)$ target PSMs with a score $w_i > \rho$. It separately estimates the expected number of false discoveries due to foreign spectra, $F_0(\rho) = |\{i: \sigma_i \in \Sigma_0, y_i > \rho\}|$, and the expected number of false discoveries due to native spectra, $F_1(\rho) = |\{i: \sigma_i \in \Sigma_1, y_i > \max(x_i, \rho)\}|$.

Let $\hat{\pi}_0$ be the estimated proportion of foreign spectra (mix-max uses Storey's estimators as described in Section 3.2.2 above), then mix-max estimates $E[F_0(\rho)]$ as

$$E[\widehat{F}_0(\rho)] = \hat{\pi}_0 \cdot n \cdot P(\widehat{Z}_i > \rho) = \hat{\pi}_0 \cdot \sum_{j=1}^n 1_{z_j > \rho}. \quad (6)$$

Recall that Y_i and Z_i are identically distributed and that z_i are the optimal PSM scores when searching the randomly drawn decoy database (Section 2).

The estimator of $E[F_1(\rho)]$ is more involved. First, we prove that for any native spectrum $\sigma_i \in \Sigma_1$

$$E[F_1(\rho)] = n_1 \cdot \int_{y > \rho} \frac{P(W_i < y)}{P(Y_i < y)} dG_*(y), \quad (7)$$

where $n_1 = |\Sigma_1|$ and G_* is the cumulative distribution function of the optimal null score (Y_i or Z_i). Then, relying on this identity, we estimate $E(F_1)$ by estimating the right-hand side of (7) to get

$$\begin{aligned} \widehat{\text{FDR}}_m(\rho) &:= \frac{E[\widehat{F}_0(\rho)] + E[\widehat{F}_1(\rho)]}{D(\rho)} \\ &= \frac{\hat{\pi}_0 \cdot \sum_{j=1}^n 1_{z_j > \rho} + (1 - \hat{\pi}_0) \cdot \sum_{z_j > \rho} \left[\frac{\sum_k 1_{w_k \leq z_j}}{(1 - \hat{\pi}_0) \cdot \sum_k 1_{z_k \leq z_j}} - \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right]_{[0, 1]}}{\sum_i 1_{w_i > \rho}}, \end{aligned}$$

where $1_{z_i > \rho}$, $1_{w_i > \rho}$, $1_{w_k \leq z_j}$, $1_{z_k \leq z_j}$ are 1 or 0, depending on whether the corresponding inequality holds, and $[x]_{[0, 1]} := \max \{0, \min \{1, x\}\}$ ensures that x remains an acceptable probability value [13].

Again, mix-max attempts to control the FDR by employing the estimator $\widehat{\text{FDR}}_m$ in combination with the general q-value based protocol for controlling the FDR at level α (Section 3.1). We used simulation studies to show the resulting procedure generally controls the FDR at the prescribed level with some liberal bias (comparable to that of T-TDC) for small spectra sets and very small FDR thresholds.

3.3.3 e-mix-max—Our new procedure is also designed to control the FDR among all $D(\rho)$ target PSMs with a score $> \rho$, and again it does so by devising a point estimator for $\text{FDR}(\rho)$ which is then utilized in the general q-value based procedure (Section 3.1).

Unlike, T-TDC and mix-max, e-mix-max works with the p-values $p_i = P(Y_i \leq w_i)$ rather than with the scores w_i (see Section 2 for methods for computing these p-values). Like, mix-max, e-mix-max estimates $\text{FDR}(t)$ by separately accounting for false discoveries due to native and foreign spectra.

First, we estimate π_0 exactly as in the Storey estimator above (also, Appendix Section B.2). Note that mix-max uses the same estimation technique except the p-values are computed from the empirical CDF of the decoy scores, whereas here we assume that we know the p-values. Then we replace the right-hand side of (6) with

$$\widehat{E[F_0(t)]} := \hat{\pi}_0 \cdot n \cdot t = \hat{n}_0 \cdot t,$$

where $\hat{n}_0 := \hat{\pi}_0 n$, is our estimate of the expected number of foreign spectra discoveries at p-value threshold of t .

To estimate $E[F_1(t)]$ for $\sigma_i \in \Sigma_1$ we first rewrite (7) in terms of p-values. Replacing W_i with $pW := \text{p-value}(W_i)$ and Y_i with $pY := \text{p-value}(Y_i)$ we have, for any p-value significance threshold $t \in (0, 1)$,

$$E[F_1(t)] = n_1 \cdot \int_{s < t} \frac{P(pW > s)}{P(pY > s)} dF_{pY}(s). \quad (8)$$

We estimate $P(pW > s)$ and therefore $P(pX > s) = P(pW > s)/P(pY > s)$ based on the available observations w_i or, more precisely, on their p-values p_i (assumed again to be increasing: $p_i = p_{(i)}$). Specifically, with $D(s) := |\{p_i \leq s\}|$ and $\hat{n}_1 := (1 - \hat{\pi}_0)n$, for $s \leq p_1$

$$\widehat{P(pW > s)} := \max \left\{ 0, \frac{[n - D(s)] - \hat{n}_0(1 - s)}{\hat{n}_1} \right\}. \quad (9)$$

Since under our assumptions the distribution of pY is the standard uniform, we define

$$P(\widehat{pX} > s) = \min \left\{ 1, \frac{P(\widehat{pW} > s)}{1 - s} \right\}. \quad (10)$$

Using this and a Riemann sum based on the partition defined by the observed p_i to approximate the integral in (8) from below yields:

$$E[\widehat{F_1(t)}] = \hat{n}_1 \cdot \left[\sum_{j=1}^{D(t)} P(\widehat{pX} > p_{j-1})(p_j - p_{j-1}) + P(\widehat{pX} > p_{D(t)})(t - p_{D(t)}) \right], \quad (11)$$

where $p_0 := 0$. Putting it all together we get the following estimator of $\text{FDR}(t)$:

$$\begin{aligned} \widehat{\text{FDR}}_e(t) &= \frac{E[\widehat{F_0(t)}] + E[\widehat{F_1(t)}]}{D(t) \vee 1} = \widehat{\text{FDR}}_e(p_{D(t)}) \\ &+ \frac{\hat{n}_0 \cdot (t - p_{D(t)}) + \hat{n}_1 \cdot \left[\frac{n - D(t)}{\hat{n}_1 \cdot (1 - p_{D(t)})} - \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right]_{[0,1]} (t - p_{D(t)})}{D(t) \vee 1}, \end{aligned}$$

where

$$\widehat{\text{FDR}}_e(p_i) = \frac{\hat{n}_0 \cdot p_i + \hat{n}_1 \cdot \sum_{j=1}^i \left[\frac{n - j + 1}{\hat{n}_1 \cdot (1 - p_{j-1})} - \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right]_{[0,1]} (p_j - p_{j-1})}{i \vee 1}. \quad (12)$$

Note that for $t < p_1$, $D(t) = p_{D(t)} = 0$, and it is easy to check that in this case $\widehat{\text{FDR}}_e(t) = \widehat{\text{FDR}}_{\text{BH}}(t) = n \cdot t$. This makes sense: we have no evidence of correct scores for $t < p_1$ so any score pW_j in that range has to satisfy $pW_j = pY_j$. Note also that, as applied to all the point estimators $\widehat{\text{FDR}}(t)$ we discussed, technically we take the minimum of the computed expression with 1 and in particular we define $\widehat{\text{FDR}}_e \equiv 1$ when $\hat{\pi}_0 = 1$.

It is easy to see that $\widehat{\text{FDR}}_e$ is increasing for $t \in (p_i, p_{i+1})$ and therefore while $\widehat{\text{q-value}}_e$ is technically defined as in (1), using (2) we can readily compute $\widehat{\text{q-value}}_e(p_i)$ for any p_i if we first compute $\widehat{\text{FDR}}_e(p_j)$ for all observed p-values. In practice, this is what our e-mix-max procedure returns.

More precisely, its input is n , the estimated $\hat{\pi}_0$, the ordered list of unique p-values and their multiplicities. The latter is used to allow e-mix-max to handle p-values computed from an empirical CDF which, depending on the sample size from which they are computed, can typically include ties among the estimated p-values. The procedure first estimates $P(\widehat{pW} > p_j)$

for all j using (9) (in $O(n)$ steps). It then proceeds to compute $P(\widehat{pX} > p_j)$ for all j using (10) (again, in $O(n)$ steps). The process is completed with computing $E[F_1(p_j)]$ using (11) (also, in $O(n)$ steps) and finally adding to it $E[F_0(p_i)] = \hat{\pi}_0 \cdot n \cdot p_i$ and dividing by the number of discoveries, which is i for p_i . Thus, the runtime and space complexity of e-mix-max estimation of the n -dimensional vector of FDRs, $\widehat{\text{FDR}}_e(p_i)$, is $O(n)$.

The general protocol for utilizing the point estimators $\widehat{\text{FDR}}(t)$ to control the FDR calls for pre-specifying the tolerable FDR level α and then declaring as discoveries all i for which $\widehat{\text{q-value}}(p_i) \leq \alpha$. Regardless of whether this procedure actually controls the FDR at level α , in practice, scientists engaging in exploratory data analysis often employ the statistically riskier practice of committing to a certain α only in hindsight.

Storey, Taylor and Siegmund suggest that showing that the estimator $\widehat{\text{q-value}}(t)$ is almost surely asymptotically simultaneously (for all t) conservative justifies this retrospective choice of the threshold α [20]. In what follows we give empirical evidence suggesting that (i) $\widehat{\text{q-value}}_e$ allows us to control the FDR at a pre-specified level α , and (ii) $\widehat{\text{q-value}}_e$ is simultaneously consistent. This evidence suggests that it is reasonable, at least for a large number of spectra, to use $\widehat{\text{q-value}}_e$ for retrospectively selecting the FDR level as well.

4 Results

We compare the aforementioned FDR controlling procedures using both real and simulated data. The latter in particular allows us to examine the procedures from multiple angles: do they control the FDR at a prescribed level α ? Are the estimators they are derived from unbiased, consistent, and simultaneously asymptotically consistent? We can look at these questions while varying the parameters of the problem such as the number of spectra n , the proportion of native spectra π_1 , and the separation between correct and incorrect PSM scores. Additionally, we can compare the number of discoveries reported by each of our procedures, and unlike FDR control, this comparison can extend to real data as well.

4.1 Controlling the FDR at a predetermined level

Recall that Benjamini and Hochberg proved that their procedure controls the FDR regardless of the p-values of the false null hypotheses. A natural extension of this to our context is to show that the FDR is controlled regardless of the values of the native matches x_i .⁷

Therefore, we first examined our FDR controlling procedures in this conditional setting: we randomly drew native match scores x_i for the $n_1 = \pi_1 \cdot n$ native spectra. For each of these sets of native scores we repeatedly randomly drew the n foreign match/null scores y_i , thereby obtaining a resulting set of n “target PSM” scores $w_i = \max \{x_i, y_i\}$ ($x_i := -\inf$ for $i > n_1$).

⁷It is not difficult to show that other natural extensions to this PSM context, such as conditioning on w_i , the scores (or p-values) of the target matches, or on the set of native spectra, or conditioning on the scores/p-values of the correct discoveries, do not allow non-trivial control of the FDR.

We then applied the BH, Storey, and e-mix-max FDR controlling procedures to each of the corresponding sets of p-values (see Appendix Section B.3 for details).

Figure 1 (1L), as well as Appendix Figures 1 and 2 (examining additional parameter settings), confirm our anticipated liberal bias of the Storey procedure, as well as the conservative bias of BH. In contrast, our new e-mix-max procedure seems essentially unbiased: the actual FDR in its list of discoveries, as approximated by the empirical mean of the corresponding FDP, largely coincides with the nominal FDR level α .

To gain further confidence in our findings we carried out another set of experiments where we repeatedly drew *both* the x_j and the y_j . The results (Figure 1 (1R), as well as Appendix Figures 3 and 4) are qualitatively the same: Storey is too liberal, BH is too conservative and e-mix-max is essentially unbiased.

As expected, under both types of simulations the biases of the BH and Storey methods increase with the proportion of native spectra π_1 . It is also apparent from the figures that these biases are largely invariant with respect to the size of the set of spectra.

Finally, to verify our assertion that $\hat{\pi}_0$ estimates the fraction of native spectra, we generated 100,000 independent samples of our simulated data with $n = 10000$, $\pi_1 = 0.5$ and estimated $\hat{\pi}_0$ for each such data set using the `smoother` option of the `qvalue` package in R. The median value of $\hat{\pi}_0$ was 0.4955, which is much closer to the proportion of foreign spectra (set to 0.5 in this case) than to the median value of the proportion of incorrect target matches, which was 0.5193.

4.2 Bias of the point estimators of FDR(t)

Following Storey, Taylor and Siegmund we presented the BH, Storey and e-mix-max FDR controlling procedures as random thresholding of the corresponding point estimators of FDR(t): the procedure rejects all null hypotheses with corresponding p-value $p \leq \tau(\alpha)$, where $\tau(\alpha) = \max \{t: \widehat{\text{FDR}}(t) \leq \alpha\}$, or equivalently, $\widehat{\text{q-value}}(p) \leq \alpha$.

Therefore, as a complementary analysis to the previous section, we empirically studied the bias of the corresponding estimators of FDR(t) by looking at the ratio of the empirical mean of $\widehat{\text{FDR}}(t)$ and FDR(t) (Appendix Section B.3.2). Appendix Figures 5 and 6 are consistent with our observations from the previous section: $\widehat{\text{FDR}}_S(t)$ (Storey) has a liberal bias, $\widehat{\text{FDR}}_{\text{BH}}(t)$ (BH) has a conservative bias, while $\widehat{\text{FDR}}_e(t)$ (e-mix-max) seems unbiased.

Consistently with our previous finding we also note that the bias of both $\widehat{\text{FDR}}_S(t)$ and $\widehat{\text{FDR}}_{\text{BH}}(t)$ increase with π_1 , albeit in opposite directions, and the bias seems largely unaffected by the spectra set size.

As the spectra set size, n , and π_1 get smaller, e-mix-max has less sample data to estimate the density of native match scores, particularly for small values of p . Hence, it is not surprising that when both the spectra set and π_1 are relatively small ($n \in \{200, 500\}$ and $\pi_1 \in \{0.1,$

0.3}), e-mix-max is increasingly conservatively biased as p gets smaller. In particular, recall that when t is less than the smallest native match p-value, p_1^x , then $\widehat{\text{FDR}}_{\text{BH}}(t) = \widehat{\text{FDR}}_e(t)$.

4.3 Consistency of the e-mix-max point estimators

Having established that only the e-mix-max estimators, $\widehat{\text{FDR}}_e(t)$ and $\widehat{\text{q-value}}_e(t)$, appear unbiased in the context of our model, we next examined whether or not they appear to be consistent (Appendix Section B.3.2).

Figure 1 (2L), as well as the left column of Appendix Figure 7 show that, regardless of the value of π_1 , as the number of spectra increases, the maximal deviations $\max_{t \in T} \widehat{\text{FDR}}_e(t) - \text{FDR}(t)$ and $\min_{t \in T} \widehat{\text{FDR}}_e(t) - \text{FDR}(t)$, shrink to 0.⁸ This suggests that $\widehat{\text{FDR}}_e$ is simultaneously (i.e., for all t) consistent. As noted by Storey, Taylor and Siegmund, this property is important for establishing that the corresponding q-value derived procedure controls the FDR at any prescribed level—a point which we independently empirically demonstrated earlier.

Repeating this analysis with $\widehat{\text{q-value}}_e(t)$ we got almost identical looking plots, suggesting that the same holds for that estimator as well. However, instead of displaying these twin figures, we chose to further examine this issue using a different perspective. Rather than examining the deviations between $\widehat{\text{q-value}}_e(t)$ and $\text{q-value}(t)$ as n increases we looked at them using the largest n we considered.

The middle column of Appendix Figure 7 shows, for each fixed $t \in T$, the maximum and minimum of $\log[\widehat{\text{q-value}}_e(t)/\text{q-value}(t)]$ across all 100K samples with $n = 70000$ spectra. The graphs suggest that, for large n , not only is the deviation between $\widehat{\text{q-value}}_e(t)$ and $\text{q-value}(t)$ small, but it is small in *relative* terms. This is particularly reassuring for those cases where we resort to selecting a very small FDR.

Finally, in order to lend credence to the practice of retrospectively selecting the FDR threshold we examined the maximal possible deviation, observed among all our 100K samples with $n = 70000$ spectra, between an e-mix-max estimated q-value and the corresponding real q-value (Appendix Sections B.3.2 and B.5). Figure 1 (2R) as well as the right column of Appendix Figure 7 indicate that, in the context of our model and a large spectra set, simultaneously for all possible values there is little difference between an estimated q-value and the actual one. In particular, this suggests that using e-mix-max we can safely engage in exploratory data analysis selecting the desired FDR threshold in hindsight.

4.4 Varying the separation between the native and null match scores

We next looked at the effect of varying the separation between the native and null match scores has on the FDR control of the BH, Storey, and e-mix-max procedures. Specifically,

⁸For computational efficiency we limited the size of the set T to 248 preselected points in $(0, 1)$.

we varied the mean of the distribution that was used to sample the native match scores x_i . Using again the conditional setup, we next generated 100K samples of the $n = 10K$ null match scores y_i to generate 100K sampled sets of target PSM scores and their corresponding p-values per each separation level (Appendix Section B.3.3).

It is evident from Appendix Figure 8 that when the separation between the native and foreign matches increases, the difference between e-mix-max and Storey diminishes: we observe fewer incorrect matches due to native spectra. At the same time, the difference between BH and e-mix-max increases with increasing separation: because BH does not account for native spectra, that method becomes increasingly conservative as more of the native spectra turn out to be correct. Regardless, e-mix-max control of the FDR seems extremely tight: the empirical mean of the FDP in its list of discoveries is almost identical to the selected FDR threshold.

4.5 Number of discoveries

The flip side of controlling the FDR is the number of discoveries the method reports. Unlike accuracy of FDR estimation, this number is something we can examine in real data. We investigated the reported number of discoveries of the BH, Storey, and e-mix-max methods over a range of FDR thresholds for three data sets (yeast, worm and *Plasmodium*) analyzed using two different search engines (Tide and MS-GF+, Appendix Section B.6). In Figure 1 (3L) and Appendix Figure 9, BH and Storey are plotted relative to e-mix-max, because our analysis hitherto showed that e-mix-max is apparently unbiased. The results of these analyses are not surprising: because Storey underestimates the FDR it reports more discoveries than the apparently unbiased e-mix-max, and vice-versa for BH. Assuming that our simulation is qualitatively correct, the respective increase and decrease in the numbers of discoveries reported by BH and Storey in the real data is due to their respective conservative and liberal biases.

4.6 Comparison with decoy-dependent estimation

Finally, it is instructive to compare the accuracy in FDR estimation of the e-mix-max procedure with its decoy-dependent counterpart, mix-max, as well as the widely used T-TDC procedure. We do so using simulated data and a similar methodology as in Figure 1 (1R). This experiment (Figure 1 (3R) as well as Appendix Figures 10 and 11) shows that e-mix-max is essentially unbiased, whereas, as we noted in [14], both mix-max and T-TDC tend to be too liberal for small FDR thresholds. The liberal bias of T-TDC for small FDR threshold is also suggested by the analysis of [1].

In addition to the mean of the actual FDP (the empirical FDR), the above mentioned figures also include the 0.05 and 0.95 quantiles, thereby quantifying the variability of these three methods. As expected in the context of our model, the variability of the decoy-dependent methods is consistently larger than the variability of e-mix-max. Furthermore, as the number of spectra increases, the overall variability that all three procedures exhibit diminishes, as expected. At the same time, it is clear from these figures that the variability also decreases with increasing proportion of native spectra π_1 .

Note that the results in these figures primarily aim to contrast e-mix-max's FDR control with that of the two decoy-dependent methods. As such, these results do not highlight the main advantages of mix-max relative to T-TDC, which we have pointed out previously [13]: reduced variability in the discovery list and a slight increase in the number of discoveries per FDR threshold.

5 Discussion

We tackle the problem of controlling the FDR in the context of imperfect matching to an incomplete database. This problem differs from the classical multiple testing one in that, due to the incomplete database, our products fall into one of two categories: the native products that match sources in the target database and the complementary, unmatchable foreign products. These categories lead to two different types of false discoveries: any foreign product that is matched is a false discovery, but due to the imperfect matching procedure native products can be incorrectly matched, thereby generating a false discovery.

Scientists have been practically dealing with this problem in the context of tandem mass spectrometry spectrum identification by peptide database search. In practice, control of the FDR in this setup is mostly achieved by TDC, which estimates the FDR by searching against a competing database of decoy peptides. The main advantage of TDC is that it has essentially no restrictions on the PSM score function. In particular, TDC does not require computing a p-value of the PSM, something that is essential for the hypothesis testing context. The downsides of TDC are that it randomly eliminates a subset of high-scoring correct matches, its reported list of discoveries is consequently quite variable, and it is liberally biased for small FDR thresholds.

We recently developed mix-max, an FDR controlling procedure that is less variable and more accurate than TDC but requires a calibrated score. Because in practice we calibrated the score by computing an empirical p-value, it was natural to ask whether we can improve our FDR control by using these p-values as such, rather than as calibrated scores.

We stress that the p-values of the matches are computed relative to a model in which the database is randomly drawn according to a null distribution (e.g., shuffling the peptides). This corresponds to a null assumption that the spectrum is foreign, rather than the more desirable hypothesis that the PSM is incorrect. Indeed, it is impossible to compute a p-value relative to the assumption that the PSM is incorrect, since the answer would be different depending on whether the spectrum is foreign or native. (We will return to this point below.)

It follows that the fraction of true null hypotheses, usually denoted by π_0 , coincides here with the fraction of foreign spectra (products) rather than incorrect matches. This observation implies that the adaptive BH procedures that factor an estimate of π_0 into the FDR control (such as the version we refer to as "Storey") are too liberal in the context of spectrum identification, because these methods fail to account for false discoveries due to native spectra. At the same time the BH procedure, which essentially ignores native spectra altogether, is still conservative in this PSM context. Our novel e-mix-max is designed to find

the middle ground between BH and Storey by separately accounting for the two types of false discoveries: the ones due to native and to foreign spectra.

Clearly, more work is required to establish our empirical findings, possibly using a more tractable estimator of π_0 in e-mix-max. However, our evidence suggests that the popular procedures of BH and Storey are, respectively, not optimal or inappropriate for controlling the FDR in the context of searching an incomplete database. At the same time, e-mix-max provided very tight control of the FDR in our simulations. Moreover, the apparent simultaneous consistency of the point estimators $\widehat{q\text{-value}}_e(t)$ suggests that when analyzing a large spectra set e-mix-max is well suited for the practice of retrospectively selecting the desired FDR threshold—a practice that is commonly used in exploratory data analysis in general, and in spectra identification in particular. Our findings in simulated data are echoed in the number of discoveries reported by these methods when applied to real mass spectrometry data (where the accuracy cannot be verified).

While e-mix-max itself has a linear time and space complexity, computing the p-values might be a much more computationally demanding task. The method we used here relies on brute force construction of empirical CDFs generated by extensive MC sampling. However, more efficient dynamic-programming based methods have recently been suggested in the literature [16, 11].

In this paper we concentrated on comparing the FDR controlling procedures that use p-values: Storey, BH, and e-mix-max. If the p-values are not computable, then the decoy dependent estimators, particularly T-TDC and mix-max, could be used to control the FDR. However, such estimation is inevitably subject to the variability of the randomly chosen decoy. This effect is particularly strong for T-TDC, where some high scoring target PSMs are randomly filtered out as they lose to corresponding decoy PSMs. Thus, when p-values are available, e-mix-max should be preferred for its reduced variability as well as its better FDR control, particularly for smaller FDR thresholds where T-TDC and mix-max are somewhat liberally biased.

Going back to the problem with computing the p-value relative to the null hypothesis that the match is incorrect, it is fair to ask if we can compute such a p-value assuming that we know the proportion of native products, as well as the distribution of the native matches. The short answer to this is that, in this case, this is possible; however, even ignoring the fact that in practice we do not know π_1 nor the distribution of x_j , there is another obstacle to applying known FDR-controlling procedures. Namely, in this new theoretical context the hypotheses are not independent: if we see many rejections among the first hypotheses we test, then the following ones are less likely to represent native products (whose total number is assumed known). Moreover, this dependence seems to rule out the PRDS condition of [4], and at the same time, because of the additional false discoveries due to native products, condition (8) of [20] is also violated. Thus, it seems there is no obvious “ready made solution” to this problem, even under our extra assumptions above.

Finally, it is important to keep in mind that while for most of the paper we used the terminology of tandem mass spectrometry, the problem we address here can be more

generally formulated as controlling the FDR among imperfect matches to an incomplete database.

An R package called `dbSearchFDR` containing functions for FDR control in the context of imperfect matches to an incomplete database, including the novel `e-mix-max`, is available for download at <http://www.maths.usyd.edu.au/u/uri/dbSearchFDR/>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful for the anonymous reviewers and associate editor for their numerous comments and insightful suggestions: addressing the raised issues significantly improved the manuscript. This work is supported in part by NIH award P41 GM103533.

References

1. Barber RF, Candès Emmanuel J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*. 2015; 43(5):2055–2085.
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*. 1995; 57:289–300.
3. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational Behavioral Statistics*. 2000; 25(60–83)
4. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Statist*. 2001; 29(4):1165–1188. 08.
5. Benjamini YoavKrieger Abba M, Yekutieli Daniel. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006; 93(3):491–507.
6. Cerqueira Fabio R, Graber ArminSchwikowski BennoBaumgartner Christian. Mude: A new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification. *Journal of Proteome Research*. 2010; 9(5):2265–2277. [PubMed: 20199108]
7. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*. 2007; 4(3):207–214. [PubMed: 17327847]
8. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology*. 2010; 604(55–71)
9. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*. 1994; 5:976–989. [PubMed: 24226387]
10. Genovese ChristopherWasserman Larry. A stochastic process approach to false discovery control. *Ann Statist*. 2004; 32(3):1035–1061. 06.
11. Howbert JJ, Noble WS. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular and Cellular Proteomics*. 2014; 13(9):2467–2479. [PubMed: 24895379]
12. Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. *BMC Bioinformatics*. 2012; 13(Suppl. 16):S2.
13. Keich U, Kertesz-Farkas A, Noble WS. Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of Proteome Research*. 2015; 14(8):3148–3161. [PubMed: 26152888]
14. Keich U, Noble WS. On the importance of well calibrated scores for identifying shotgun proteomics spectra. *Journal of Proteome Research*. 2015; 14(2):1147–1160. [PubMed: 25482958]
15. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of Proteome Research*. 2008; 7:3354–3363. [PubMed: 18597511]

16. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJ, Pevzner PA. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular and Cellular Proteomics*. 2010; 9(12):2840–2852. [PubMed: 20829449]
17. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*. 2010; 73(11): 2092–2123. [PubMed: 20816881]
18. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*. 2002; 64:479–498.
19. Storey JD. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*. 2003; 31(6):2013–2035.
20. Storey JD, Taylor JE, 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*. 2004; 66:187–205.
21. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100:9440–9445. [PubMed: 12883005]

Figure 1.

(1L) Conditional FDR control: $\log$ of the mean of the FDP over the selected FDR threshold (10K spectra, $\pi_1 = 0.5$). (1R) FDR control: same as (1L) except the data was generated by sampling both the native scores x_i and the null scores y_j each time. (2L) Examining extreme values of $\max_{t \in T} \widehat{\text{FDR}}_e(t) - \text{FDR}(t)$ and $\min_{t \in T} \widehat{\text{FDR}}_e(t) - \text{FDR}(t)$ over 100K samples suggests the simultaneous consistency of $\widehat{\text{FDR}}_e(\pi_1 = 0.5)$. (2R) Lower and upper bounds on the actual q-value of any e-mix-max estimated q-value (the line $y = x$ is added for reference, 70K spectra, $\pi_1 = 0.5$). (3L) The ratio of the number of e-mix-max to Storey/BH discoveries as a function of the selected FDR threshold (yeast dataset searched with Tide). (3R) The mean as well as the 0.05 and 0.95 quantiles of the ratio of the actual FDP to the selected FDR threshold for e-mix-max, mix-max and T-TDC (10K spectra, $\pi_1 = 0.3$).