Journal of proteome research

Article

pubs.acs.org/jpr

Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics

Uri Keich,*[,†] Attila Kertesz-Farkas,[‡] and William Stafford Noble*[,‡,§]

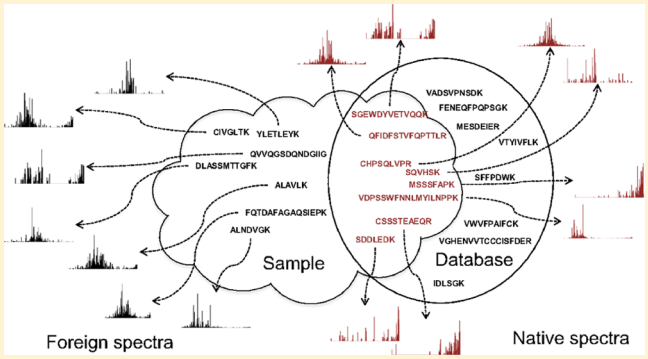
[†]School of Mathematics and Statistics F07, University of Sydney, Sydney, New South Wales 2006, Australia

[‡]Department of Genome Sciences, University of Washington, Foege Building S220B, 3720 15th Avenue North East, Seattle, Washington 98195-5065, United States

[§]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-5065, United States

Supporting Information

ABSTRACT: Interpreting the potentially vast number of hypotheses generated by a shotgun proteomics experiment requires a valid and accurate procedure for assigning statistical confidence estimates to identified tandem mass spectra. Despite the crucial role such procedures play in most high-throughput proteomics experiments, the scientific literature has not reached a consensus about the best confidence estimation methodology. In this work, we evaluate, using theoretical and empirical analysis, four previously proposed protocols for estimating the false discovery rate (FDR) associated with a set of identified tandem mass spectra: two variants of the target-decoy competition protocol (TDC) of Elias and Gygi and two variants of the separate target-decoy search protocol of Käll et al. Our analysis reveals significant biases in the two separate target-decoy search protocols. Moreover, the one TDC protocol that provides an unbiased FDR estimate among the target PSMs does so at the cost of forfeiting a random subset of high-scoring spectrum identifications. We therefore propose the mix-max procedure to provide unbiased, accurate FDR estimates in the presence of well-calibrated scores. The method avoids biases associated with the two separate target-decoy search protocols and also avoids the propensity for target-decoy competition to discard a random subset of high-scoring target identifications.



KEYWORDS: *mass spectrometry, spectrum identification, false discovery rate*

1. INTRODUCTION

A typical shotgun proteomics produces thousands of tandem mass spectra (hereafter referred to simply as "spectra"), each of which can be tentatively assigned a corresponding peptide using a database search procedure (reviewed in ref 1). Some of those assignments will be correct, while others will be false, and the statistical problem we face is estimating the proportion of false peptide assignments among all assignments whose quality exceeds a given threshold. This proportion is commonly referred to as the false discovery rate (FDR), and the list of reported discoveries is typically set so that the estimated FDR is less than some desired threshold, say 0.05 (5%). In practice, solutions to this problem commonly rely on comparing the spectrum identifications obtained from searching a real (target) peptide database with those obtained from searching against a decoy database of shuffled or reversed peptides.

At least four distinct, decoy-based FDR estimation protocols have been advanced in the literature. The first, proposed by Elias and Gygi, finds the best matching peptide for each spectrum relative to a concatenated target-decoy database and estimates the FDR among all peptide-spectrum matches (PSMs) above a specified score threshold.[2] In the example shown in Figure 1, each of the five observed spectra is associated with a top-scoring target peptide and a corresponding top-scoring decoy peptide. In some cases, the top score is less than a specified score threshold, in which case no peptide is indicated (e.g., the third spectrum has no corresponding decoy peptide). A key component of the Elias and Gygi strategy is target-decoy competition (TDC), in which the top-scoring target and decoy peptides compete with one another and only the higher scoring of the two peptides is retained in the final list. In practice, this competition is carried out by searching the spectra against a concatenated database containing the target and decoy peptides. Thus, in Figure 1, the reported list ("C-TDC" for combined TDC) contains three target peptides and two decoys. The estimated FDR is simply twice the number of decoys in the list, divided by the total length of the list (in this case, $(2 \times 2)/5 = 0.8$).

Received: January 29, 2015
Published: July 8, 2015

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

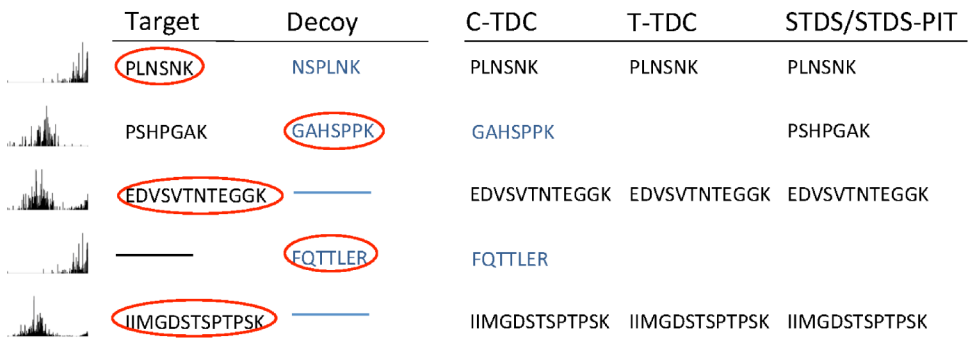


Figure 1. Comparison of target-decoy FDR estimation procedures. Each spectrum is associated with a top-scoring target peptide and decoy peptide, although only peptides that score above a specified threshold are displayed. The higher scoring of the two peptides is circled. The corresponding lists of PSMs for C-TDC, T-TDC, and STDS/STDS-PIT are shown.

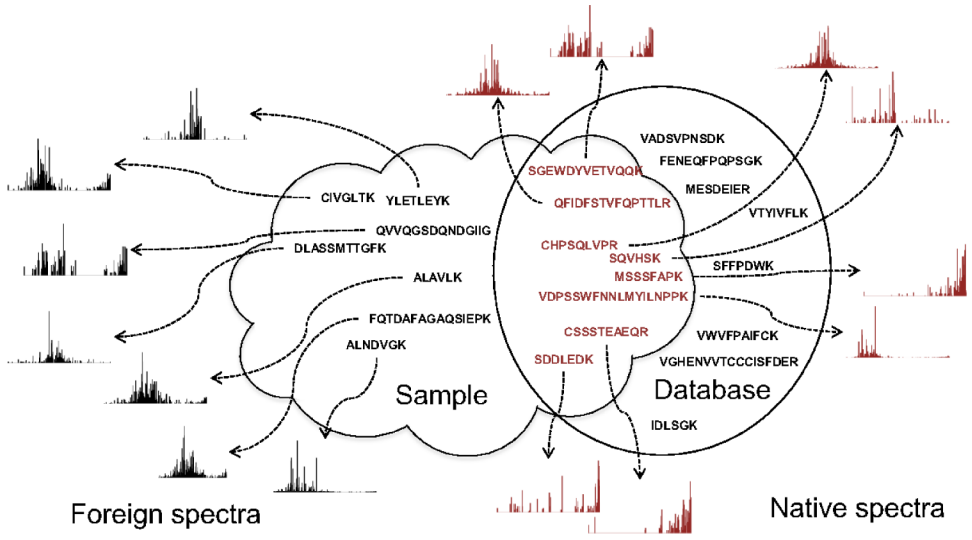


Figure 2. Foreign and native spectra. The sets of peptides present in the sample and the peptides in the database overlap one another. Only peptides present in both (red) generate native spectra that have a chance to be correctly identified by the database search algorithm.

One apparent drawback of the C-TDC protocol is that the reported list of identified spectra contains a mixture of target and decoy peptides. In practice, of course, the user is typically interested only in the spectra that match a target peptide. Accordingly, the target-only variant of target-decoy competition (T-TDC) eliminates decoy identifications from the reported list and adjusts the FDR estimate accordingly.[3,4] The FDR estimate is simply the number of decoys divided by the number of targets (i.e., 2/3 = 0.67). Hence, for a fixed score threshold, the T-TDC protocol yields the same number of target identifications as C-TDC but a lower estimated FDR.

Unfortunately, the target-decoy competition that is the basis for both of these methods leads to two closely related problems. First, the competition occasionally eliminates a high-scoring target PSM because the corresponding decoy PSM happened to achieve an even higher score. This happened, for example, in Figure 1 for the target peptide "PSHPGAK," which matches the second observed spectrum with a high score but is not reported by either TDC method. Second, when using randomly generated decoy peptides, the TDC method exhibits an undesirable variability because the filtered target PSMs differ each time the procedure is run. Note that the use of reversed, rather than shuffled, decoy peptides simply hides this problem by arbitrarily fixing the decoys and the corresponding filtered peptides.

To avoid randomly discarding a small proportion of the high-scoring target PSMs, Käll et al. proposed an alternative method, which we call "separated target-decoy search" (STDS), in which the decoy PSMs are used separately to estimate the FDR among the target PSMs.[5] In STDS, all target PSMs above a specified threshold are reported to the user. For the small set of PSMs in Figure 1, the corresponding FDR is the total number of above-threshold decoy PSMs divided by the number of above-threshold target PSMs (i.e., 3/4 = 0.75). A second, more sophisticated approach proposed by Käll et al., which we call "STDS-PIT," involves estimating one additional parameter, the "percentage of incorrect targets" (PIT), from the data. In STDS-PIT, the final FDR estimate is the STDS estimate multiplied by the PIT; however, as we discuss later, the inclusion of this parameter is problematic.

We recently argued that because the STDS methods estimate the significance of each target PSM using the set of all decoy PSMs, their use should be restricted to search engines that use fairly well-calibrated scores.[6] Intuitively, a PSM score function is well-calibrated with respect to spectra (and a null peptide database model) if a score of x assigned to spectrum σ_i has the same meaning or significance as a score of x assigned to spectrum σ_j. More precisely, if S_i is the score of the best match to spectrum σ_i in a randomly drawn database then the score is calibrated if for any spectra σ_i and σ_j, $P(S_i \geq S_j) = P(S_j \geq S_i)$.

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

Moreover, we demonstrated that when the score is calibrated, STDS-PIT typically reports substantially more discoveries than T-TDC does at a given FDR threshold. In the current work, to better understand the source of the apparent power advantage displayed by STDS-PIT, we introduce a formal statistical model that allows us to rigorously evaluate the aforementioned FDR estimation methods: C-TDC, T-TDC, STDS, and STDS-PIT.

We show in the context of our model that the two protocols based on target-decoy competition are asymptotically accurate in estimating the FDR within their respective lists of PSMs. (Although, for C-TDC, that list is not one we are typically interested in.) On the other hand, the STDS procedure is conservative (overestimating the true FDR) and the STDS-PIT method is anticonservative (underestimating the true FDR). Consequently, motivated by the desire for a statistically unbiased method that does not arbitrarily discard some high-scoring PSMs, we designed a novel method, the mix-max procedure, that extends the STDS-PIT method and is demonstrably unbiased. Similar to STDS-PIT, the mix-max procedure requires calibrated scores, which can be obtained using existing, calibrated score functions[7,8] or by postprocessing scores using our recently described nonparametric calibration procedure.[6] Open-source implementations of the T-TDC and mix-max estimation procedures are available as part of the Crux mass spectrometry software toolkit (http://cruxtoolkit.sourceforge.net, assign-confidence command).[9]

2. RESULTS

2.1. Theoretical Model of the Spectrum Identification Problem

To rigorously evaluate the four FDR estimation methods, we designed a simple probabilistic model of decoy-based FDR estimation. A key component of the model is its division of the set of spectra into two subsets (Figure 2): the "native" spectra that were generated by a peptide present in the target database and the "foreign" spectra that were generated by contaminant peptides, peptide variants that are not in the given database, peptides with unexpected post-translational modifications and nonpeptide species, as well as spectra for which the charge state was not correctly identified. Note that this native/foreign distinction applies to spectra and is hence orthogonal to the more familiar target/decoy distinction, which applies to peptides. As will become apparent, the distinction between foreign and native spectra is a critical component of our model.

Given the foreign/native distinction, the optimal target PSM score of a native spectrum σ_i can be defined as $W_i = \max\{X_i, Y_i\}$, where X_i is the score of σ_i relative to the peptide that generated it and Y_i is the score of the best match of σ_i to the remainder of the target database. (For reference, all of our notation is summarized in Table 1.) Note that W_i is observed but X_i and Y_i are not. We denote by Z_i the observed score of the (optimal) PSM between σ_i and the decoy database. We can then distinguish between three different FDRs, relative to a given score threshold:

(1) F_C/D_C, the FDR in the combined list of target and decoy PSMs after TDC, which is what C-TDC aims to estimate.

(2) F_T/D_T, the FDR in that same list with the decoy PSMs removed, which is what T-TDC aims to estimate.

(3) F/D, the FDR in the complete set of target PSMs, which is what the STDS and STDS-PIT (as well as our newly proposed mix-max) procedures aim to estimate. Among these three options, we argue that the third (F/D) is the most useful to ascertain. The first FDR (F_C/D_C) includes decoys, which typically are not of direct scientific interest. And both the first and second FDRs (F_C/D_C and F_T/D_T) exclude potentially valuable, high-scoring target PSMs if they happen to lose the target-decoy competition.

Table 1. Variables and Their Definitions[a]

variable	definition
db	target peptide database
dc	decoy peptide database
$db \oplus dc$	combined target and decoy database
Σ	all spectra
n_Σ	number of spectra
Σ_1	native spectra
Σ_0	foreign spectra
π_1	proportion of native spectra
π_0	proportion of foreign spectra
σ_i	single spectrum
x_i/X_i	score of the match between the ith spectrum and the peptide that generated it ($-\infty$ if the spectrum is foreign)
y_i/Y_i	score of the best match between the ith spectrum and the irrelevant part of the target database
z_i/Z_i	score of the best match between the ith spectrum and the decoy database
w_i/W_i	$\max(X_i, Y_i)$
T	score threshold
α	FDR threshold
$\mathcal{F}$	the event the PSM between σ and its best match in the concatenated database is a false positive
D	number of discoveries in a search of the target database
F	number of false positive discoveries in a search of the target database
D_C	number of discoveries in the combined list of discoveries when searching the concatenated database
F_C	number of false positive discoveries in the combined list of discoveries when searching the concatenated database
D_D	number of decoy discoveries in a search of the concatenated database
F_D	number of false positive decoy discoveries in a search of the concatenated database (same as D_D)
D_T	number of target discoveries in a search of the concatenated database
F_T	number of target false discoveries in a search of the concatenated database
$GP(\sigma)$	the peptide that generated the native spectrum σ
$S_{DB}(\sigma, l)$	score of the match between spectrum σ and peptide $l \in$ DB in the context of database DB
$S(\sigma, \mathrm{DB})$	score of the best match of σ in the peptide database DB: $\max_{l \in db} S_{DB}(\sigma, l)$

[a]When two versions are listed, the capitalized form stands for the random variable, whereas the lowercase is a specific realization thereof.

2.2. Two out of These Four Existing Estimation Methods Are Asymptotically Unbiased

Using our model, we first investigated the Elias and Gygi target-decoy competition (C-TDC) and its target-only variant (T-TDC). Both of these methods rest upon the assumption that for each i the distributions of Y_i and Z_i are identical and independent given X_i, and, in particular, (assuming no ties or that ties are randomly broken) $P(Z_i > Y_i \mid \mathcal{F}) = 1/2$, where $\mathcal{F} = \{\max(Y_i, Z_i) > \max(X_i, T)\}$ is the event: the PSM between σ_i and its best match in the concatenated database $db \oplus dc$ is a false positive. We used this assumption to prove (see Methods 4 and Supplementary Note 1 in the SI) that both methods consistently estimate the FDR for their respective lists of discoveries: C-TDC for the concatenated list of target and

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

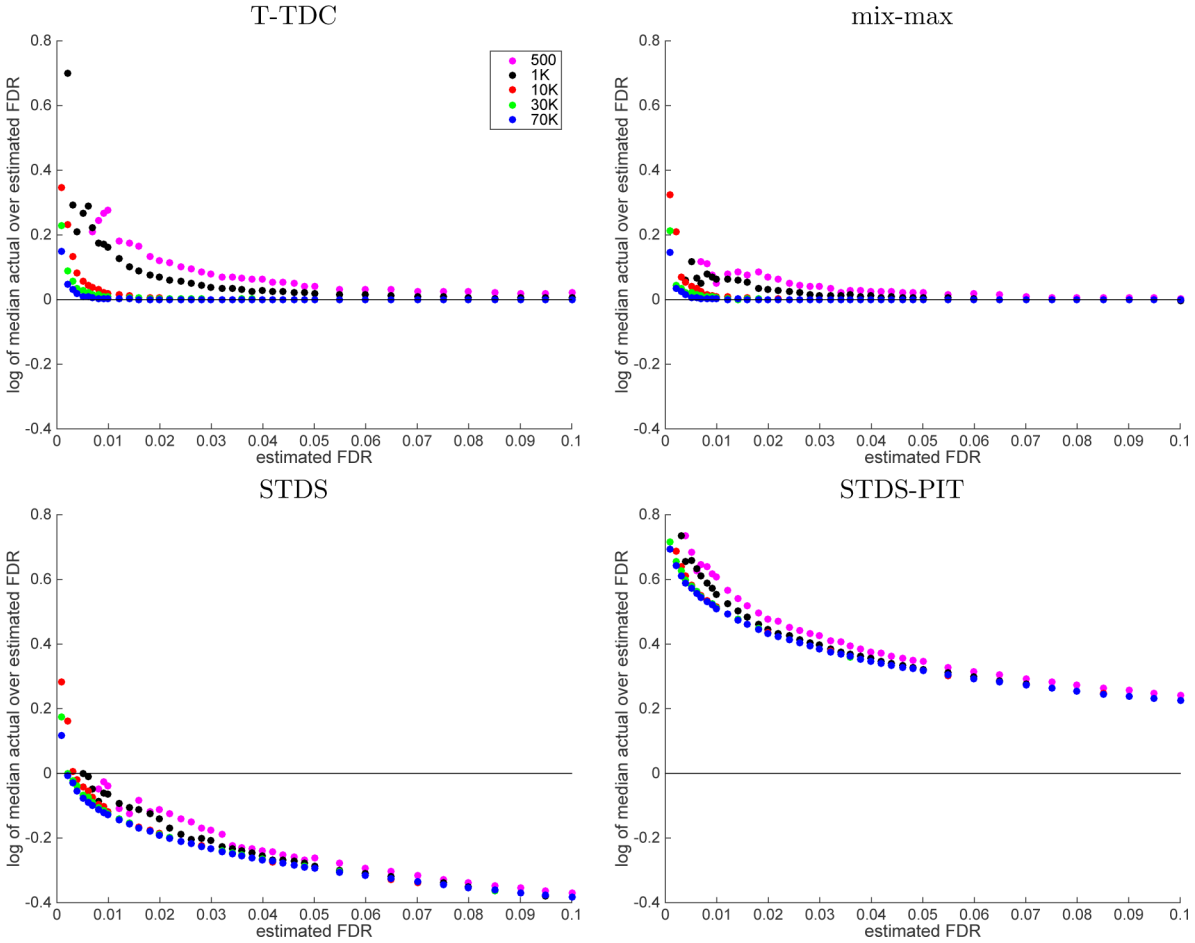


Figure 3. Accuracy of estimated FDR (mixture model). Each panel plots, as a function of estimated FDR, the logarithm of the median ratio between the actual FDR and the nominal one (so a value of 0 means perfect median estimation). The data were generated using the normal mixture model (see Methods 4), and the number of spectra increased from 500 to 70 000, keeping the native spectra rate at 0.5. The medians are calculated at each FDR value with respect to 10K random draws of both native and foreign spectra. Because all of the plots are on the same scale, it is easy to see that STDS overestimates the true FDR while STDS-PIT underestimates it. Both T-TDC and mix-max become increasingly more accurate as the spectrum set or the nominal FDR level becomes larger, but mix-max seems slightly more accurate. In the case of both T-TDC and mix-max and small spectrum sets (500 and 1000) the median estimated FDR jumps from 0 to a number greater than 0.001; hence, the logarithm of the ratio to the nominal FDR is not defined for some small nominal FDR values. When the average separation between the correct PSM scores and the false PSM scores is further increased we noted similar results, albeit STDS-PIT suffers a reduced bias whereas the opposite holds for STDS (Supplementary Figure 2 in the SI).

decoy PSMs and T-TDC for the filtered list of target PSMs. This finding suggests that C-TDC should not be used to estimate the target-PSM FDR in the concatenated list (F_T/D_T) and is consistent with recently reported results suggesting that C-TDC is more conservative than T-TDC.[3] (See Supplementary Figure 1 in the SI.) We also showed analytically that C-TDC conservatively estimates the FDR in the target-only search (F/D).

We then used our model to analyze STDS and STDS-PIT, both of which implicitly require that all of the incorrect PSM scores Y_i and Z_i are drawn independently from the same null distribution. Note that this is a much stronger assumption than the one required by the TDC methods, and it essentially amounts to using a calibrated score. We argue that, in this context, STDS-PIT underestimates the true FDR (Supplementary Note 2 in the SI and Methods 4.1). Specifically, STDS-PIT relies on estimating the "percentage of incorrect target" PSMs, which is supposed to be the overall rate of false discoveries; however, we demonstrate that what the method actually estimates is the proportion π_0 of foreign spectra. Essentially, the STDS-PIT estimate ignores all of the incorrect targets that are attributed to the native spectra.

We note that in a followup work Käll et al. proposed estimating the PIT using an alternative method,[10,11] which can improve the results of STDS-PIT because it often yields a lower estimate of the PIT compared with the original approach; however, as we explain in the Methods 4 section, the specific implementation of that revised method is not theoretically sound. Moreover, as we show in Supplementary Figure 3 in the SI, in the context of our model, the revised method, which we refer to as STDS-PIT+qvality, still shows considerable liberal bias, which is quite close to that of the original method. Therefore, the analysis we present here concentrates on the original STDS-PIT method.

At the same time, it is clear that the simple STDS procedure overestimates the number of false discoveries. Indeed, the STDS estimate, eq 4, corresponds to the case where there are no native spectra (because it assumes $\pi_0 = 1$) so all spectra are foreign. Because the estimated FDR among the native spectra PSMs should generally be much lower than the estimated rate

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

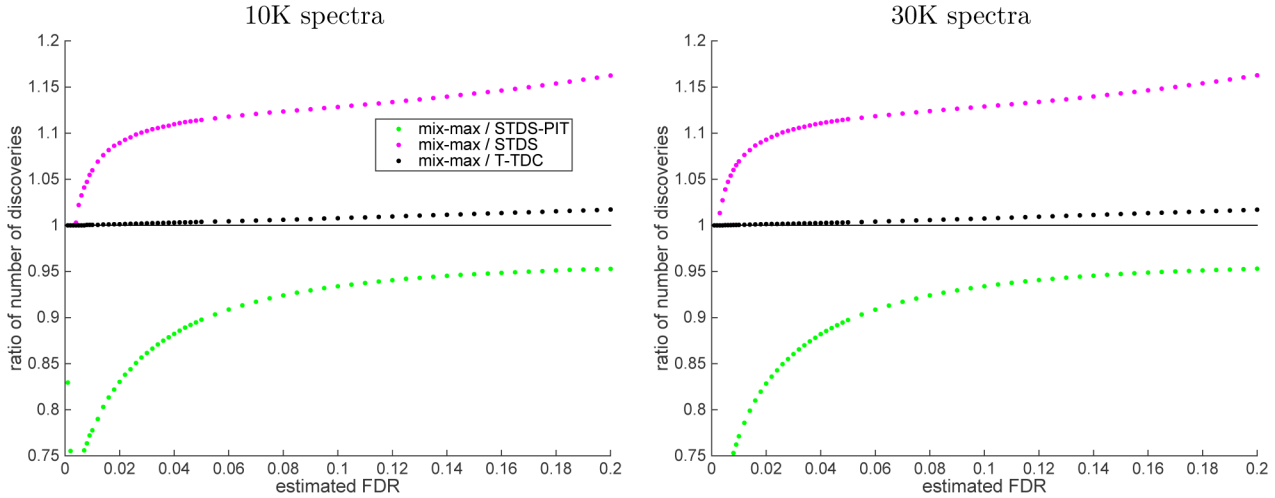


Figure 4. Median ratios of number of discoveries. Both panels plot, as a function of estimated FDR, the median ratio of the number of mix-max to T-TDC/STDS/STDS-PIT discoveries. The data were generated using the normal mixture model (see Methods 4), and the number of spectra was 10K for the left panel and 30K for the right panel. The native spectra rate was set to 0.5 with each spectra set drawn 10K times. Consistent with STDS overestimating the true FDR (Figure 3), it reports fewer discoveries than mix-max, which estimates the FDR quite accurately. Conversely, STDS-PIT underestimates the FDR; hence, it reports more discoveries than mix-max. Less obvious is the subtle but consistent trend of increasingly more mix-max than T-TDC discoveries (black) as the FDR level increases. The results are qualitatively similar for the other sizes of spectrum sets we looked at: 500, 1K, and 70K.

among the foreign spectra (the true rate of which is by definition always 100%), it follows that the STDS method is overestimating the FDR and hence is overly conservative.

2.3. Mix-Max Approach

Motivated by these observations and by our desire for a method that avoids the drawbacks of target-decoy competition, we designed a mixture-maximum (mix-max) FDR estimation procedure that reports all sufficiently high scoring target PSMs while consistently estimating the FDR under the same, stricter assumption that Y_i and Z_i are drawn independently from the same null distribution. The mix-max approach separately estimates the number of false discoveries due to foreign spectra and due to native spectra. The first part essentially follows the STDS-PIT approach; that is, like STDS-PIT, mix-max estimates properties of the null distribution from the decoy set. The second is a bit more involved and requires estimating the distribution of W_i for a native spectrum (Methods 4). We define the resulting mix-max FDR estimation as

$$\frac{\hat{\pi}_0 \cdot \sum_{j=1}^{n_\Sigma} 1_{z_j > T} + (1-\hat{\pi}_0) \cdot \sum_{z_j > T} \left[\frac{\sum_k 1_{w_k \le z_j}}{(1-\hat{\pi}_0)\cdot \sum_k 1_{z_k \le z_j}} - \frac{\hat{\pi}_0}{1-\hat{\pi}_0} \right]_{[0,1]}}{\sum_i 1_{w_i > T}}$$

where $1_{z_j>T}$, $1_{w_i>T}$, $1_{w_k \le z_j}$, and $1_{z_k \le z_j}$ are 1 or 0, depending on whether the corresponding inequality holds, and where $\hat{\pi}_0$ is the estimated proportion of foreign spectra, T is the score threshold, w_i and z_j refer to the observed target and decoy PSM scores, and $[x]_{[0,1]} := \max\{0, \min\{1,x\}\}$ ensures that x remains an acceptable probability value. In the equation, the numerator is the sum of two terms, corresponding to the estimated number of false positives due to foreign and native spectra, respectively, and the denominator is the observed number of accepted (target) PSMs.

2.4. Simulation Results

We first sought to experimentally validate our theoretical analysis using simulated data. The advantage of using simulated results is that we know which PSMs are "true" (drawn from the true PSM distribution) and which are "false" (drawn from the false PSM distribution); therefore, we can gauge the accuracy of our estimates in the context of our model.

In the simulations we drew target and decoy "optimal PSM" scores from two different Gaussian distributions (Methods 4). More specifically, each "native PSM" score was set to the maximum of a randomly drawn false PSM score (Y_i) and a randomly drawn true PSM score (X_i), whereas each "foreign PSM" score and each decoy PSM score (Z_i) were drawn according to the same distribution of false PSM scores as (Y_i).

We first examined the issue of the PIT, or percentage of incorrect target PSMs, versus the percentage of foreign spectra. To do so we used the recipe of Käll et al. (also described here in the Methods 4 section), which estimates what is known in the FDR literature as the proportion of true null hypotheses, π_0. Our claim is that because in this context the null distribution is estimated only from decoy PSMs it corresponds to a null hypothesis that the spectrum is foreign. Thus, the estimated value $\hat{\pi}_0$ corresponds to the proportion of foreign spectra rather than the PIT.

Simulating 100 K as well as 10 K spectra, of which 50% are native and 50% are foreign, the median estimated value of $\hat{\pi}_0$ across 10 K independent experiments was 0.496 for both the 10 K and 100 K spectrum sets, whereas the median of the actual proportion of false PSMs among all target PSMs ($X_i < Y_i$) was 0.519 for both spectrum sets. This result agrees with our claim that $\hat{\pi}_0$ represents the estimated proportion of foreign spectra, rather than the overall fraction of incorrect target PSMs, as originally claimed.

We next looked at the accuracy of the FDR estimation procedures when applied to data randomly generated according to the same mixture model. Figure 3 shows that, consistent with our analysis, the accuracy of T-TDC in estimating the FDR in the filtered list of target PSMs (F_T/D_T) improves with the size of the spectrum set. Noticeably, T-TDC underestimates the true FDR for small spectrum sets, a fact that we will return to later on. Similarly, as our analysis predicts, STDS consistently overestimates and STDS-PIT consistently underestimates the

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

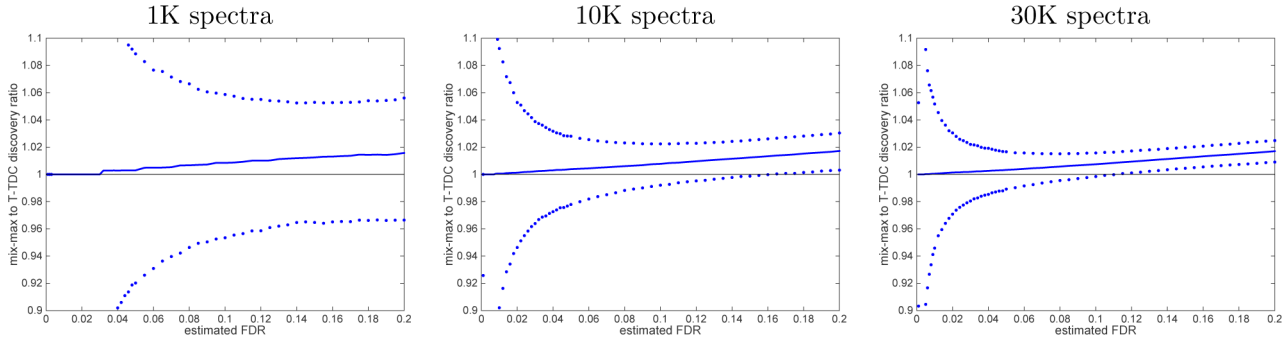


Figure 5. Quantiles of ratios of mix-max to T-TDC number of discoveries. Each panel plots, as a function of estimated FDR, the 0.05/0.5/0.95 quantiles of the ratio of the number of mix-max to T-TDC discoveries. The data were generated using the normal mixture model (see Methods 4), and the number of spectra increased from 1K for the left panel to 30K for the right panel. The native spectra rate was set to 0.5, and the number of draws of both target and decoy PSM scores was 10K for all spectrum sets. The results are qualitatively similar for the other sizes of spectrum sets we looked at: 500 and 70K.

true FDR in the unfiltered list of target PSMs (F/D). The mix-max procedure seems to offer the best accuracy throughout in estimating F/D: it is even somewhat more accurate than T-TDC, which is estimating the FDR in the filtered list of target PSMs (F_T/D_T), and the procedure is only somewhat inaccurate for small spectrum sets and small FDR rates. As expected, we also find that C-TDC is increasingly more accurate in predicting the FDR in the combined list of target and decoy discoveries (F_C/D_C) as the size of the spectrum set increases, whereas C-TDC is consistently overestimating the FDR in the target-only list of discoveries (Supplementary Figure 1 in the SI). Finally, we note that STDS-PIT+qvality[10,11] improves very little on the liberal bias of STDS-PIT in this setup (Supplementary Figure 3 in the SI).

In light of the results in Figure 3, it is not surprising that STDS yields fewer discoveries and STDS-PIT yields more discoveries than mix-max at a given FDR (Figure 4, magenta and green, respectively), corresponding to the two STDS methods' respective tendencies to over- and underestimate the FDR; however, the fact that mix-max yields a small but consistently increasing larger percentage of discoveries than T-TDC does not seem to be a result of a difference in the accuracy of the FDR estimation (Figure 3). Instead, this trend most likely can be attributed to the fact that mix-max does not randomly filter its target PSMs, which, given that the score is calibrated, means swapping the filtered target PSMs with ones that are less likely to be "correct," or drawn from the native distribution in our case.

The exact values we observe in Figure 4 depend on several parameters of our model, including, for example, the proportion of native spectra. The latter was set to 0.5 in that Figure. By varying this proportion we observed that the difference between the methods diminishes as the proportion of native spectra decreases (Supplementary Figure 5 in the SI). Similarly, the difference between the number of discoveries is smaller when the average separation between the correct PSM scores and the false PSM scores is increased. Specifically, by changing the parameters of the normal distribution from which the alternative scores are drawn we notice the discovery ratios are closer to 1 (Supplementary Figure 4 in the SI). We note that in terms of the overall accuracy of the FDR estimation methods such larger separation seems to mostly benefit STDS-PIT, whose liberal bias is reduced whereas the opposite holds for STDS (Supplementary Figure 2 in the SI).

The discussion so far overlooked an inherent feature of any decoy-based FDR estimation procedure, namely, the variability in the reported discoveries. In our simulations, this variability is due to both the target and decoy PSMs being drawn in each "experiment." As such, it is instructive to gauge this variability as an indication of what we might see in any given experiment with real data. Taking a closer look at the ratio of mix-max to T-TDC discoveries, we see that while the median value of this ratio stays more or less constant, the variability for the smaller sets of spectra is substantial (Figure 5). On the basis of the fact that for small sets of spectra (~1000 spectra) the subtle advantage that mix-max offers over T-TDC is much smaller than the observed variability, one might be tempted to conclude that there is no point to using mix-max in such a setting; however, one should keep in mind that (a) for such size sets T-TDC is too liberal (quite a bit more than mix-max: Figure 3) and (b) mix-max still offers the advantage of reduced variability in the list of discoveries, which is discussed later.

2.5. Analysis of Real Data

We next examined the behavior of all five FDR estimation procedures, C-TDC, T-TDC, STDS, STDS-PIT, and mix-max, using three real data sets, derived from a yeast whole cell lysate,[12] a *C. elegans* (worm) digest,[13] and an analysis of the erythrocytic cycle of the malaria parasite *Plasmodium falciparum*.[14] Searches were carried out using two different search engines, MS-GF+[15] and Tide,[16] and each tool was applied to each target database as well as to 1000 randomly drawn decoy databases. Each FDR estimation method was applied to each of these 1000 sets of paired optimal target and decoy PSM scores. More precisely, because STDS, STDS-PIT, and mix-max require calibrated scores (and T-TDC gains significantly from calibration), we calibrated both Tide's XCorr score and MS-GF+'s E value score using spectrum-specific empirical distributions constructed from 10K decoys before applying any FDR estimation procedure. (See Methods 4.)

In this real data setting, we no longer know which PSMs are false; however, we can still draw plots similar to the ones in Figure 4 and compare their general aspects with the ones that we drew based on the simulated data. Overall, the results obtained on real data sets agree with our simulation based analysis. In particular, STDS consistently yields fewer discoveries than mix-max, whereas STDS-PIT consistently yields more discoveries than mix-max (Figure 6, magenta and green curves, respectively). Furthermore, mix-max and T-TDC yield very similar numbers of discoveries, with mix-max

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

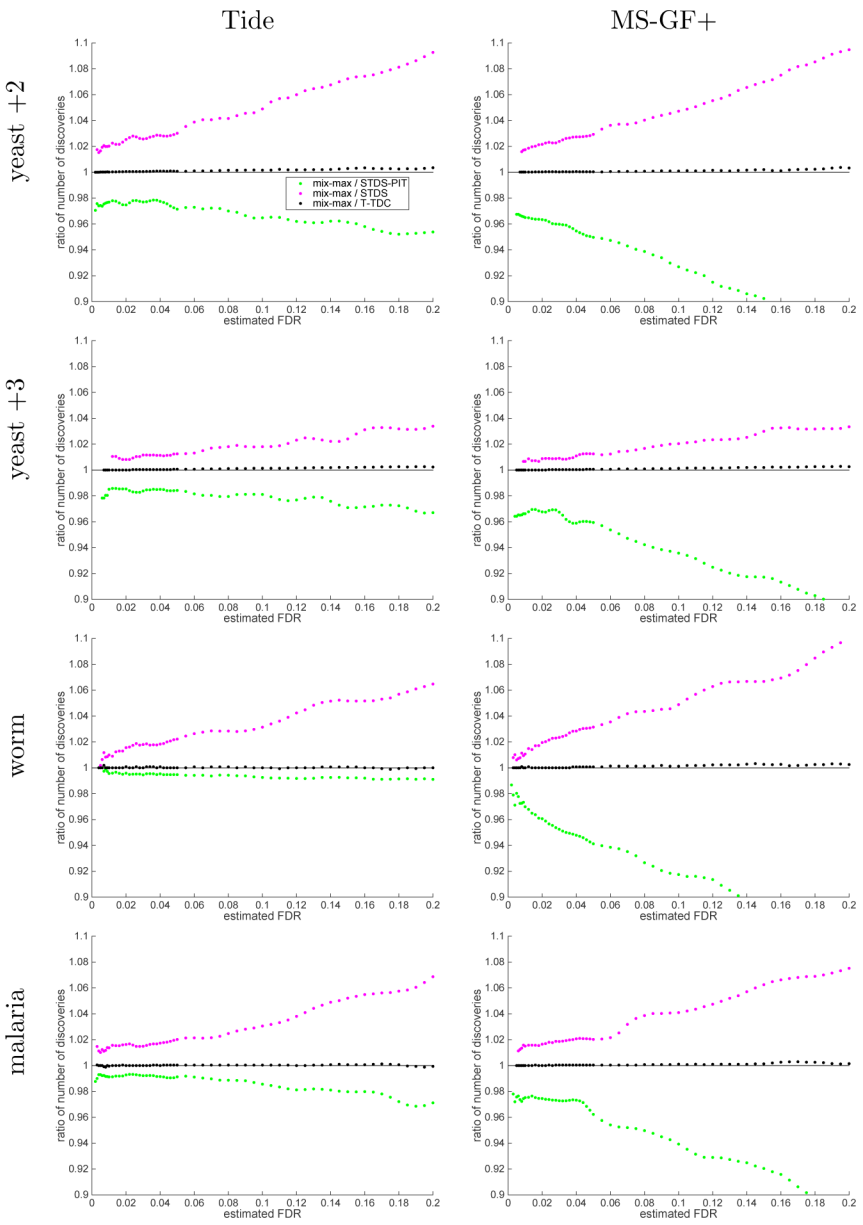


Figure 6. Median ratios of number of discoveries in the yeast data set. Each panel plots, as a function of estimated FDR, the median ratio of the number of mix-max to T-TDC/STDS/STDS-PIT discoveries. (For reference, the number of T-TDC discoveries is given in Supplementary Figure 6 in the SI.) The spectrum sets are the yeast, worm, and malaria data sets. The yeast data are separated by charge state, whereas the significantly smaller worm and malaria data sets are aggregates of both charge states. In each plot, the medians were taken with respect to 1000 corresponding discovery ratios, estimated using that many randomly drawn decoy databases. Each pair of target-decoy databases was searched using two different search engines: Tide and MS-GF+. In all cases, the scores were calibrated using spectrum-specific empirical distributions constructed from 10K randomly drawn decoy databases, as described in ref 6. Overall, the graphs are qualitatively similar to the results from the simulated data (Figure 4). The yeast +2 set has a lower estimated rate of foreign spectra than the corresponding +3 set (e.g., using Tide $\hat{\pi_0} = 0.66$ for charge 2 but $\hat{\pi_0} = 0.83$ for charge 3), which probably explains the larger differences between mix-max and both the STDS and STDS-PIT (cf. Supplementary Figure 5 in the SI).

showing a small but consistent gain in the larger spectrum sets of the yeast data (Figure 6, black curve, yeast rows). In the smaller malaria and worm data sets, except for very small FDRs, mix-max is almost tied with T-TDC, although it does have a miniscule overall advantage in terms of the number of discoveries (Figure 6, black curve, worm and malaria rows). Finally, consistent with our analysis that C-TDC is conservative when applied to the target-only discoveries (Supplementary Figure 1 in the SI), we find it is reporting significantly fewer discoveries than T-TDC does (Supplementary Figure 6 in the SI)

The overall smaller differences in the real compared with the simulated data between mix-max and the other three FDR estimation methods are somewhat explained when we examine results from simulations that set $\pi_0 = 0.7$, which seems more appropriate for this real data (Supplementary Figure 5 in the SI, top panels). Also, recall that a potentially larger separation between the native and null scores could also have an impact (Supplementary Figure 4 in the SI). Finally, the real data differs from the simulated one in that it uses our 10K calibration procedure. In Supplementary Note 4 in the SI we look at how incorporating this calibration procedure with the simulated data impacts the FDR estimation procedures. We find that

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

Table 2. Discrepancy in PSM Discoveries Reported by Different Applications of T-TDC and Mix-Max[a]

		% only in one T-TDC			% only in one mix-max		
set	FDR	0.01	0.05	0.10	0.01	0.05	0.10
yeast	0.05 quantile	0.0	0.1	0.3	0.0	0.0	0.0
	median	0.7	0.5	0.7	0.7	0.4	0.4
	0.95 quantile	3.1	1.7	2.1	2.9	1.5	1.7
worm	0.05 quantile	0.0	0.0	0.1	0.0	0.0	0.0
	median	1.2	0.9	1.3	1.2	0.7	0.9
	0.95 quantile	5.7	4.2	4.5	5.6	3.7	3.8
Plasmodium	0.05 quantile	0.0	0.0	0.1	0.0	0.0	0.0
	median	0.7	0.3	0.6	0.6	0.2	0.3
	0.95 quantile	3.5	1.9	2.3	3.3	1.6	1.8

[a]For each of 2000 pairs of applications of T-TDC/mix-max to analyze the Tide searches of the target database, coupled to two independently drawn decoys, we found the percentage of PSM discoveries (across the two largest charge sets of each species spectra sets) that were reported at the given FDR by only one of the two T-TDC/mix-max runs. The Table gives the quantiles of these percentages. The results show that mix-max consistently exhibits less decoy-dependent variability than T-TDC.

calibration has a fairly minor effect, making both T-TDC and mix-max slightly conservative for moderately large spectrum sets (≥10K) and small FDR value (<0.01) (Supplementary Figures 7 and 8 in the SI).

As noted in the Introduction, the use of randomly drawn decoy sets imparts an undesired variability on the reported list of discoveries. This is true of all five FDR estimation methods that we have considered; however, while in the case of mix-max, STDS, and STDS-PIT this decoy-induced variability manifests itself only in the selection of the cutoff, in the TDC methods the variability also causes some high-scoring target PSMs to be randomly eliminated. To quantify the magnitude of this phenomenon, we compared the decoy-induced variability of mix-max and T-TDC using both search engines applied to all three data sets. The results for Tide (Table 2) show that mix-max's list of discoveries exhibits a small but consistent reduced decoy-dependent variability compared with the list provided by T-TDC. (The same qualitative result holds for MS-GF+; data not shown.)

3. DISCUSSION

Introducing the distinction between foreign and native spectra has allowed us to formulate a theoretical model of decoy-based FDR estimation for spectrum identification. This model, in turn, suggests a variety of deficiencies of existing decoy-based FDR estimation procedures (summarized in Table 3). Perhaps most significantly, our analysis suggests that the STDS and STDS-PIT procedures, which were previously proposed by one of us (Noble), lead to conservative and liberal FDR estimates, respectively. To address this problem, we therefore proposed the mix-max procedure, which also employs separate target and decoy searches. Our theoretical analysis and empirical results suggest that the mix-max procedure estimates the FDR fairly accurately for spectrum sets that are larger than 1000−2000 spectra.

Table 3. Comparison of FDR Estimation Methods

	C-TDC	T-TDC	STDS	STDS-PIT	mix-max
estimates the FDR in an unbiased fashion	×	×			×
estimates the FDR in a conservative fashion			×		
estimates the FDR in a liberal fashion				×	
includes decoys PSMs in the list of discoveries	×				
excludes a random subset of high-scoring correct target PSMs	×	×			
requires calibrated scores			×	×	×

On the other hand, when the number of spectra is small, any of the decoy-based FDR estimation procedures we have considered face a challenge. For the mix-max procedure, a small set of spectra implies that the estimation of π_0, the proportion of foreign spectra, might be compromised. However, our analysis also suggests that T-TDC consistently underestimates the true FDR for small spectrum sets. To understand why this is the case, consider a spectrum set made of a single foreign spectrum. In this case, the true FDR is 1.0, but it is easy to see that T-TDC will grossly underestimate the true FDR with probability 0.5. We thus propose that small spectrum sets present a challenge that requires further research.

The mix-max procedure is only valid in the context of calibrated scores; however, given the advantages of calibration in general[6] we recommend calibrating whenever it is computationally feasible, in which case mix-max could be applied. If one is forced to use uncalibrated scores, then one should use the T-TDC method, which produces acceptable FDR estimation even with uncalibrated scores, albeit at a nontrivial loss of power.[6] Indeed, this feature is one of the strong suits of target-decoy competition, as pointed out by Elias and Gygi: the per spectrum competition with a decoy PSM provides a "built-in" level of calibration.

Finally, it is worth pointing out that both mix-max and T-TDC are only asymptotically accurate. Researchers would do well to remember that for any given set of target and decoy PSMs the estimated FDR is just that—an estimate which can deviate from the true FDR by a nontrivial amount. This deviation (Figure 7) is particularly large for small FDR thresholds, which are, ironically, typically of most interest to experimentalists. In practice, this inherent estimation error should be factored into any downstream analysis.

4. METHODS

4.1. Decoy-Based FDR Estimation Procedures

For a spectrum σ and a peptide $l \in DB$, where DB is a peptide database, let $S_{DB}(\sigma,l)$ be the score of the match between σ and l in the context of the database DB. Some scores can be invariant

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

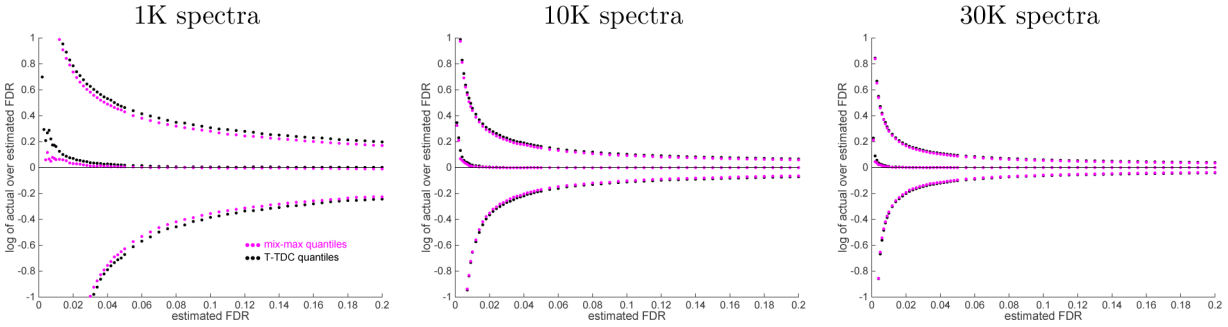


Figure 7. Accuracy of estimated FDR in simulated data. Each panel plots, as a function of estimated FDR, the 0.05/0.5/0.95 quantiles of the logarithm of the ratio between the actual FDR and the mix-max (magenta) or the T-TDC (black) estimated FDR. The data are the same as in Figure 3. As expected, the estimated FDR converges to the true value as the number of spectra increases; however, for sets of 1K spectra and FDR of 0.05, the mix-max or T-TDC estimated FDR can easily be ±50% off the true value, and even for sets of 30K spectra the estimated FDR can be roughly ±20% off.

with respect to the database *DB* (e.g., Xcorr), whereas others can depend on *DB*, for example, through the number of candidate peptides (e.g., *E* values). We assume that $S_{DB}(\sigma,I)$ is invariant of the spectrum set Σ from which σ came.

Given a target peptide database *db* and a set of observed spectra $\Sigma = \{\sigma_i: i = 1, ..., n_\Sigma\}$, let Σ_1 be the set of "native" spectra $\sigma \in \Sigma$, which were generated by a peptide $GP(\sigma) \in db$ and let $\Sigma_0 := \Sigma\backslash\Sigma_1$ be the "foreign" spectra. To simplify our notation, we assume each native spectrum is generated by at most one peptide, although this is not an inherent restriction of our model. We denote by $\pi_1 = n_1/n_\Sigma$ the proportion of native spectra among all spectra in our input data set Σ and by $\pi_0 = 1 - \pi_1$ the fraction of foreign spectra. (See Table 1 for a summary of our notation.)

For $\sigma_i \in \Sigma_1$ let $x_i = S_{db}(\sigma_i, GP(\sigma_i))$ be the score of the match between the spectrum σ_i and the peptide that generated it $GP(\sigma_i)$. For each foreign spectrum $\sigma_i \in \Sigma_0$ we define $y_i = S(\sigma_i, db) = \max_{I\in db} S_{db}(\sigma_i, I)$ as the score of the best match of this foreign spectrum to *db*. Similarly, for $\sigma_i \in \Sigma_1$ we define $y_i = \max_{I\in db\backslash GP(\sigma_i)} S_{db}(\sigma_i, I)$ as the score of the best match of σ_i to the "random" or irrelevant part of *db*.

In general, the scores x_i and y_i depend on some random effects that are not accounted for in our model of a random database, so we shall treat them alternatively as unknown parameters of the problem or as unobserved random variables X_i and Y_i. Unless otherwise stated we do not assume that the X_i and Y_i are identically distributed.

Finally, in the context of this theoretical model we assume that no ties are observed among the competing top scores x_i and y_i (and later z_i as well): $x_i \neq y_i$ 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.

A false positive or a false discovery at threshold *T* occurs when either $\sigma_i \in \Sigma_1$ and $y_i > \max(x_i, T)$, or $\sigma_i \in \Sigma_0$ and $y_i > T$.We summarize this as $y_i > \max(x_i, T)$ where $x_i := -\infty$ for $\sigma_i \in \Sigma_0$.

Note that while x_i and y_i are unobservable, $w_i := \max(x_i, y_i)$ is observable because it is the score of the optimal match to σ_i in the target database *db* (loosely referred to here as the "target PSM" or "target PSM score"). Hence, the number of "discoveries" $D := |\{i: w_i > T\}|$ is also observable but *F*, the number of false positives/discoveries, is not.

We are interested in gauging *F*/*D*, the FDR in our target list of discoveries. Elias and Gygi refer to a closely related figure (see later) as the "false-positive rate." We avoid this term here because it is also used for the *p* value of the match between a single spectrum against a random database. Below, we compare several methods of estimating *F*/*D*.

C-TDC. In their 2007 paper, Elias and Gygi suggest that instead of estimating the error rate *F*/*D* in the target database *db* we estimate a related error rate, F_C/D_C, where F_C and D_C are the analogues of *F* and *D* with respect to the concatenated database $db \oplus dc$. Here *dc* is the decoy database, which is of the same size as *db* and can be thought of as a particular sample from our null distribution of random databases.

To rigorously define the Elias and Gygi (C-TDC) procedure, we introduce the random variables $Z_i := S(\sigma_i, dc) = \max_{I\in dc} S_{dc}(\sigma_i, I)$, where $\sigma_i \in \Sigma$. A false positive in the concatenated search occurs if $\max(y_i, Z_i) > \max(x_i, T)$, and we define F_C as the (unknown) number of false positives we encounter in this search. The number of discoveries $D_C = |\{i: \max\{W_i, Z_i\} > T\}|$ is obviously observable; therefore, estimating F_C is essentially equivalent to estimating F_C/D_C. F_C is estimated by Elias and Gygi as

$$\widehat{F_C} = 2F_D \Rightarrow \frac{\widehat{F_C}}{D_C} = \frac{2F_D}{D_C} \quad (1)$$

where

$$F_D = D_D = |\{i: Z_i > \max(W_i, T)\}|$$

is the number of (false) discoveries that fall in *dc*.

Note that, in practice, the list of discoveries provided by the Elias and Gygi scheme would typically be reduced to those discoveries that fall within the target database. In other words, it makes sense to throw out all F_D decoy discoveries, leaving us with a filtered set of discoveries that has

$$D_T = |\{i: W_i > \max(Z_i, T)\}| = D_C - F_D \quad (2)$$

PSMs in it.

T-TDC. The list of discoveries reported by T-TDC is a subset of the list reported by C-TDC, consisting only of the PSMs from the target database. The T-TDC estimator of the FDR in this list is defined as

$$\text{T-TDC} := \frac{F_D}{D_T} \quad (3)$$

where F_D is the number of decoy discoveries in the concatenated search.

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

STDS and STDS-PIT. Käll et al. propose two methods for estimating the FDR.[5] In the first, which we call "separated target-decoy search" (STDS), they estimate the number of false discoveries that exceed score T as the number of decoy PSMs that exceed that score: $|\{i\colon z_i > T\}|$. This estimate is unbiased only if we assume there are no native spectra because the estimate gauges the number of false discoveries by counting the number of discoveries among spectra that are *entirely* foreign by definition. Hence, in general, STDS is overestimating the FDR, which it estimates as

$$\frac{\widehat{F}}{D} = \frac{|\{i:\ z_i > T\}|}{|\{i:\ w_i > T\}|} \tag{4}$$

Arguing along similar lines, Käll et al. also noted that STDS should, in general, be conservative; hence, they proposed to correct the method by estimating what they termed the "percentage of incorrect target" or PIT, which is the (unknown) number of all false target identifications divided by the number of spectra. Note that this is not the same as the fraction of foreign spectra: some of the native spectra PSMs are typically also false.

Specifically, implicitly assuming a calibrated score, Käll et al. use the decoy PSMs $\{z_i\}$ to estimate the p values of $\{w_j\}$, the set of optimal target PSMs scores. Subjecting these estimated p values to the standard FDR analysis of Storey,[17] they estimate their PIT by identifying it with π_0 the proportion of true null hypotheses among the set of target PSMs; however, the p values of these target PSMs are computed based on the empirical CDF compiled purely from decoy scores; hence, these p values are computed relative to the null hypothesis that the spectrum is foreign. Thus, $\hat{\pi_0}$ estimates the fraction of foreign spectra rather than the PIT as claimed. For more on this point, see Supplementary Note 2 in the SI.

Regardless, Käll et al. then use this estimate to adjust the STDS estimate with the following FDR estimate, which we refer to as STDS-PIT

$$\frac{\widehat{F}}{D} = \widehat{\pi}_0 \frac{|\{i:\ z_i > T\}|}{|\{i:\ w_i > T\}|} \tag{5}$$

In this work we estimated π_0 using the R qvalue package (http://www.bioconductor.org/packages/release/bioc/html/qvalue.html with the option "smoother" which is equivalent to the way it is estimated in STDS-PIT.

Note that the difference between STDS (4) and STDS-PIT (5) is similar to the difference between the Benjamini–Hochberg's procedure of controlling the FDR and that of Storey.[17]

Note that the previously described version of STDS-PIT is the original one presented in ref 5. Because of difficulties in estimating π_0 using the standard method (which in hindsight were mostly due to the use of uncalibrated scores[6]), in a followup work Käll et al. proposed estimating π_0 using Storey's bootstrap method;[10,11] however, while, in principle, this is a theoretically sound approach and indeed the preferred one for small sets, Käll et al. introduced an undocumented twist to the estimation that is not theoretically justified. Specifically, the maximal value of λ was set to 0.5 even though Storey used 0.95 when he originally proposed this bootstrap estimation method,[17] as well as in a closely related variant[18] and the qvalue package itself uses a default of 0.90 for the "bootstrap" option. Setting the maximal value of λ to 0.5 tends to inflate the estimation of the proportion of true null hypotheses in the sample, thereby reducing the liberal bias that is inherent to STDS-PIT. Thus, the revised method is, in general, less biased than the one presented here, although this modification is not theoretically supported. At any rate, it can be shown that even if the PIT could have been correctly estimated, STDS-PIT would still show a tendency to be liberally biased.

Mix-Max. If you believe that your PSM score is well calibrated, then a better scoring PSM should imply a better match. Therefore, for any desired level of FDR you would prefer taking all of the target PSMs scoring above the appropriately computed threshold. The T-TDC procedure, however, does not allow one to do this because it presents us with a filtered list of discoveries. The method that we propose next is therefore preferable for well-calibrated scores because, like the methods of Käll et al., it is designed to estimate the FDR in the entire set of target discoveries rather than a filtered one. Let

$$F_0 = |\{i:\ \sigma_i \in \Sigma_0\,,\ y_i > T\}|$$

$$F_1 = |\{i:\ \sigma_i \in \Sigma_1\,,\ y_i > \max(x_i, T)\}|$$

be the number of foreign, respectively, native spectra generating a false positive in the target database (when used on its own). Clearly, $F = F_0 + F_1$ and our mixture-maximum (mix-max) procedure estimates each term separately.

It is relatively easy to estimate F_0 if we adopt the same homogeneous null assumption of Käll et al. (Z_i and Y_i are independent of X_i and share the same null distribution for all i). Indeed, by applying their method to estimate the proportion of true null hypothesis π_0, which is in fact the fraction of foreign spectra, we can estimate

$$\widehat{F_0} = \widehat{\pi}_0 \cdot n_\Sigma \cdot P(\widehat{Z_i > T}) = \widehat{\pi}_0 \cdot \sum_{j=1}^{n_\Sigma} 1_{z_j > T} \tag{6}$$

The following claim motivates our estimator of $E(F_1)$:

Claim 1. Let G_i be the cumulative distribution function (CDF) of Y_i (or Z_i), which under the homogeneous null assumption is independent of X_i. Then

$$E(F_1) = \sum_{\sigma_i \in \Sigma_1} \int_{y>T} \frac{P(W_i < y)}{P(Y_i < y)}\, \mathrm{d}G_i(y)$$

Proof. Because we assume that X_i and Y_i are independent we have

$$P(Y_i > \max(X_i, T)) = \int_{y>T} P(y > X_i | Y_i = y)\, \mathrm{d}G_i(y)$$

$$= \int_{y>T} P(y > X_i)\, \mathrm{d}G_i(y)$$

and since $W_i = \max(X_i, Y_i)$

$$P(W_i < y) = P(X_i < y) \cdot P(Y_i < y)$$

Therefore

$$P(Y_i > \max(X_i, T)) = \int_{y>T} \frac{P(W_i < y)}{P(Y_i < y)}\, \mathrm{d}G_i(y)$$

The claim now follows immediately from

$$E(F_1) = \sum_{\sigma_i \in \Sigma_1} P(Y_i > \max(X_i, T))$$

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148–3161

If in addition to the homogeneous null assumption (Z_i and Y_i are independent of X_i and share the same null distribution for all i), we make the reasonable assumption that the random variables $\{X_i: \sigma_i \in \Sigma_1\}$ are identically distributed, then so are the random variables W_i, and it follows from the previous claim that

$$E(F_1) = n_1 \cdot \int_{y>T} \frac{P(W_i < y)}{P(Y_i < y)}\, dG(y) \tag{7}$$

where $n_1 = |\Sigma_1|$, G is the CDF of any Y_i, and W refers to any W_i with $\sigma_i \in \Sigma_1$.

The discussion so far assumed there are no ties; however, in reality, especially if a procedure such as the nonparametric calibration described in ref 6 is applied, ties will be present. In the presence of ties, eq7 slightly underestimates the false discoveries due to the native spectra because it implicitly resolves all ties in favor of the correct identification (X_i). Conversely, replacing the strict inequalities of eq 7 with weak ones as in eq 8 below implicitly resolves ties in favor of the incorrect identification (Y_i) and hence slightly overestimates the false discoveries due to the native spectra. Following the standard statistical practice we therefore prefer to use the right-hand side of eq 8 below rather than risk being too liberal:

$$E(F_1) \geq n_1 \cdot \int_{y>T} \frac{P(W_i \leq y)}{P(Y_i \leq y)}\, dG(y) \tag{8}$$

We estimate the right-hand side of eq 8 using the decoy PSM scores z_j, as explained next. Because $Z \sim Y$ we can estimate

$$\widehat{P(Y \leq z_j)} := \frac{\sum_k 1_{z_k \leq z_j}}{n_\Sigma}$$

For W_i with $\sigma_i \in \Sigma_1$ we can estimate $P(W_i \leq y)$ by accounting for the number of events $\{W_k \leq y\}$ that are due to foreign spectra. Specifically, the latter number can be estimated from the decoy set because the distribution of W_k for $\sigma_k \in \Sigma_0$ is identical to the distribution of Z_k

$$\widehat{P(W_i \leq y)} := \frac{\sum_k 1_{w_k \leq y} - \hat{\pi}_0 \cdot \sum_k 1_{z_k \leq y}}{(1 - \hat{\pi}_0) n_\Sigma}$$

and with $y = z_j$ we have

$$\widehat{P(W_i \leq z_j)} = \frac{\sum_k 1_{w_k \leq z_j} - \hat{\pi}_0 \cdot \sum_k 1_{z_k \leq z_j}}{(1 - \hat{\pi}_0) n_\Sigma}$$

Therefore, we estimate

$$\widehat{E(F_1)} := (1 - \hat{\pi}_0) \cdot n_\Sigma \cdot \sum_{z_j > T} \left[\frac{\sum_k 1_{w_k \leq z_j} - \hat{\pi}_0 \cdot \sum_k 1_{z_k \leq z_j}}{(1 - \hat{\pi}_0) n_\Sigma} \Big/ \frac{\sum_k 1_{z_k \leq z_j}}{n_\Sigma} \right]_{[0,1]} \cdot \frac{1}{n_\Sigma}$$

where $[x]_{[0,1]} := \max\{0, \min\{1,x\}\}$ ensures that x remains an acceptable probability value. Thus, our mix-max estimator is defined as

$$\frac{\widehat{F}}{D} := \frac{\hat{\pi}_0 \cdot \sum_{j=1}^{n_\Sigma} 1_{z_j > T} + (1 - \hat{\pi}_0) \cdot \sum_{z_j > T} \left[\frac{\sum_k 1_{w_k \leq z_j}}{(1 - \hat{\pi}_0) \cdot \sum_k 1_{z_k \leq z_j}} - \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right]_{[0,1]}}{\sum_i 1_{w_i > T}} \tag{9}$$

It is important to stress that the mix-max method explicitly requires the homogeneous null assumption. (Z_i and Y_i are independent of X_i and share the same null distribution for all i.) Although this is not a realistic assumption when using raw scores,[19] using a calibrated score as suggested here would essentially satisfy this requirement.

In Supplementary Note 3 in the SI we provide pseudocode showing how mix-max can be adapted for simultaneously computing the FDR for all possible thresholds of interest.

4.2. Simulated Data

We performed simulations where we drew n_Σ independent triplets of independent random variables X_i, Y_i, Z_i with $Y_i, Z_i \approx N(0,1)$ (corresponding to the use of calibrated scores). With $(1 - \pi_0)$ being the fraction of native spectra, $(1 - \pi_0) \cdot n_\Sigma$ of the X_i were drawn from the $N(2.5,1)$ distribution, whereas the other $\pi_0 \cdot n_\Sigma$ of the X_i, corresponding to foreign spectra, were defined as $-\infty$. We then defined $W_i = \max\{X_i, Y_i\}$ as the "target" PSM score of the ith "spectrum" and Z_i as the corresponding "decoy" PSM score. Note that for the foreign spectra, by our choice of X_i, $W_i = Y_i$.

We then applied the five FDR estimation methods to the pairs of target-decoy PSMs, noting the number of discoveries at a selected set of FDR values as explained in Section 4.5.

Because the data are simulated, we know which target discovery is correct ($W_i = X_i$) and which one is incorrect or false ($W_i = Y_i$). Note that because the data are sampled from a continuous distribution there are no ties. We therefore kept a record of the number of false as well as true discoveries at each of the selected FDR values. Using these data we could readily calculate the total number of discoveries as well as the actual FDR at every FDR threshold.

We made one set of experiments where we drew 10K "datasets" with $n_\Sigma = 10^4$ as well as with $n_\Sigma = 10^5$ to corroborate our claim that $\hat{\pi}_0$ estimates the proportion of foreign spectra rather than the incorrect PSMs. Our main set of experiments was repeated 10K times each so that we could gauge the variability in the results. Specifically, the spectrum set sizes we used were 500, 1K, 10K, 30K, and 70K, and for each size we drew 10K pairs of target and decoy PSMs as previously described.

To analyze the effect our nonparametric 10K-decoys calibration procedure has on FDR estimation, we also generated data where we added an artificial calibration step. Because the data were perfectly calibrated to begin with, this step is not only redundant but also in fact compromises the perfect calibration the data enjoyed initially. Specifically, for each "spectrum" (really just an index i) we drew 10K "null PSM scores" according to $N(0,1)$ and used this 10K sample to construct an empirical distribution function (ECDF) that was specific to this spectrum. Next, every drawn score, X_i, Y_i, Z_i, was transformed to a p value using the corresponding ECDF (technically, the score is minus the p value). Note that the ECDF was drawn only once, so all randomly drawn pairs of n_Σ target-decoy sets of optimal PSMs were transformed using a fixed set of n_Σ ECDFs.

We also conducted another simulation where we increased the average separation between the native scores X_i and the null drawn ones Y_i, Z_i. Specifically, while drawing the null scores from the same $N(0,1)$ distribution, we increased the mean of the "correct PSM" scores from $\mu = 2.5$ to 3.0.

4.3. Real Data

Analysis was performed using three previously described sets of spectra (Table 4). All three data sets and their associated protein databases are available at http://noble.gs.washington.edu/proj/calibration.

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

Table 4. Properties of the Three Data Sets

data set	yeast	worm	*Plasmodium*
precursor resolution	low	high	high
fragment resolution	low	low	high
+1 spectra	737	1423	
+2 spectra	34,499	7891	1311
+3 spectra	34,469	4646	8441
+4 spectra		1683	2372
+1 PSMs	737	241	
+2 PSMs	34,499	4494	790
+3 PSMs	34,467	3173	8382
+4 PSMs		1288	2362
enzyme	trypsin	trypsin	lys-c
peptides in database	165,930	462,523	223,602
precursor m/z tolerance	±3 Th	±10 ppm	±50 ppm
fragment m/z bin width	1.0005079	1.0005079	0.10005079
average candidates/spectrum	955.7	22.0	48.7

The yeast data set was collected from *S. cerevisiae* (strain S288C) whole cell lysate.[12] The cells were cultured in YPD media and grown to mid log phase at 30 °C, then lysed and solubilized in 0.1% RapiGest. Digestion was performed with a modified trypsin (Promega), and the sample was subsequently microcentrifuged at 14 000 rpm to remove any insoluble material. Microcapillary liquid chromatography tandem mass spectrometry was performed using 60 cm of fused silica capillary tubing (75 μm I.D.; Polymicro Technologies), placed in-line with an Agilent 1100 HPLC system and an LTQ ion trap mass spectrometer. MS/MS spectra were acquired using data-dependent acquisition with a single MS survey scan triggering five MS/MS scans. Precursor ions were isolated using a 2 m/z isolation window. The charge state of each spectrum was estimated by a simple heuristic that distinguishes between singly charged and multiply charged peptides using the fraction of the measured signal above and below the precursor m/z.[20] No attempt to distinguish between 2+ or 3+ spectra were made other than limiting the database search to peptides with a calculated M+H mass of 700 to 4000 Da. Thus, of the 35 236 spectra, 737 were searched at 1+ charge state, 30 were searched at 2+ charge state, and the remaining (34 469) were searched at both 2+ and 3+ charge states.

The worm data set is derived from a *C. elegans* digest.[13] *C. elegans* were grown to various developmental stages on peptone plates containing *E. coli*. After removal from the plate, bacterial contamination was removed by sucrose floating. The lysate was sonicated and digested with trypsin. The digest (4 μg) was loaded from the autosampler onto a fused-silica capillary column (75 μm i.d.) packed with 40 cm of Jupiter C12 material (Phenomenex) mounted in an in-house constructed microspray source and placed in line with a Waters NanoAcquity HPLC and autosampler. The column length and HPLC were chosen specifically to provide highly reproducible chromatography between technical replicates, as previously described.[13] Tandem mass spectra were acquired using data-dependent acquisition with dynamic exclusion turned on. Each high-resolution precursor mass spectrum was acquired at 60 000 resolution (at m/z 400) in the Orbitrap mass analyzer in parallel with five low-resolution MS/MS spectra acquired in the LTQ. Bullseye[21] was then used to assign charges and high-resolution precursor masses to each observed spectrum on the basis of persistent peptide isotope distributions. Because a single precursor m/z range may contain multiple such distributions, Bullseye frequently assigns more than one distinct precursor charge and mass to a given fragmentation spectrum. The final data set consists of 7557 fragmentation spectra, with an average of 2.10 distinct precursors per spectrum: 1423 +1, 7891 +2, 4646 +3, 1683 +4, and 228 +5. The +5 spectra were discarded from the analysis.

The *Plasmodium* data set is derived from a recent study of the erythrocytic cycle of the malaria parasite *Plasmodium falciparum*.[14] *P. falciparum* 3D7 parasites were synchronized and harvested in duplicate at three different time points during the erythrocytic cycle: ring (16 ± 4 h postinvasion), trophozoite (26 ± 4 h postinvasion), and schizont (36 ± 4 h postinvasion). Parasites were lysed and duplicate samples were reduced, alkylated, digested with Lys-C, and then labeled with one of six TMT isobaric labeling reagents. The resulting peptides were mixed together, then fractionated via strong cation exchange into 20 fractions, desalted, and then analyzed via LC−MS/MS on an LTQ-Velos-Orbitrap mass spectrometer. All MS/MS spectra were acquired at high resolution in the Orbitrap. We focused on one of these fractions (number 10), consisting of 12 594 spectra, and we discarded 470 spectra with charge state > +4, leaving 12 124 spectra.

The analysis of the real data was limited to charge sets that contained at least 2000 spectra, which, in practice, included exactly the two larger charge sets in each data set: charges 2 and 3 of the yeast data, charges 2 and 3 of the worm data, and charges 3 and 4 of the malaria data. Results for other charge states were not included in the analysis.

4.4. Assigning Peptides to Spectra

Searches were carried out using two different search engines: the Tide search engine[16] as implemented in Crux v2.0[9] and MS-GF+.[7]

The yeast spectra were searched against a fully tryptic database of yeast proteins. The trypsin cleavage rule did not include suppression of cleavages via proline.[22] The precursor m/z window was ±3.0 Th. No missed cleavages were allowed, and monoisotopic masses were employed for both precursor and fragment masses. A static modification of C+57.02146 was included to account for carbamidomethylation of cysteine. For Tide, the mz-bin-width parameter was left at its default value of 1.0005079, and PSMs were ranked by the XCorr score. For MS-GF+, the -inst parameter was set to 0, and no isotope errors were allowed.

The worm spectra were searched against a fully tryptic database of *C. elegans* proteins plus common contaminants. Searches were performed using the same parameters as for the yeast data set, except that candidate peptides were selected using a precursor tolerance of 10 ppm. For MS-GF+, the -inst parameter was set to 1, and no isotope errors were allowed. Note that because of the Bullseye processing of the worm spectra a single spectrum may have be assigned multiple high-resolution precursor windows with the same charge state. In such cases, we identified the maximum scoring PSM per charge state. Eliminating spectra with no Bullseye-assigned precursor window or no candidate peptides within the assigned precursor range yielded a total of 9312 worm PSMs.

The *Plasmodium* spectra were searched against a database of *Plasmodium* peptides, digested using Lys-C. In addition to C +57.02146, static modifications of +229.16293 were applied to lysine and to the peptide N-termini to account for TMT labeling. All searches were performed using a 50 ppm precursor range. For Tide, the mz-bin-width parameter was set to

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

0.10005079. For MS-GF+, the -inst parameter was set to 1, and no isotope errors were allowed. For some spectra, no candidate peptides occur within the specified precursor tolerance window; hence, the number of PSMs (11 625) is smaller than the total number of spectra (12 594).

Decoy databases were generated by independently shuffling the nonterminal amino acids of each distinct target peptide. For each database, the decoy creation procedure was repeated 11 000 times, creating a 10K decoy set (used for calibration as explained next) and an independent 1K decoy set (used for evaluating the performance of the search methods).

Calibrating the Real Data Scores. We performed empirical score calibration using our previously described procedure.[6] For each spectrum σ and the target database db as well as each decoy database dc in the 1K decoy set we replaced its reported optimal PSM score (XCorr or E value) $S(\sigma,db)$ (or $S(\sigma,dc)$) with its calibrated score computed using the 10K decoys. To calibrate the score, we searched each spectrum σ against each decoy database dc in the 10K decoy set, and the optimal PSM for that database $z = S(\sigma,dc)$ was noted. (We loosely refer to it as the decoy PSM.) We next used this null sample of $N = 10\,000$ decoy PSM scores $\{z_i\}_{i=1}^{N}$ to construct a spectrum-specific empirical null distribution. This distribution was used to assign the per-spectrum p value to any PSM $s = S(\sigma,DB)$ involving the considered spectrum σ and a database DB. The calibrated score is the negative of the value. For each data set, the spectra were divided by charge state, and the previously described procedure was carried out separately on each of the resulting sets of spectra.

4.5. Number of Discoveries

Number of Discoveries at a Given FDR. The number of discoveries at a given FDR level α was defined as the largest number of discoveries reported by the method for which the estimated FDR was still $\leq\alpha$. For computational efficiency we only computed this number for 120 selected values of α: from 0.001 to 0.01 in increments of 0.001, from 0.012 to 0.05 in increments of 0.002, and from 0.055 to 0.5 in increments of 0.005.

Evaluating the Differences in Discovery Lists. The differences in the discovery lists between methods A and B were evaluated at a given FDR level $\alpha \in \{0.01,0.05,0.1,0.2\}$ as follows. First, we identified the largest discovery list reported by each method for which the FDR was still $\leq\alpha$. Then, the two lists were compared to see which PSMs appear only in A and not in B and vice versa. Finally, the number of PSMs present only in A's list was expressed as a percentage of the total number of PSMs in this list (and vice versa).

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Note 1: The C-TDC and T-TDC procedures are unbiased estimators of FDR. Supplementary Note 2: STDS-PIT underestimates the true FDR. Supplementary Note 3: suggested implementation of mix-max. Supplementary Note 4: the effects of our 10K-calibration procedure on FDR estimation. Supplementary Figure 1: Accuracy of C-TDC estimated FDR (mixture model). Supplementary Figure 2: Accuracy of estimated FDR (mixture model, larger separation). Supplementary Figure 3: Accuracy of STDS-PIT+qvality. Supplementary Figure 4: Median ratios of number of discoveries, larger separation. Supplementary Figure 5: Median ratios of number of discoveries. Supplementary Figure 6: Median number of T-TDC and C-TDC discoveries in the yeast data set. Supplementary Figure 7: Accuracy of estimated FDR (mixture model, 10K calibrated scores). Supplementary Figure 8: Median ratios of number of discoveries, 10K calibrated scores. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00081.

■ AUTHOR INFORMATION

Corresponding Authors

*U.K.: Phone: 61 2 9351 2307. E-mail: uri@maths.usyd.edu.au.
*W.S.N.: Phone: 1 206 221-4973. Fax: 1 206 685-7301. E-mail: william-noble@uw.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank J. Jeffry Howbert for thoughtful comments on the draft manuscript. U.K. thanks Pavel Pevzner for suggesting this general direction of research and Nuno Bandeira for helpful discussions.

■ ABBREVIATIONS

FDR, false discovery rate; PSM, peptide-spectrum match; TDC, target-decoy competition; C-TDC, original Elias and Gygi combined target-decoy competition; T-TDC, target-only target-decoy competition; STDS, separated target-decoy search; PIT, percentage of incorrect targets

■ REFERENCES

(1) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092−2123.

(2) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.

(3) Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinf.* **2012**, *13* (Suppl. 16), S2.

(4) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604* (55−71), 55−71, DOI: 10.1007/978-1-60761-444-9_5.

(5) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29−34.

(6) Keich, U.; Noble, W. S. On the importance of well calibrated scores for identifying shotgun proteomics spectra. *J. Proteome Res.* **2015**, *14* (2), 1147−1160.

(7) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7*, 3354−3363.

(8) Howbert, J. J.; Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **2014**, *13* (9), 2467−2479.

(9) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; MacCoss, M. J.; Noble, W. S. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **2014**, *13* (10), 4488−4491.

(10) Käll, L.; Storey, J.; Noble, W. S. Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008**, *24* (16), i42−i48.

(11) Käll, L.; Storey, J. D.; Noble, W. S. QVALITY: Nonparametric estimation of q values and posterior error probabilities. *Bioinformatics* **2009**, *25* (7), 964−966.

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161

(12) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−25.
(13) Hoopmann, M. R.; Merrihew, G. E.; von Haller, P. D.; MacCoss, M. J. Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. *J. Proteome Res.* **2009**, *8* (4), 1870−1875.
(14) Pease, B. N.; Huttlin, E. L.; Jedrychowski, M. P.; Talevich, E.; Harmon, J.; Dillman, T.; Kannan, N.; Doerig, C.; Chakrabarti, R.; Gygi, S. P.; Chakrabarti, D. Global analysis of protein expression and phosphorylation of three stages of *Plasmodium falciparum* intraerythrocytic development. *J. Proteome Res.* **2013**, *12*, 4028−4045.
(15) Kim, S.; Pevzner, P. A. MS-GF+ makes progress toward a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(16) Diament, B.; Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **2011**, *10* (9), 3871−3879.
(17) Storey, J. D. A direct approach to false discovery rates. *J. R Stat Soc. Series B* **2002**, *64*, 479−498.
(18) 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* **2004**, *66*, 187−205.
(19) Granholm, V.; Käll, L. Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* **2011**, *11* (6), 1086−1093.
(20) Klammer, A. A.; Wu, C. C.; MacCoss, M. J.; Noble, W. S. Peptide charge state determination for low-resolution tandem mass spectra. *Proc. IEEE Comput. Syst. Bioinform. Conf.* **2005**, 175−185.
(21) Hsieh, E.; Hoopmann, M.; Maclean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138−1143.
(22) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7* (1), 300−305.

DOI: 10.1021/acs.jproteome.5b00081
J. Proteome Res. 2015, 14, 3148−3161