Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics

Yohann Couté[♯], Christophe Bruley[♯], Thomas Burger*
Univ. Grenoble Alpes, CNRS, CEA, INSERM, IRIG, BGE
F-38000 Grenoble, France
[♯]Equal contribution
*thomas.burger@cea.fr

December 16, 2019

Abstract

In bottom-up discovery proteomics, target-decoy competition (TDC) is the most popular method for false discovery rate (FDR) control. Despite unquestionable statistical foundations, this method has drawbacks, including its hitherto unknown intrinsic lack of stability *vis-à-vis* practical conditions of application. Although some consequences of this instability have already been empirically described, they may have been misinterpreted. This article provides evidence that TDC has become less reliable as the accuracy of modern mass spectrometers improved. We therefore propose to replace TDC by a totally different method to control the FDR at spectrum, peptide and protein levels, while benefiting from the theoretical guarantees of the Benjamini-Hochberg framework. As this method is simpler to use, faster to compute and more stable than TDC, we argue that it is better adapted to the standardization and throughput constraints of current proteomic platforms.

1 Introduction

Most projects involving mass spectrometry (MS)-based discovery proteomics use data-dependent acquisition workflows in which tandem mass (MS/MS) spectra are produced from isolated peptides. Then, peptide identification is performed by database search engines which match the experimental spectra acquired with theoretical spectra derived from a list of protein sequences. This methodology has been widely adopted, and it has been accepted that it could lead to false positive identifications. Indeed, among the tremendous number of spectra generated by a peptide mixture prepared from a complex biological sample at least a few of them are expected to match an erroneous sequence, by chance. To avoid corrupting the biological conclusions of the analysis, researchers have come to rely on statistical procedures to limit the False Discovery Proportion (FDP) – *i.e.* the proportion of random mismatches among all the peptide spectrum matches (PSMs) which look correct. As this quality control problem is ubiquitous in science, statisticians have extensively studied it. The main conclusions of these studies (See [1] for a proteomic-oriented summary) are as follows: *(i)* Due to the random nature of the mismatches, it is impossible to precisely compute the FDP; *(ii)* However, it can be estimated, as an FDR (False Discovery Rate); *(iii)* Depending on the experiment, the FDR will provide a more or less accurate estimate of the FDP; *(iv)* Therefore, practitioners should carefully select the FDR methodology, and interpret its result cautiously, making an educated guess (*e.g.*, like a political poll before an election).

Target-decoy competition (TDC) has emerged as the most popular method to estimate the FDP in MS-based discovery proteomics [2]. Its success is a marker both of its conceptual simplicity and of its broad scope of application. The principle of TDC is to create artificial mismatches by searching a specific ("decoy") database of random sequences which differ from the sequences of interest (or "target" sequences) and to organize a competition between target and decoy assignments. Under the so-called Equal Chance Assumption [2] (stating that target mismatches and decoy matches are equally likely), it is possible, for any given cut-off score, to estimate the number of target mismatches that will be validated. Like any other estimator, TDC-FDR can lead to inconsistent estimates if the theoretical assumptions on which it is based do not hold in practice. Notably, the quality of TDC-FDR is strictly linked to the validity of

the Equal Chance Assumption, *i.e.* the decoy's capacity to adequately fool the database search engine. If it fools it too much, the TDC-FDR will overestimate the FDP; whereas if it is too unrealistic to fool the search engine, the FDP will be underestimated [3]. For this reason, decoy database construction and conditions of application have been extensively studied. Results from these studies indicate that: *(i)* the search engine must be compliant with TDC [4]; *(ii)* In theory, the larger the decoy database, the more precise the mismatch score distribution [5, 6] (like political polls, accuracy depends on the number of citizens surveyed); *(iii)* The decoys must respect the cleavage sites [7] to avoid systematic target matching regardless of spectrum quality; *(iv)* The influence of randomness in the construction of the decoy database can be counter-balanced by boosting strategies, leading to less volatile FDRs [8]; *(v)* Decoy counting also has an influence [6]. In addition to these restrictions, numerous parameters have been reported and discussed to control their relative importance [9]. This extensive body of literature has notably contributed to installing the competition step of TDC as essential, and today, target-decoy searches without competition [10] are scarcely ever reported. Despite TDC wide acceptance, a series of letters from Bret Cooper [11, 12] initiated a controversy regarding the observed downfall of TDC validation levels with higher resolution instruments. He provided experimental arguments to reject the idea that such downfall was simply a positive consequence of instrument evolution, leading to an increase in the numbers of peptides identified. Notably, he pointed out that very low-quality spectra incompatible with confident peptide identifications could be validated despite application of a stringent FDR cut-off. Moreover, as this phenomenon was observed with multiple widely-used search engines (Mascot, X!tandem and MS-GF+), he concluded that there was an "*inherent bias*" of "*peptide presumption*" (i.e., only peptides already listed in a target database could be identified). As this stance contradicted both empirical and theoretical evidence, few articles were published arguing against this view [13, 14] while others confirmed [15, 16], maintaining the *statu quo*.

However, Cooper's observations can be reconciled with statistical theory. In fact, the correctness of any statistical estimate is only asymptotic: if the quality of the empirical model depicting the mismatches is improved (for instance, by increasing the size of the decoy database [5, 6] or by averaging a growing number of TDC-FDRs resulting from randomly generated decoy databases, in a boosting-like strategy [8]), we should end-up with a series of estimates that theoretically converges towards the FDP. Although essential, this asymptotic property is unfortunately not sufficient for practitioners, who work with a finite number of decoy databases of finite size (classically, a single decoy database of the same size as the target database). Thus, the convergence speed and stability of the TDC estimator must be verified: If the convergence is very slow or if the TDC provides volatile estimates (when two slightly different conditions provide estimates of very different quality), it is possible, in given application conditions, to obtain inaccurate FDRs in practice.

In this article, we first present results that seriously question how MS/MS identification results for discovery proteomics assays are generally validated using TDC. We notably shed new light on Cooper's observations, which reconcile opposing opinions: While we believe target and decoy searches can be used to accurately compute FDRs, we uphold his concerns by showing that, with state-of-the-art high-resolution instruments, the risk that the TDC strongly underestimates the FDP increases. We then describe a series of mathematical transformations of classical identification scores, to which the well-known Benjamini-Hochberg (BH) procedure [17] and its numerous variants [18] can be applied at spectrum, peptide and protein levels. This leads to an original and powerful framework that demonstrably controls the FDR without requiring construction and searching of decoy databases. Altogether, the results presented demonstrate that making TDC-FDR compliant with instrument improvements requires unexpected efforts (careful implementation, fine-tuning, manual checks and computational time), whereas proteomics results can be simply and accurately validated by applying alternative strategies.

2 Results

2.1 TDC accuracy depends on mass tolerances set during database searching

The mechanism underlying the mass tolerance problem raised in [12] is as follows: Depending on the instrument's accuracy when measuring the precursor mass, due to variable resolving powers, and assuming the search engine is tuned accordingly, a larger or smaller number of decoys are considered possible competitors for a given spectrum. Thus, MS data acquired with high-resolution mass spectrometers and analyzed using search engines in which narrow mass tolerances are set *de facto* leads to smaller numbers of decoy matches, and therefore, to lower FDR thresholds. Cooper highlighted this mechanism, and concluded that "*there may be no remedy*". In contrast, it is well-established that from a theoretical

viewpoint, TDC provide a good estimate of the FDP [19, 5, 20].

To the best of our knowledge, TDC stability with respect to preliminary filters applied to reduce the number of decoy competitors in a non-uniform manner, has never been demonstrated. We would therefore like to defend an intermediary viewpoint: If used appropriately, decoy searches could lead to accurate estimates, even with high-resolution instruments; However, the lack of demonstrated stability of TDC is consistent with potential inaccurate FDRs, as reducing the mass tolerance during database searching would increase the chance that the Equal Chance Assumption [2] no longer applies.

To illustrate TDC instability, we compared the behavior of the TDC estimate with another estimate for which stability has been established, the Benjamini-Hochberg (BH) FDR [17]. This comparison has already been performed (*e.g.* [21, 22]), but the discrepancies observed were not studied in terms of estimate stability. To compute a BH-FDR, the search engine must provide PSM scores which can be related to p-values. Fortunately, numerous state-of-the-art search engines do so [23]: For instance, Mascot provides scores in the form

$$S = -10 \cdot \log_{10}(p) \quad \text{or} \quad p = 10^{-\frac{S}{10}} \tag{1}$$

where p is a p-value. Andromeda provides a similar calculation, although the score is not directly accessible as it is only an intermediate computation (see Methods 4.3). PepProbe, InsPecT and MyriMatch directly provide p-values as scores, and SEQUEST scores can be transformed into p-values through the application of dedicated wrappers *e.g.* [21, 22]. More generally, p-values can be derived from any scoring system by using a decoy search which is not in competition with the target search [10]. Moreover, for the BH-FDR to be accurate, the p-values must be well-calibrated – i.e., mismatch p-values should uniformly distribute across the [0,1] interval [1]. According to [24, 25], it is possible to assess how well a search engine is calibrated by using *entrapment databases*. Using these databases, it is possible to empirically estimate the quantiles of the mismatch p-value distribution thanks to artefactual amino acid sequences, as described in [10].

We applied both TDC and BH methodologies to results acquired with a Q-Exactive Plus instrument on ten analytical replicates of an *E. coli* lysate, which is classically used as a quality control sample on our platform (see Methods 4.1 and 4.2). PSMs were identified using Mascot and a reverse database to perform TDC. Figure 1A shows the score thresholds (see Methods 4.3) obtained at an FDR of 1% as a function of the various mass tolerances set during database searching (see Supplementary Table 1 for numerical values). MS and MS/MS spectra were acquired at relatively high resolutions (70,000 and 17,500 at m/z 200, respectively), and we considered four combinations of mass tolerance tuning at the precursor and fragment levels: LL, HL, LH and HH (where L stands for low precision or large tolerance, and H for high precision or narrow tolerance), the final combination (HH) corresponds to the tolerance levels generally used on our platform for Q-Exactive data. Several conclusions can be drawn from the results obtained: First, for each tuning taken individually, the TDC threshold is less stable than its BH counterpart (the set of ten cut-off scores was more dispersed with TDC). Second, the TDC threshold is always less conservative than its BH counterpart. Third, depending on the mass tolerances applied during database searching, the TDC threshold on the Mascot score varied from 1.11 to 20.73, whereas its BH counterpart was more stable (between 20.67 and 23.43). Fourth, if the different database search sets are interpreted as surrogates for the recent evolution of instrumental capabilities, the discrepancy between TDC and BH is seen to increase. Thus, when TDC was first applied to data from low-resolution instruments, and for which TDC and BH led to roughly similar results, the resolution of mass spectrometry instruments has progressively increased, and now TDC and BH diverge considerably. To guarantee that modifying how mass tolerance is tuned allows results obtained with lower resolution instruments to be approximated, we analyzed another batch of ten analytical replicates of the *E. coli* lysate using a LTQ-Orbitrap Velos Pro, on which it is possible to use mixed resolutions: high resolution in MS mode (Orbitrap analysis) and lower resolution in MS/MS mode (linear ion trap analysis). Database searches were conducted using the HL and LL tunings, and FDR thresholds were computed as above. Interestingly, the mapping between the thresholds is excellent, which justifies our methodology: from an FDR viewpoint, switching to analysis of a lower resolution dataset using an appropriately-tuned search engine, or retaining the higher resolution data while substantially increasing the mass tolerance, produces roughly equivalent outputs. To better capture the influence of mass tolerance tuning, we then conducted more detailed experiments: Starting from the HH setting, we progressively broadened the mass tolerance range, either for the precursor (Figure 1B) or for fragment masses (Figure 1C). The results obtained support the four conclusions derived from the results shown in Figure 1A.

A part of the TDC's relative instability can naturally be explained by the random nature of decoy sequence generation [8], regardless of the search engine used. However, at first glance, there is no reason

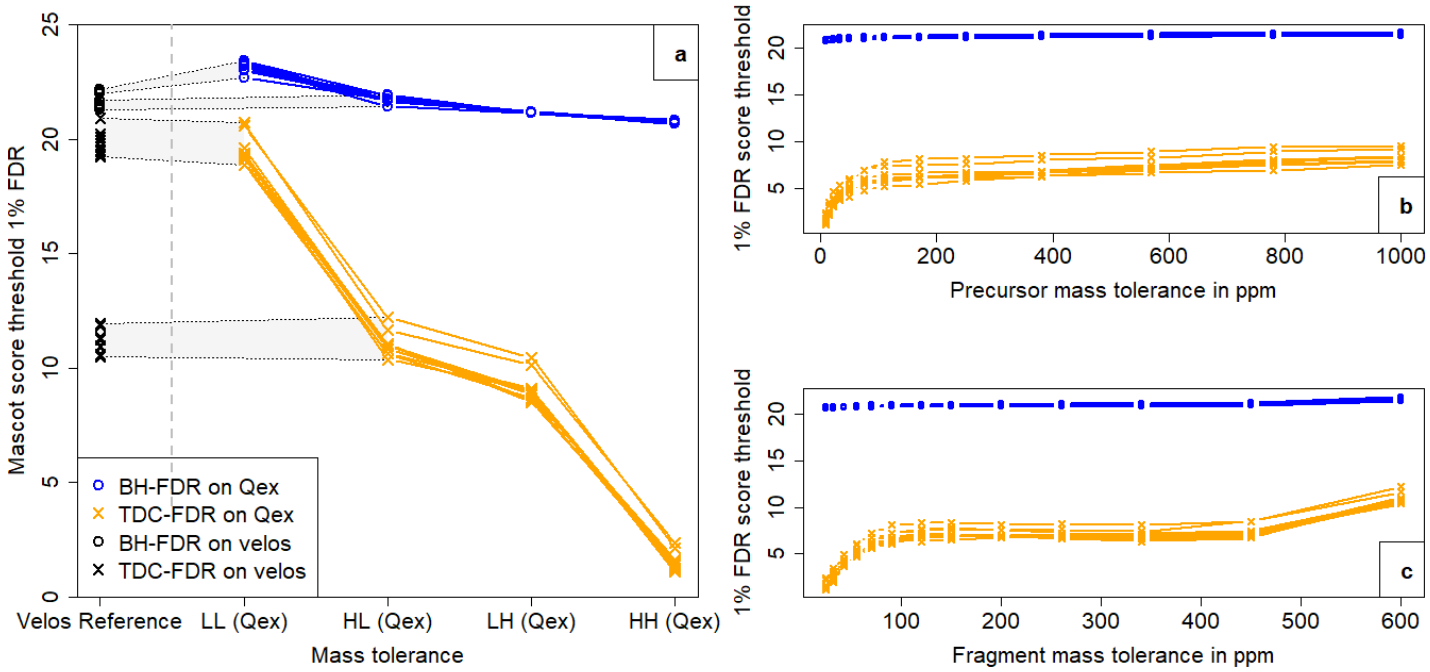


Figure 1: Score thresholds obtained when applying TDC (Orange) and BH (Blue) filtering at an FDR of 1%, as a function of the search engine mass tolerance parameters, for 10 samples analyzed with a Q-Exactive Plus (Qex) instrument. (**a**) Precursor and fragment mass tolerances are tuned to the LL, LH, HL and HH settings: LL assumes the MS and MS/MS data were acquired at low resolutions for the precursor and fragment masses (1 Da and 0.6 Da, respectively); HL uses mass tolerances of 10 ppm and 0.6 Da, respectively; LH uses mass tolerances of 1 Da and 25 mmu; and finally, HH uses mass tolerances of 10 ppm and 25 mmu (which corresponds to classical parameters for database searches performed with Qex data). The black lines encompass thresholds resulting from similar analyses performed on an LTQ-Orbitrap Velos Pro (Velos) with LL and HL settings. (**b**) Refined analysis of the FDR threshold's sensitivity to precursor mass tolerance tuning (Qex data, fragment tolerance= 25 ppm). (**c**) Refined analysis of the FDR threshold's sensitivity to fragment mass tolerance tuning (Qex data, precursor tolerance= 10 ppm).

to assume that the remaining reported instability (notably the drop in score) is specifically linked to the TDC. Consequently, it could also make sense to question the algorithmic specificities of the search engine (here Mascot) (see Methods 4.3). Unfortunately, similar pitfalls (at least for the precursor tolerance parameter) were reported in the supplementary materials of [12] with X!tandem and MS-GF+ (formerly referred to as MS-GFDB). In addition, we discovered similar effects with Andromeda (Maxquant environment, see Supplemental Materials B). Consequently, we believe that it is legitimate to question the TDC procedure itself.

In conclusion, our benchmark confirms the collapse of validation thresholds described by Cooper. Moreover, it proves that not only the precursor mass tolerance, but also, the fragment mass tolerance is involved. More generally, any instrumental or software tuning that could reduce the number of fair decoy competitors in the matching process should be investigated to ensure it does not have spurious consequences. In addition, the TDC-FDR appears to be more sensitive than the BH-FDR, in the sense that reducing the mass tolerances increases its anti-conservativeness faster. Finally, the relative instability of TDC-FDR with respect to BH-FDR is obvious.

2.2 Recovering stable target decoy FDR is possible, but cumbersome

Let us extrapolate an extreme case where, for a long peptide, the precursor mass tolerance window is so narrow that only a single target sequence is eligible, albeit incorrect. Then, in any random case where the single decoy sequence generated within the corresponding precursor mass window is too unrealistic, it is possible to end up with an almost systematic target assignment relying on only a poor recovery of fragment matches, just because the decoy alternative produces even fewer fragment matches. In this situation, the Equal Chance Assumption is clearly violated. Thus, to increase the chances of obtaining decent decoys despite excess narrow mass tolerance parameters, it intuitively makes sense to enlarge the decoy database. Although the link between mass tolerance and decoy database size can be efficiently exploited to limit the computational cost of repeated searches [26], it will, unfortunately, be inefficient in our case: In principle, if d (resp. t) stands for the number of decoys (resp. targets) that have passed the validation threshold, and r is the ratio between the sizes of the whole decoy and target databases, the FDR should read $\frac{d+1}{r\times t}$. However, as demonstrated in the supplemental material to [6], this approximation only holds when $t \to \infty$ and $d/t \leq 5\%$. Consequently, when the validation threshold is set to an FDR of 1%, r must be smaller than 5. Therefore, to increase r in proportion to the reduction in precursor mass tolerance, the FDR should only be controlled at immaterial levels. Alternatively, the TDC can be refined by a procedure akin to *empirical null estimation* [27]. This type of approach has been termed *entrapment FDRs* in the proteomic context [28, 29], because they rely on the quantile distribution of mismatches in an entrapment database to estimate the FDR [24, 25]. We implemented this type of strategy (see Methods 4.4 for details) and found that a fair (but still unstable) empirical estimate of the null distribution could be obtained by performing a separate decoy search, without competition. To stabilize the estimations, we averaged 10 repetitions, each based on a different shuffled database, thus performing a mixed strategy between [10] and [8]. For this reason, for each of the 10 sample replicates of *E. coli* lysate analysed with the Q-Exactive Plus, Figure 2 shows the 10 entrapment FDRs corresponding to the cut-off scores obtained using the two validation methods (BH and TDC, cf. Figure 1A and Supplementary Table 1) in the HH setting.

The difference is striking: although BH thresholds (Mascot scores between 20.67 and 20.84) produce entrapment FDRs slightly below 1% (between 0.53% and 1.29%, with an average ≈0.84%), those obtained with TDC (scores between 1.11 and 2.37) lead to 10-fold larger entrapment FDRs (between 7.72% and 11.71%, with an average ≈9.1%). This result is insightful for three reasons: *(i)* it confirms that TDC can lead to considerable under-estimations of the FDR if used inappropriately; *(ii)* it shows that in contrast to Cooper's concerns, the concept of "peptide presumption" is not inherently biased, since by applying an appropriate decoy search strategy, it is possible to recover coherent FDRs; *(iii)* Using an appropriate strategy is more complex than determining a simple TDC-FDR from a reverse or shuffled database, or than performing a BH procedure. Beyond these observations, if one compares the entrapment FDRs derived from the BH-thresholds with the directly obtained BH-FDR (i.e. 1%), BH estimates appear overall to be slightly larger than entrapment estimates. This result is most probably due to the conservative property of the BH estimator. Note that many methods can be used to limit over-conservativeness [18]. Finally, let us observe that the entrapment FDRs are highly dependent on the entrapment databases: From one randomly generated version to another, the FDR estimated varies significantly. At first glance, FDRs around 1% seem slightly more stable than those around 10%. However, after normalization relative to the mean FDR value, it is actually the opposite that occurs (mean coefficient of variation of 13.90%

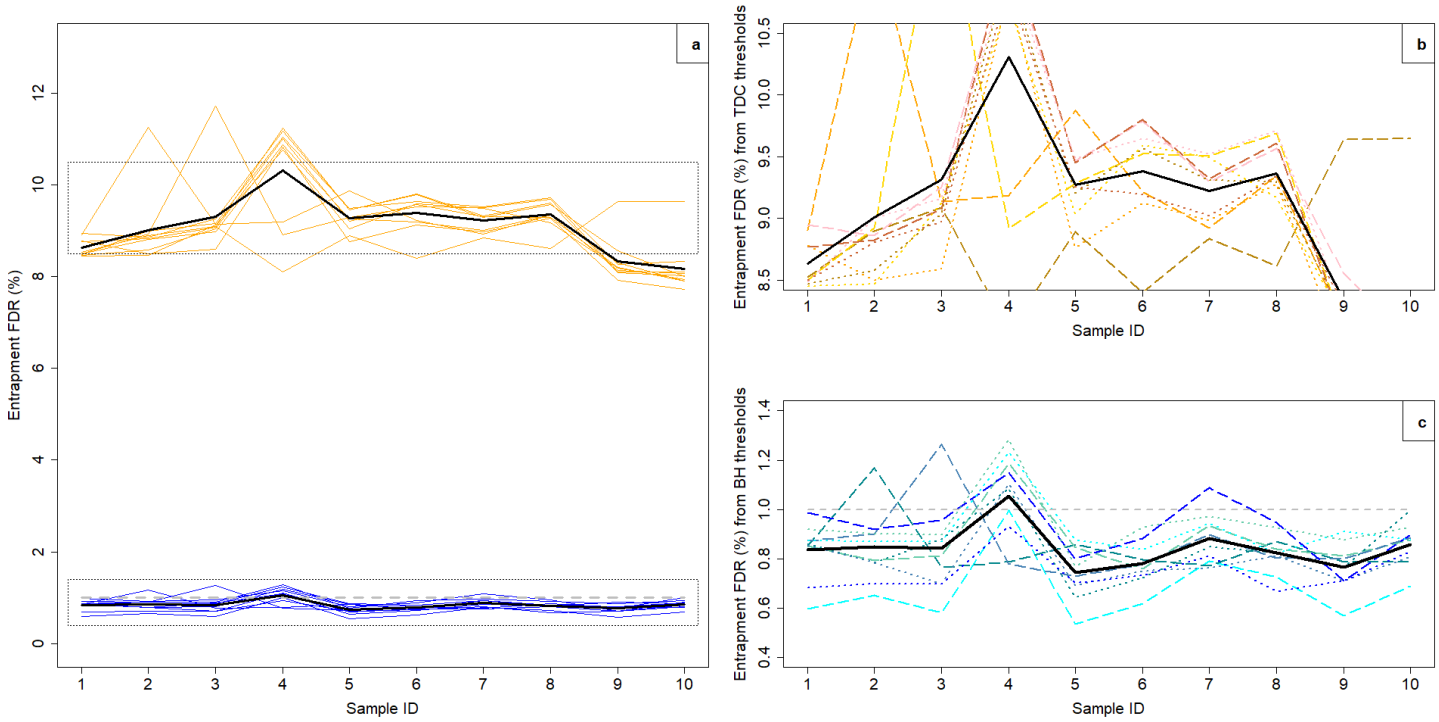


Figure 2: (**a**) Entrapment FDRs computed for each replicate matched against 10 randomly generated (shuffled) entrapment databases, according to its 1% FDR validation score threshold (gray dash line), computed by applying BH (blue) and TDC (orange) methodologies; The black continuous lines depict the average of the 10 entrapment FDRs (boosted estimate). (**b**) and (**c**) Zooms of the two framed areas in (**a**), with different shades of blue (resp. orange) and of line types (dot or dash) for better shuffle discrimination.

around 1% entrapment FDR, versus 5.82% around 10% entrapment FDR). This observation can be easily explained: With lower FDR thresholds, fewer decoys passed the threshold, and as a result, the statistics were computed on smaller sample sizes, which are more sensitive to randomization. This conclusion explains why it is necessary to rely on a boosting strategy (i.e. averaging multiple runs, see above), despite its additional computational complexity.

In conclusion, these experiments show that target and decoy searches can be implemented to deliver accurate FDRs. However, to do so requires a number of strategies to be refined (notably omitting the competition step of the TDC while stabilizing the FDR by averaging multiple independent estimates). These strategies are in fact less intuitive, more hand-crafted and more computationally demanding. The BH procedure, in contrast, is appealing for its simplicity and stability.

2.3 FDR control at peptide and protein levels using the BH procedure

The difficulty of inferring peptide- and protein-level knowledge from spectrum-level information, while applying quality control criteria, has been widely addressed in the literature (see *e.g.* [30, 31] for surveys). However, to our knowledge, all available inference systems require a preliminary decoy search to propose a peptide- or protein-level FDR. Today, combining multiple levels of FDR control has become accepted standard good practice. We therefore propose a generic procedure to extend the BH-FDR approach to peptide and protein levels. Moreover, the proposed method is independent of the chosen inference rules (see Methods 4.5). Hereafter, we assume that the inference rules selected unambiguously define which PSMs should be used in peptide scoring, as well as which peptides contribute to protein group scoring [32, 28, 29], and we focus on the scoring methods applied.

The most classical peptide scoring methods assume that each peptide is identified by the spectrum with the highest PSM score amongst the Q matching spectra [25, 33, 28, 29]. In this setting, it makes sense to define the peptide score as equal to the best PSM score [25]. Formally, if the **PSM score** between peptide sequence $\mathbf{seq}_i$ and spectrum q is referred to as S_{iq}, then, the **best-PSM score** can be defined as

$\max_{q\in[1,Q]} S_{iq}$. This score can potentially be used to compute a TDC-FDR, but not a BH-FDR. Indeed, its probabilistic counterpart cannot be well-calibrated (the minimum of several calibrated p-values is non-uniformly distributed, as illustrated in Supplementary Figure 1). Fortunately, it is possible to modify the best-PSM score by applying a formula conceptually similar to the Šidák correction [34] (although it relies on different mathematical hypotheses), and thus to recover correct calibration:

Proposition 1 *Let $S_1, \ldots, S_n$ be a set of n scores of the form $S_\ell = -10\log_{10}(p_\ell)$, $(\ell \in [1,n])$ where p_ℓ is realizations of n i.i.d. $\mathbb{R}_+$ random variables, $X_1, \ldots, X_n$. If $X_\ell \sim \mathcal{U}[0,1]$ $\forall \ell$, then,*

$$Y = 1 - \left(1 - 10^{-\frac{\max_\ell(S_\ell)}{10}}\right)^n$$

uniformly distributes over the range $[0,1]$.

Proof: See Methods 4.6.

Therefore, (See Methods 4.6 for the full derivations), the **peptide p-value** of peptide sequence $\mathbf{seq}_i$ can be defined as:

$$p_i^\diamond = 10^{-\frac{S_i^\diamond}{10}} \tag{2}$$

and its **peptide score** as:

$$S_i^\diamond = -10 \cdot \log_{10}\left(1 - \left(1 - 10^{-\frac{\max_q(S_{iq})}{10}}\right)^Q\right) \tag{3}$$

To define protein-level scores and p-values, fragment matches for PSM scores were considered equivalent to what peptide matches are for protein scores. This equivalence led us to rely on Fisher's test to define protein scores/p-values from the scores for the best subset of peptides. Similar approaches have frequently been investigated in the literature (see [26, 32, 28, 29]), and the full derivation is presented in Methods 4.8. To the best of our knowledge, we are the first to discuss the adaptation of Fisher's methodology from its original context (meta-analysis) to proteomics by explicitly considering (i) risks of anti-conservativeness due to dependent peptides (see Methods 4.9); (ii) the impact of the poorly conclusive peptide-level evidence in an open-world assumption context (see Methods 4.10). Finally, for a protein sequence $\mathbf{seq}_\pi$ identified by k specific peptides with scores $S_1^\diamond, \ldots, S_K^\diamond$, the **protein p-value** is defined as:

$$p_\pi^\star = 10^{-\frac{S_\pi^\star}{10}} \tag{4}$$

and the **protein score** $S_\pi^\star$ as:

$$S_\pi^\star = -10\log_{10}\left(\min_{A\in 2^{\{1,\ldots,K\}}}\left[\int_{0.2\ln(10)\cdot\sum_{k\in A} S_k^\diamond}^{\infty} f_{2\cdot|A|}(x)dx\right]\right) \tag{5}$$

where: $2^{\{1,\ldots,K\}}$ is the powerset of the set of K peptides identified; A is a peptide set with cardinality $|A| \leq K$; and $f_{2\cdot|A|}$ is the density function of the χ^2 distribution with $2\cdot|A|$ degrees of freedom.

The ten replicate analyses of *E. coli* lysate were validated at 1% FDR by applying the BH procedure to the PSM, peptide and protein scores. To do so, only a target database search was necessary. However, and because it delivered a striking illustration of the capacity of the proposed framework to distinguish false identifications, we introduced shuffled sequences in the searched database to assess the results (see Methods 4.4). We considered a challenging scenario where the number of decoys was set to five times the number of target sequences. Table 1 summarizes the average (across the 10 replicates) cut-off scores as well as the average counts for validated PSMs, peptides and proteins in both target and fivefold shuffled databases (see Supplementary Figure 2 for detailed values). Although the corresponding proportion must not be interpreted as FDRs, it is interesting to discuss them: First, despite the fivefold decoy over-representation, each of the three validation levels (PSM, peptide or protein) taken individually was sufficient to provide a decoy ratio below the FDP expectation of 1% at any level. Second, the three validation strategies provided broadly concurring filters and validated protein list sizes. Third, some discrepancies between the three validation strategies exist (for instance, when filtering at protein level, PSMs with low score are validated because they belong to proteins which are confirmed by other high scoring peptides), leaving room to refine validation with appropriate multi-level filters, as discussed below.

	No validation			PSM validation			Peptide validation			Protein validation		
	PSM	Pep.	Prot.	PSMs	Pep.	Prot.	PSM	Pep.	Prot.	PSM	Pep.	Prot.
Min score	0.002	0.002	0.002	21	19.676	21.187	0.023	20.963	21.163	0.013	0.021	22.534
#Targets	12020.2	9351.3	1466.5	10233.6	8180.3	1297.7	10736.2	8159.5	1297.9	11836.6	9169.4	1294.6
#Decoys	873.3	817.2	786	11.5	10.9	10.7	11.1	10.2	10	13.8	12	9.2

Table 1: Average (across the 10 *E. coli* replicates) minimum score and PSM, peptide and protein counts assigned as target and decoy in the raw dataset (No validation), as well as after validation by one of the three following rules: 1% BH-FDR at PSM level (PSM validation), 1% BH-FDR at peptide level (Peptide validation) and 1% BH-FDR at protein level (Protein validation).

3 Discussion

As a whole, this work sheds new light on a crucial step in bottom-up proteomics experiments: the validation of identification results. First, it illustrates that the TDC and BH estimates of the FDP have progressively diverged as MS accuracy has improved. Our results demonstrated that this divergence originated in the TDC's lack of stability with respect to the precursor and fragment mass tolerances set during database searches. Although this lack of stability can be partially counteracted by suppressing the competition step of TDC [10], the instability induced by the random generation of decoy sequences remains [8]. Therefore, even though target-decoy strategies can be refined to partially cope with this instability, the results are not satisfactory as these strategies: *(i)* are not as stable as BH, in particular at lower FDRs; *(ii)* are more complex to organize (implementation, computational cost) and require additional manual checks; *(iii)* do not provide any guarantee of reliable FDR estimates in the future (on datasets acquired with even higher resolution next generation instruments for which narrower mass tolerances can be expected); *(iv)* fail to provide any guarantee with respect to pipeline modifications that change the number of target and decoy candidates (e.g. [35]).

Second, this work provides new peptide and protein scores which demonstrably respect the calibration conditions of the BH procedure. Indeed, implementing BH-FDR at PSM-, peptide- and protein-level is straightforward (see Code Availability 5) and its practical use within a preexisting platform pipeline requires no precise tuning. Moreover, our results highlighted that, despite slightly different behavior, any of these scores alone is sufficient to conservatively validate a proteomics dataset at PSM, peptide and protein levels. This finding suggests that various strategies could be developed to comply with different objectives: If the expected output is a protein list, then it is probably most appropriate to control the FDR at protein-level. However, in studies seeking to refine discrimination between proteoforms sharing many subsequences, it may be more relevant to validate at peptide level. Finally, when quantifying proteins, extracting the ion current for misidentified spectra produces erroneous results, making validation at PSM level necessary. Beyond these considerations, acting at different levels of filtering may also improve the quality of the validated identifications, although this assertion requires further investigation. For example, multiple FDRs are classically used sequentially, following the inference process (starting at PSM level and ending at protein level); using a reverse order or parallel filtering may also be of interest to preserve the distribution assumption of the BH procedure.

Based on these results, we propose an overhaul of how FDR is estimated in discovery proteomics using database searching and suggest replacing TDC by BH-FDR. Nevertheless, as a theoretical research field, TDC remains of interest. The original idea proposed by Elias and Gygi [2] has stimulated the field of theoretical biostatistics and led to the idea that simulating null tests from the data could produce efficient alternatives to BH procedures, which demonstrably control the FDR [19, 20]. Transferring these theoretical results into biostatistics routines that can be applied on a daily basis still requires some investigation [36]. However, they will hopefully contribute to computational proteomics in the future, as an example of an interdisciplinary virtuous circle.

4 Methods

4.1 Sample preparation and nanoLC-MS/MS analyses

For this work, we used the data obtained with a quality control standard composed of *E. coli* digest, analyzed very regularly in our platform to check instrument performances.

Briefly, competent *E. coli* DH5α cells transformed with pUC19 plasmid were grown at 37°C in LB medium containing carbenicillin before harvesting during exponential phase ($OD_{600} \sim 0.6$). After centrifugation at $3'000 \times g$ during 10 min, the pellet was washed 3 times with cold PBS before lysis of cells

using Bugbuster Protein Extraction Reagent (Novagen) containing cOmplete^{TM}, EDTA-free Protease Inhibitor Cocktail (Roche) and benzonase (Merck Millipore). After centrifugation at $3'000 \times g$ during 30 min and at 4°C, the supernatant was recovered and the protein amount was measured, before protein solubilisation in Laemmli buffer.

Proteins were stacked in a single band in the top of a SDS-PAGE gel (4-12% NuPAGE, Life Technologies) and stained with Coomassie blue R-250 before in-gel digestion using modified trypsin (Promega, sequencing grade) as described in [37].

Resulting peptides were analyzed by online nanoliquid chromatography coupled to tandem MS (UltiMate 3000 and LTQ-Orbitrap Velos Pro, or UltiMate 3000 RSLCnano and Q-Exactive Plus, Thermo Scientific). The equivalent of 100 ng of starting protein material was used for each injection. Peptides were sampled on 300 μm $\times$ 5 mm PepMap C18 precolumns (Thermo Scientific) and separated on 75 μm $\times$ 250 mm C18 columns (Reprosil-Pur 120 C18-AQ, Dr. Maisch HPLC GmBH, 3 μm and 1.9 μm porous spherical silica for respectively UltiMate 3000 and UltiMate 3000 RSLCnano). The nanoLC method consisted of a linear 60-min gradient ranging from 5.1% to 41% of acetonitrile in 0.1% formic acid.

For LTQ-Orbitrap Velos Pro analyses, the spray voltage was set at 1.5 kV and the heated capillary was adjusted to 200°C. Survey full-scan MS spectra ($m/z = 400 - 1600$) were acquired with a resolution of 60'000 at m/z 400 after the accumulation of 10^6 ions (maximum filling time 500 ms). The twenty most intense ions from the preview survey scan delivered by the Orbitrap were fragmented by collision-induced dissociation (collision energy 35%) in the LTQ after accumulation of 10^4 ions (maximum filling time 100 ms). MS and MS/MS data were acquired using the software Xcalibur (Thermo Scientific). For Q-Exactive Plus analyses, the spray voltage was set at 1.5 kV and the heated capillary was adjusted to 250°C. Survey full-scan MS spectra ($m/z = 400 - 1600$) were acquired with a resolution of 60'000 at m/z 400 after the accumulation of 10^6 ions (maximum filling time 200 ms). The ten most intense ions were fragmented by higher-energy collisional dissociation (normalized collision energy 30%) after accumulation of 10^5 ions (maximum filling time 50 ms) and spectra were acquired with a resolution of 15'000 at m/z 400. MS and MS/MS data were acquired using the software Xcalibur (Thermo Scientific).

4.2 MS data analysis

Data were processed automatically using Mascot Distiller software (version 2.6, Matrix Science). Peptides and proteins were identified using Mascot (version 2.6) through concomitant searches against *Escherichia coli* K12 reference proteome (20180727 version downloaded from Uniprot), and/or custom made decoy databases (reversed or shuffled sequences - see Methods 4.4 for details). Trypsin/P was chosen as the enzyme and 2 missed cleavages were allowed. Precursor and fragment mass error tolerance has been variably adjusted as described in the manuscript. Peptide modifications allowed during the search were: carbamidomethylation (C, fixed), acetyl (Protein N-ter, variable) and oxidation (M, variable). The Proline software (`http://proline.profiproteomics.fr`) was used to filter the results: conservation of rank 1 peptide-spectrum match (PSM) and single PSM per query. 1% FDR control was performed with various methods, according to the Results (see Section 2).

4.3 Selection of the adequate score

Initially, TDC was designed to operate on raw PSM scores (see Definition 1 in Methods 4.6), *i.e.* on scores which individually quantify the similarity between an experimental spectrum and a theoretical one; irrespective of any additional information concerning the distribution of other (real or tentative) matches. However, the last decade has witnessed the publication of many "contextualized scores": Despite being rather diverse, these scores all leverage the same idea of relying on the target and the decoy score distributions to improve the discrimination between correct and incorrect matches. This can be concretely achieved by defining a delta score, *i.e.* a difference between the best PSM score and a statistics summarizing the score distribution (the second best candidate in MS-GF+ [38] and in Andromeda/Maxquant, the homology threshold in Mascot) an e-value (X!tandem) [39] or a posterior probability (in Andromeda/Maxquant or in the Prophet post-processing tool suites [40]). Many comparisons have experimentally assessed the improved discrimination capabilities of these thresholds. In a nutshell, by helping discarding the high scoring decoy matches, lower FDRs can be achieved.

Among these contextualized scores, it is not clear which ones require a TDC, which ones require a decoy database but not necessarily a competition step, and which ones are only defined on the target match distribution. This is not necessarily an issue, as long as TDC is assumed to fairly represent the mismatch distribution. However, in this work we hypothesized that the Equal Chance Assumption could be violated in some practical conditions, so that no assumption can be made on the decoy distribution

correctness. This is a real issue for our experiments: If one uses a contextualized score to assess the quality of the TDC-FDR, while the contextualized score is built on top of TDC, one faces a self-justifying loop: It is essentially the same issue as over-fitting some data or as using the same dataset to learn and to test a machine learning algorithm.

To cope for this, we have decided to evaluate the quality of TDC-FDR by relying on individual / raw scores. The rational is the following: If TDC-FDR is stable with respect to the tuning of various mass tolerance parameters in the search engine, then, the validated PSMs with the lowest scores should have roughly the same "absolute" quality (*i.e.* irrespective of the other scores) whatever the search engine tuning. Contrarily to a well-spread belief, when one filters a PSM list at 1% FDR, it does not mean that we allow 1% of poor matches in the dataset. On the contrary, it means that, despite all the validated PSMs apparently depict matches of sufficient quality, 1% of them are spurious. In other words, the validated PSMs with the lowest scores are not randomly selected mismatches that make the list longer because 1% of false discoveries are tolerated; but borderline PSMs that nonetheless meet the quality standard of a 1% FDR validation. In this context, it makes sense to assume their quality should remain roughly constant whatever the search engine tuning.

Once it has been decided to use the lowest individual PSM scores to evaluate the stability of TDC across various conditions of applications, one has to select a subset of search engines to perform the experiment. This is a touchy subject as any TDC criticism can be read as a strike against a given search engine [11, 41, 12]. In our view, the five most widely used search engines are the following [23]: Andromeda (from Maxquant suite), Mascot, MS-GF+, SEQUEST and X!tandem. Among them, Sequest is more a core algorithm that derives in a multitude of tools [42] which have different implementations, optimizations and control parameters. As for Andromeda, it possesses many layers of scores that cannot be accessed, and which results in behaviour that should be questioned before using it in a TDC evaluation. Notably in case of very long peptides, it is customary to observed validated PSMs with fairly high posterior probability, while the similarity score is zero. Obviously, this questions the prior distributions which are involved in turning a zero score into an almost certain match, and consequently the possible construction of these posteriors from decoy matches. Finally, the last three search engines (*i.e.* X!tandem, MS-GF+ and Mascot) have already been reported to lead to similar score downfalls [12]. As a result, we have focused on Mascot (which is the most popular among the three of them and for which p-values can be straightforwardly derived) and we have postponed the study of Andromeda (see Supplemental Materials B).

As a side note, let us stress that this evaluation protocol should not be over-interpreted. Its context of use is the following, strictly: We aim at evaluating the stability of TDC, independently of the search engine or the scoring methodology. Considering, the presented evaluation protocol should not be understood as a prejudiced view on any search engine, or on any contextualized score. Notably, some contextualized score could as-a-matter-of-factly over-exploit TDC, *de facto* leading to larger FDR under-estimations (as demonstrated by [14]), while on the contrary, some others may partially cope for the problem by stabilizing the FDR. Although these questions are of interest, they stand beyond the scope of this work.

4.4 Decoy database generation

For classical TDC experiments (Figure 1), we used the following procedure: The target database was reversed by using the Perl script (`decoy.pl`) supplied with Mascot software and the generated decoy database was appended to the target one before concatenated search. From our observations, slightly different procedures (shuffled vs. reversed, accounting for trypsic cleavage site, etc.) yields similar results, which concurs with the observations described in [9].

To compute an entrapment FDR, we simplified the various methods developed in the recent years [24, 43, 44, 25, 28, 29], for in our case, we did not need an assessment of the calibration of the search engine (as in [25]), nor a combined estimation of a classical TDC-FDR and of an entrapment FDR (as in [28]). Concretely, we only needed a sufficiently realistic empirical model of the null distribution [10, 26, 24, 27], notably one which accounts for the increment of the search engine scores as a consequence of the competition between several decoy sequences (and which presence was the justification to design a tool to entrap them). However, this competition being of different nature than the competition between target and decoy assignments, avoiding interferences between both types of competition is crucial to accurate FDR computation. In this context, we have followed the proposal from [10], [26] and [14] which have long ago pinpointed that a realistic empirical model of the null distribution could be derived from a classical decoy database search; and that this derivation was straightforward in absence of competition between the target and decoy assignments. Although this observation blurs the line between the entrapment

procedure and classical decoy search, the proximity of the two concepts has already been mentioned in [43].

However, we remarked that this strategy was successful only if the decoy database was of exactly the same size of the target. Otherwise, the competition model was not exactly the same (roughly, the larger the database, the greater the chances that the best mismatch score increases), leading to a biased empirical estimate of the null. A practical consequence of using a decoy database of the same size as the target is that the resulting (discrete) null distribution is rather unstable [10]. Thus, to cope for this instability we have relied on a boosting strategy [8] where one averages the FDR estimates of multiple shuffle decoy databases. In our case, we used 10 shuffles, and the shuffling procedure respects the cleavage site so as to respect the precursor mass distribution [7].

4.5 Inference rules

Peptide inference or *protein inference* are umbrella terms which encompass several distinct notions: Inference rules, scoring methodologies and quality control metrics. The **inference rules** define which pieces of information should be considered and how they should be used in the inference process, regardless of their quality or reliability. For the spectrum-to-peptide inference, this notably refers to the possibly multiple ranked interpretation of a spectrum. For the peptide-to-protein inference, this refers to the minimum number of peptides per protein, the processing of shared peptides and the definition of protein groups. The **scoring methodology** refers to the definition of a confidence value for each entity (defined thanks to the inference rules), in order to rank them from the most confident to the least one. Finally, the **quality control metrics** is used to filter out some insufficiently reliable entities so as to keep a shorter list of validated ones. The metrics can either be individual (each entity is considered independently of the others) such as with Posterior Error Probability [45, 40]; or associated to the entire filtered list (typically, an FDR, but other multiple test corrections methods exist [46]).

Although conceptually distinct, these notions can overlap in practice, see [32, 28, 29]: Some inference rules directly involve the scoring methodology; Quality control metrics may tightly relate to the scoring methodology; Inference rules and scoring systems are compared so as to find the combination leading to the lowest FDRs; Etc. However, for sake of generality, we kept here a clear distinction. Concretely: *(i)* We did not address the definition of inference rules, and we considered the most simple one (*i.e.* a single peptide interpretation per spectra and only protein-group specific peptides, regardless their number and the protein grouping), and leave to future work (or to any inspired reader) the application of our procedure to more elaborated inference rules; *(ii)* We focused on the scoring methodology, with the objective to preserve the $[0, 1]$-uniform distribution, so as to call them well-calibrated peptide/protein p-values; *(iii)* Regarding the quality control metrics, we obviously relied on BH procedure, which becomes possible thanks to the calibration correctness.

This complete separation provides two advantages. First, it makes it possible to reason on each step independently of the others. Notably, this article focuses on the scoring methodology independently of the inference rules and the quality control metrics. Second, it enables a distinction between the quality control level and and the nature of the controlled entities: While it is customary to validate a list of proteins with an FDR of 1%, it is not as classical to validate a list of PSMs with the criterion that less than 1% of the proteins they map on are putatively false discoveries. However, as illustrated in the Discussion (see Section 3), such options are really insightful.

4.6 Peptide score definition

Definition 1 *Let us have a peptide sequence* $\mathbf{seq}_i$*, a spectrum* $\mathbf{spec}_j$ *and a score reading*

$$S^\circ_{ij} = \texttt{Score}(\mathbf{seq}_i, \mathbf{spec}_j) \tag{6}$$

that is provided by a search engine. The triplet $(\mathbf{seq}_i, \mathbf{spec}_j, S^\circ_{ij})$ *formally defines a Peptide-Spectrum-Match (or a PSM). To avoid ambiguity with other scoring system,* S°_{ij} *is referred to as the* **PSM score**.

In the rest of this article, we make the following assumption:

Assumption 1 *The search engine provides a PSM score* S°_{ij} *of the form* $S^\circ_{ij} = -10\log_{10}(p^\circ_{ij})$ *where* p°_{ij} *is probability of a random match.*

In the setting of Ass. 1, by construction, p°_{ij} is the p-value of a test with the following null hypothesis:

$$\mathbf{H}_0^{ij} : \mathbf{spec}_j \neq \mathbf{seq}_i$$

which simply means that the peptide sequence and the observed spectrum do not correspond. A direct consequence of Ass. 1 reads:

Corollary 1 *Under* $\mathbf{H}_0^{ij}$ *(*i.e. *when considering only false PSM),* p_{ij}° *is known to distribute uniformly.*

Remark 1 *See [1] for justifications of Cor. 1.*

In other words, if symbol $\approx$ is used to represent the term "look like", then p_{ij}° corresponds to the following conditional probability:

$$p_{ij}^\circ = \mathbb{P}(\mathbf{spec}_j \approx \mathbf{seq}_i \mid \mathbf{spec}_j \neq \mathbf{seq}_i).$$

In practice, several spectra are acquired for each precursor ion, so that several PSMs participate to the identification of a same peptide sequence. This is why, one classically defines the **best-PSM score** of sequence $\mathbf{seq}_i$ (noted $S_i^\top$) as the maximum PSM score among the PSMs involving that peptide sequence:

$$S_i^\top = \max_{q\in[1,Q]} S_{iq}^\circ \tag{7}$$

where Q is the number of spectra that are possibly considered for a match onto $\mathbf{seq}_i$. Let us denote by $p_i^\top$ the corresponding probability, linked to $S_i^\top$ by Ass. 1. It rewrites as:

$$p_i^\top = \min_{q\in[1,Q]} p_{iq}^\circ \tag{8}$$

In other words, $p_i^\top$ is the minimum value of a set of p-values. We would like to interpret $p_i^\top$ as the p-value resulting from testing of the following null hypothesis:

$$\mathbf{H}_0^i : \forall q \in [1, Q],\ \mathbf{spec}_q \neq \mathbf{seq}_i$$

or with a more compact notation,

$$\mathbf{H}_0^i : \mathbf{seq}_i^?$$

the interrogation mark simply indicating that $\mathbf{seq}_i$ does not corresponds to any observed spectrum. Unfortunately, this is not possible: taking the minimum promotes small p-values, so that one should not expect the $p^\top$'s to distribute uniformly under the null hypothesis, which is required to have well-calibrated statistical test and to apply BH procedure. Fortunately, it is possible to recover exact calibration thanks to Prop 1.

Proposition 1 *Let* $S_1, \ldots, S_n$ *be a set of* n *scores of the form* $S_\ell = -10\log_{10}(p_\ell)$, $(\ell \in [1, n])$ *where the* p_ℓ*'s are realizations of* n *i.i.d.* $\mathbb{R}_+$ *random variables,* $X_1, \ldots, X_n$*. If* $X_\ell \sim \mathcal{U}[0, 1]$ $\forall \ell$*, then,*

$$Y = 1 - \left(1 - 10^{-\frac{1}{10}\cdot \max_\ell S_\ell}\right)^n$$

uniformly distributes in $[0, 1]$.

Proof:

$$\begin{aligned}
\mathbb{P}[Y \leq t] &= \mathbb{P}\left[1 - \left(1 - 10^{-\frac{1}{10}\cdot\max_\ell S_\ell}\right)^n \leq t\right] \\
&= \mathbb{P}\left[1 - \left(1 - \min_\ell \left[10^{-\frac{S_\ell}{10}}\right]\right)^n \leq t\right] \\
&= \mathbb{P}[1 - (1 - \min_\ell[p_\ell])^n \leq t] \\
&= \mathbb{P}[1 - (\max_\ell[1 - p_\ell])^n \leq t] \\
&= \mathbb{P}[(\max_\ell[1 - p_\ell])^n \geq 1 - t] \\
&= 1 - \mathbb{P}[(\max_\ell[1 - p_\ell])^n < 1 - t] \\
&= 1 - \mathbb{P}[\max_\ell[1 - p_\ell] < (1 - t)^{1/n}] \\
&= 1 - \mathbb{P}\left[\bigcup_\ell \left\{(1 - p_\ell) < (1 - t)^{1/n}\right\}\right] \\
&= 1 - \prod_{\ell=1}^n \mathbb{P}[1 - p_\ell < (1 - t)^{1/n}] \\
&= 1 - \prod_{\ell=1}^n \mathbb{P}[p_\ell \geq 1 - (1 - t)^{1/n}]
\end{aligned} \tag{9}$$

As each p_ℓ is the realization of a $\mathcal{U}[0,1]$ random variable, one has, $\forall \ell$:

$$\begin{aligned}\mathbb{P}[p_\ell \geq 1-(1-t)^{1/n}] &= 1-(1-(1-t)^{1/n}) \\ &= (1-t)^{1/n} \end{aligned} \tag{10}$$

So that

$$\begin{aligned}\mathbb{P}[Y \leq t] &= 1-\prod_{\ell=1}^{n}(1-t)^{1/n} \\ &= t \end{aligned} \tag{11}$$

Consequently, the cumulative distribution function of Y is that of a uniform random variable. Moreover, Y takes its value in $[0,1]$, strictly. □

As well-calibration is equivalent to uniform distribution of mismatch scores, if the PSM scoring system is well-calibrated, then according to Prop 1, the best-PSM probability can be transformed to be well-calibrated too. Concretely, uniformity under the null is thus recovered by applying a transform akin to that of Šidák correction [34] to $p^\top$, where one defines the **peptide p-value** of peptide sequence $\mathbf{seq}_i$ as:

$$p_i^\diamond = 1-(1-p_i^\top)^Q \tag{12}$$

The corresponding **peptide score** (defined under Ass. 1) is noted $S_i^\diamond$.

4.7 Accounting for fragmentation multiplicity

To the best of our knowledge, the peptide score resulting from Prop. 1 has never been proposed so far. However, similar mathematical derivations (*i.e.* also rooted in the Šidák correction for multiple testing) have already been applied in the proteomic context, notably to recalibrate scoring systems in a context of multiple peptide interpretations of spectra [24] – which differs from the present context. Besides, the aggregation of PSM scores into well-calibrated peptide-level scores has been already addressed by almost the same authors, notably in [25]. This article focuses on controlling the FDR at peptide-level within a TDC context, which probably explains the numerous discrepancies between their work and ours. Essentially, the article compare three different approaches: ETWO, WOTE and Fisher's combined probability test. ETWO and WOTE are both based on the best-PSM score, their difference relying in how the filtering of the non-best PSMs interplays with the TDC. As for Fisher's method, one converts all the PSM scores into PSM p-values (using Eq. 1 or a similar formula, depending on the search engine), before applying Fisher's combined probability test [47], which returns p-values at peptide level (that can finally be transformed back into peptide level scores). As a result of the comparisons, it appears that the WOTE method is the best calibrated one, tightly followed by ETWO, while Fisher method provides miscalibrated p-values. As by construction, Fisher method should provide well-calibrated p-values when combining independent tests, the authors explain this miscalibration by underlying that different PSMs pointing toward a same peptide cannot be considered as independent. We agree with this explanation and we believe it is possible to go further. Due to the the strong dependence of PSMs, using Fisher method should lead to dubious peptide scores, because of the following undesirable effect: several PSMs with intermediate scores pointing toward a given peptide are practically considered equivalent to a single PSM with an excellent score, as illustrated in the following example:

Example 1 *Consider six spectra pointing toward a peptide sequence* $\mathbf{seq}_1$*, all with a Mascot scores of 18. Besides, one has another peptide sequence* $\mathbf{seq}_2$ *identified by a single PSM with a Mascot score of 58.084. According to Fisher method, the peptide scores of* $\mathbf{seq}_1$ *and of* $\mathbf{seq}_2$ *are equal, indicating similar confidence in both peptide identifications.*

This example contradicts with peptide scoring expectations. In fact, the PSM/peptide relationship is not the same as the peptide/protein one, so it is not surprising that that Fisher method, which is helpful to switch from peptide to protein level is not to switch from PSM to peptide level.

It is interesting to note that according to our proposal, the mathematical tool suggested in [24] is of interest to solve the question raised in [25], even though bridging them has never been proposed so far. This can be explained by a noticeable drawback of the proposed peptide score: The greater the number Q of PSMs pointing toward a given peptide, the smaller the score (see Figure 3). As this contradicts with the intuition that the more observed spectra, the likelier the peptide presence, a refined analysis is necessary.

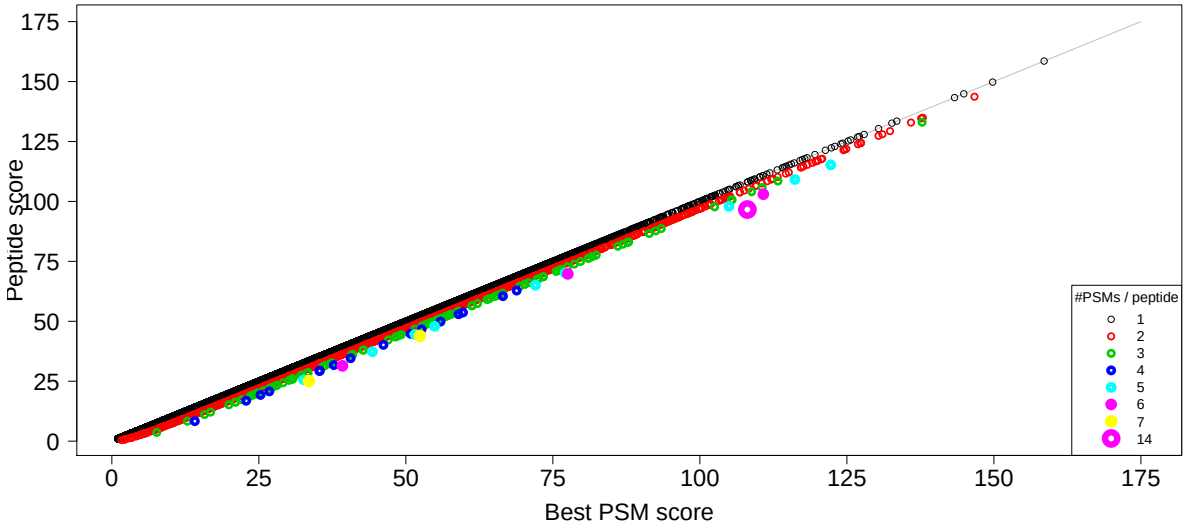


Figure 3: Peptide scores versus best-PSM scores for the *E. Coli* dataset: In the logarithmic scale induced by the conversion from p-values, the scores are deflated by a constant that is proportional to the number of matches. Relatively, the confidently identified peptides are less impacted than other ones.

From a statistical viewpoint, this penalty is well motivated: Let us consider two ions I_1 and I_2. If I_1 is fragmented two times more than I_2, it is two times more likely to reach a higher score thanks to random fluctuations (by analogy, it is easier to obtain a high score when taking the best out of 2 dices than when throwing a single dice). Thus, our Šidák-like correction is essential to avoid an increment of the peptide scores which would only result in the repetition of multiple randomized tests. Contrarily to Fisher method (discussed above through Ex. 1), the mathematical assumption underlying our correction is not that PSMs are independent; but only that their random fluctations are, *i.e.* a weaker and more realistic assumption. However, from an analytical viewpoint, this assumption is still unrealistic. In the case of high-flying ions with long elution profiles, it is possible to obtain a large number of repeated fragmentation spectra (up to 50 or 80, depending on the dynamic exclusion tuning of the mass spectrometers, on the LC length and on the complexity of the sample) with limited (and consequently correlated) random fluctuations. In such a case, the assumption on which our Šidák-like correction is based does not hold so that it should not be used to deteriorate the identification scores.

Finally, one has to find a trade-off between the necessity of correcting for multiple testing, while avoiding too systematic corrections. From our experience, such deterioration only marginally occurs and only impact ions with excellent scores which are validated regardless of a small score reduction (in fact, the higher the best-PSM score, the weaker the correction, as illustrated on Figure 4). As a result, the increased conservativeness of the correction has globally a positive effect on validation (see Section 2.3), even though more investigating for refined strategies can be of interest.

4.8 Fisher combined probability test

To define the protein-level counterpart to PSMs and peptides, we leverage the intuition that fragment matches are for PSM scores what peptide matches are for protein scores. Instead of peptide sequence $\mathbf{seq}_i$, one simply has a protein sequence $\mathbf{seq}_\pi$. As for spectrum $\mathbf{spec}_j$, one has a collection of spectra $\{\mathbf{spec}_j\}_{j\in\mathbb{N}}$ that could potentially match to any of K subsequences of $\mathbf{seq}_\pi$. The goal is thus to derive score and p-value for protein sequence $\mathbf{seq}_\pi$ as counterparts to peptide score and p-value (each score/p-value couple being linked by the $-10\log_{10}(.)$ transform). However, let us first make another assumption:

Assumption 2 *The protein sequence* $\mathbf{seq}_\pi$ *does not share any subsequence with other proteins. As a result, any identified peptide sequence corresponds to a protein-specific peptide.*

Remark 2 *Ass 2 is rather strong. In practice, one can simply restrict the analysis to peptides that are specific to a protein and discard the others, as done in many proteomic software tools. It is also classical*

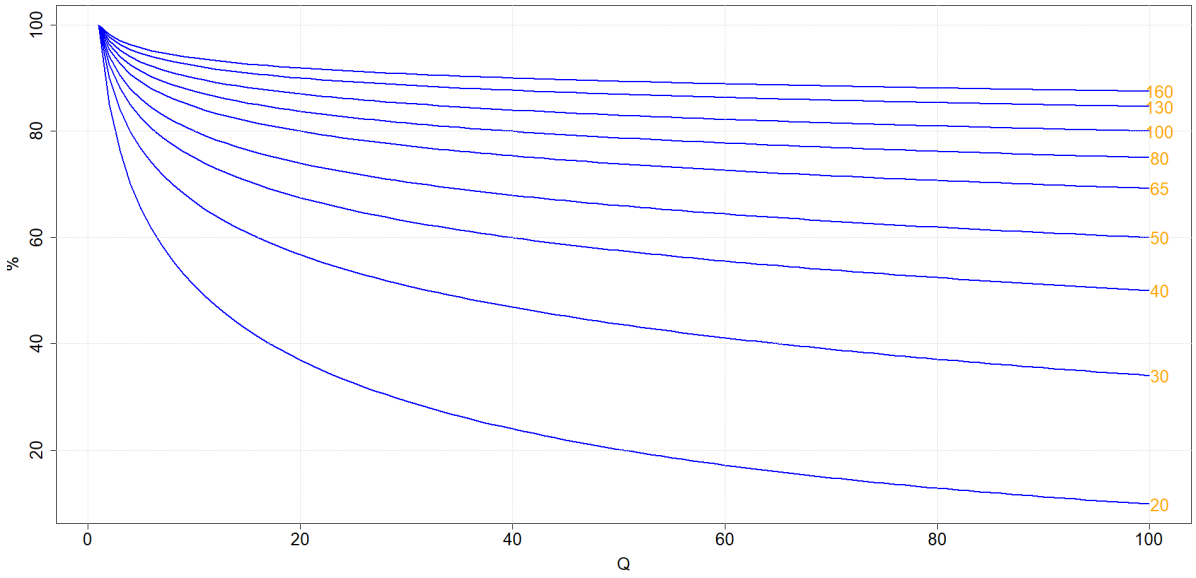


Figure 4: Peptide score expressed as a percentage of the best-PSM score in function of the number Q of PSMs: Each curve corresponds to the evolution of a given score, written in orange on the right hand side of the curve. The plot reads as follow: for Q=1, the correction is idle, so that the peptide score equates the best-PSM score (100%). The peptide score diminishes as Q grows (the plots stopped at Q=100, depicting an already extreme situation where 100 PSMs point toward the same peptide). For instance, if the best-PSM has a score of 80 (fourth curve form the top), and if reaching such a high score has required 39 additional PSMs with lower or equal scores, the peptide score is equal to 80% of the best-PSM score, *i.e.* 64.

to define equivalence classes on the peptide-protein graph, leading to so-called ***protein-groups****.*

The next step is to define $\mathbf{H}_0^\pi$, the null hypothesis which testing would result in the desired p-value. Intuitively, if the protein $\mathbf{seq}_\pi$ is in the sample, one expects at least one spectrum to match on one of the K peptide sequences; and the more matched peptide sequences, the better. However, one should not expect that all the K peptide sequences are confidently matched: For example some sequences may correspond to chemical species that are difficult to ionize. Conversely, if the protein is not in the sample, only random match(es) should occur on one or few peptide sequence(s). This leads to the following null hypothesis:

$$\mathbf{H}_0^\pi : \forall k \in [1, K], \mathbf{seq}_{\pi,k}^?$$

or

$$\mathbf{H}_0^\pi : \forall k \in [k, K], \mathbf{H}_0^k \text{ is true,}$$

where $\mathbf{seq}_{\pi,k}$ represents the kth subsequence of $\mathbf{seq}_\pi$. The corresponding alternative hypothesis $\mathbf{H}_1^\pi$ is that among the K subsequences, at least one is matched. This reads:

$$\mathbf{H}_1^\pi : \exists k \in [1, K] \text{ such that } \mathbf{H}_0^k \text{ is false.}$$

As a matter of fact, these null and alternative hypotheses are those of a combined probability test built on the K peptide p-values $p_1^\diamond, \ldots, p_K^\diamond$. In other words, a p-value at protein level can be computed according to Fisher's method [47] (or possibly a related test, see [26, 32]), which relies on the fact that:

$$-2 \sum_{k=1}^{K} \ln(p_k^\diamond) \sim \chi^2_{2K},$$

where χ^2_{2K} is the Chi-squared distribution with $2K$ degrees of freedom. This is equivalent to:

$$\frac{\ln(10)}{5} \sum_{k=1}^{K} S_k^\diamond \sim \chi^2_{2K}.$$

Thus, if f_{2K} denotes the density function of χ^2_{2K} and $S^{\oplus}_{\pi} = \sum_{k=1}^{K} S^{\diamond}_{k}$ (*i.e.* the sum of the K peptide scores for protein π), then the p-values of the combined probability test reads:

$$p^{\dagger}_{\pi} = \int_{0.2\ln(10)S^{\oplus}_{\pi}}^{\infty} f_{2K}(x)dx \tag{13}$$

Let us call $p^{\dagger}_{\pi}$ the **Fisher p-value** and $S^{\dagger}_{\pi} = -10\log_{10}(p^{\dagger}_{\pi})$ the **Fisher score**.

4.9 Enforcing the conservativeness of Fisher method

The accuracy of Fisher's method is known in the case where the K p-values derive from independent tests. However, in case of dependent tests, it can possibly be anti-conservative, *i.e.* $p^{\dagger}_{\pi}$ can be underestimated (or conversely $S^{\dagger}_{\pi}$ can be overestimated), leading to too optimistic decisions. For instance, providing a too great score to a given protein makes the practitioner overly confident on the presence of the protein.

Therefore, let us analyze the possible dependencies between two tests with null hypotheses $\mathbf{H}^{i}_{0}$ and $\mathbf{H}^{j}_{0}$, respectively. In the case where protein $\mathbf{seq}_{\pi}$ is not in the sample, which corresponds to being under $\mathbf{H}^{\pi}_{0}$, the two tests are clearly independent: The quality of any random match is not related to the existence of other spectra matching on any other sequence of the same protein. However, if protein $\mathbf{seq}_{\pi}$ is present in the sample, the two tests should tend to reject $\mathbf{H}^{i}_{0}$ and $\mathbf{H}^{j}_{0}$, respectively: independence cannot be assumed. Therefore, the independence assumption necessary to the conservativeness of Fisher's method only holds under the null (in fact, this can be easily observed in practice: false PSMs are spread on numerous proteins while true ones tend to concentrate).

As explained above, in general, anti-conservative tests cannot be used, for they lead to too optimistic decision making. This is notably why, when using Fisher's method to combine different statistical analyses into meta-analysis, the independence of the analyses is of the utmost importance: In the case were the meta-analysis confirms the discovery of the analyses (which means rejecting their null hypothesis), it may do so with a too great confidence. This is also why, numerous alternatives to Fisher's method are available in the literature, such as for instance [48, 49].

However, in proteomics, one is seldom interested in providing a confidence level for each protein separately: Or at least, if one is, then, other tools exists, such as PEP / local FDR [45]. Most of the time, the practitioner needs to provide a list of confidently identified proteins, endowed with a quality control metric, such as the FDR. If within this list, all the Fisher scores are inflated, it is ultimately not a problem, as long as the FDR is well-controlled. In other words, it is not important to be overly optimistic with true identifications, as long as one is not with false identifications. In fact, being anti-conservative with true identifications while conservative with the false ones may be a good way to help discriminating them, and thus, reduce the FDP induced by the cut-off score (roughly, this leads the score histograms of true and false discoveries to have a smaller overlap).

As a conclusion, despite the independence assumption only holds under $\mathbf{H}^{\pi}_{0}$, Fisher score will not lead to an increment of false discoveries; at least, as long as one validates the protein identification list with an FDR only.

4.10 Accounting for poorly identified peptides

A deeper look on Fisher's combined probability test behavior pinpoints an undesirable effect for proteomics: A very low p-value can be moderated by greater p-values. This concretely makes sense in a meta-analysis setting, where two poorly conclusive analyses will soften the conclusion of a third very conclusive analysis. However, in proteomics, two poorly reliable peptide identifications should not soften the parent protein identification, if the latter is supported by a third highly reliable (specific) peptide. Let us illustrate this on an example:

Example 2 *Consider a protein π with 4 peptides with score $S^{\diamond}_{1} = 44.09$, $S^{\diamond}_{2} = 1.59$, $S^{\diamond}_{3} = 1.59$ and $S^{\diamond}_{4} = 1.59$. The corresponding Fisher score is* 23.90638. *In fact, the 3 last peptides are so unreliable that they moderate the score resulting from the single observation of the best peptide. Intuitively, a score of* 44.09 *(related to highest protein evidence) would have been preferable.*

Concretely, peptides 2, 3 and 4 in Ex. 2 having small scores does not mean they are not present in the sample, but only that one did not have sufficiently good spectra.

This outlines an intrinsic limitation of Fisher score. It originates in the fact that, contrarily to the setting Fisher's method was originally designed for, mass spetrometry based proteomics relies on the Open World Assumption (OWA, [50]): an absence of observation does not mean non-existence. Concretely, a

low score does not mean the peptide is not present in the sample, but only that one does not have a sufficiently good spectrum. This is why, it intuitively makes sense to consider that, given a protein, its score should not be the combination of all its peptides, but only the combination of the scores of the peptide subset which gives the highest Fisher score. This leads to the following definition of the **protein score** $S^\star_\pi$:

$$S^\star_\pi = \max_{A \in 2^{\{1,\dots,K\}}} S^\dagger_A \tag{14}$$

where $S^\dagger_A$ denotes the Fisher score of the subset A of the set of K peptides that maps onto protein π. If one defines

$$p^\star_\pi = 10^{-\frac{S^\star_\pi}{10}} \tag{15}$$

then, the **protein p-value** $p^\star_\pi$ relates to the Fisher p-value by the following formula:

$$p^\star_\pi = \min_{A \in 2^{\{1,\dots,K\}}} p^\dagger_A \tag{16}$$

Elaborating on methods akin to Fisher test to derive a protein level score has already been proposed in the literature. Conceptually, the method closest to ours is also the oldest [26]: The authors proposed to apply Stouffer's method [51] (which is akin to that of Fisher) on the best subset of peptides to define the protein p-value. Several differences are noticeable with respect to our proposal: First, their scoring system is used in a target-decoy context. Second, the anti-conservative behaviour induced by dependencies is not discussed. Third, the restriction to the best subset of peptides is not interpreted under the open world assumption, so that it is accompanied with a multiple test correction. Fourth, the protein score is directly based on the best-PSM score, without any intermediate recalibration at peptide level. This could lead to miscalibration, however, the specific distribution of best-PSM scores is accounted for by a Gumble law fit. More recently, numerous works have investigated a related path, yet with an objective that seems closer to the comparison and the design of protein inference rules (a subject that is not investigated in this article), rather than quality control procedures. Notably, they extensively discuss the involvement of shared peptides with regards to the protein groups, and simply resort to use TDC to estimate a protein-level FDR: In [32], authors follow a path similar to that of [26]. However, several differences exists: First, another variant of Fisher method [52] is used, which makes it possible to account for shared peptides (by down-weighting them in the combination process). Second, the method is directly applied at PSM level to derive protein-level scores. Third, it does not focus on FDR, but on individual protein-level metric instead (PEP, E-value, etc.), despite the PSM scoring system remained strongly linked to TDC. In a similar trend, works from overlapping teams ([28] and [29]) have recently investigated the use Fisher method as a protein inference rule (rather than a scoring methodology) in a TDC context, and compared it with other approaches (product of PEP, best-peptide protein surrogate, two peptide rules, etc.).

5 Code availability

Implementing BH-FDR at PSM-, peptide- and protein-level is straightforward. First, if the scores of all the PSMs indicating a given peptide sequence are stored as a vector, `psm.scores`, then, the peptide p-value `pep.pval` and peptide score `pep.score` can be determined by applying the following `R` code:

```
library(Rmpfr) # to avoid roundings in p-values
psm.pvals <- mpfr(10**(-psm.scores/10), 128)
pep.pval <- 1-(1-min(psm.pvals))^length(psm.pvals)
pep.score <- -10*log10(pep.pval)
```

Then, the protein score `prot.score` and p-value `prot.pval` for a protein for which peptide scores are stored in a vector `pep.scores` can be computed using the following code:

```
pep.scores=sort(pep.scores, decreasing=T)
nb.pep=length(pep.scores)
pep.cumscore=cumsum(pep.scores)*log(10)/5
tmp.scores=rep(0,nb.pep)
for(j in 1:nb.pep){
   tmp.scores[j]=pchisq(pep.cumscore[j],2*j,lower.tail=F,log.p=T)
   tmp.scores[j]=tmp.scores[j]/(-0.1*log(10))
}
prot.score= max(tmp.scores)
prot.pval= 10**(-prot.score/10)
```

Once the Peptide scores and Protein scores are available alongside the PSM scores provided by the search engine, the BH procedure can simply be run by applying the `p.adjust() R` function (base function).

All these scores (and the BH procedure) are implemented in Proline software (http://www.profiproteomics.fr/proline/), written in Java/Scala, so that any proteomics data analyst can use them whatever their coding skills.

6 Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [53] partner repository with the dataset identifier PXD016669 and 10.6019/PXD016669.

7 Acknowledgments

This work was supported by grants from the French National Research Agency: ProFI project (ANR-10-INBS-08), GRAL project (ANR-10-LABX-49-01), Data@UGA (ANR-15-IDEX-02) and SYMER project (ANR-15-IDEX-02). The authors are grateful to EDyP platform engineers, who performed the various quality control analyses used in this work, and notably to Alexandra Kraut for the sample preparation, as well as the ProFI developers of Proline software (http://www.profiproteomics.fr/proline/), which was used to perform all the data analyses presented.

8 Author Contributions

Y.C. proposed the experimental methods and processed the data. C.B. developed the software environment to perform the experiments and implemented the proposed methodology. T.B. proposed the theoretical framework, conducted the study and drafted the manuscript. All authors contributed and approved to the final manuscript.

9 Competing Interests

The authors declare no competing interests.

References

[1] Thomas Burger. Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *Journal of Proteome Research*, 17(1):12–22, 2018.

[2] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[3] Yulia Danilova, Anastasia Voronkova, Pavel Sulimov, and Attila Kertész-Farkas. Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *Journal of Proteome Research*, 18(5):2354–2358, 2019.

[4] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.

[5] Kun He, Yan Fu, Wen-Feng Zeng, Lan Luo, Hao Chi, Chao Liu, Lai-Yun Qing, Rui-Xiang Sun, and Si-Min He. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv preprint arXiv:1501.00537*, 2015.

[6] Lev I Levitsky, Mark V Ivanov, Anna A Lobas, and Mikhail V Gorshkov. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *Journal of proteome research*, 16(2):393–397, 2016.

[7] Luca Bianco, Jennifer A Mead, and Conrad Bessant. Comparison of novel decoy database designs for optimizing protein identification searches using abrf sprg2006 standard ms/ms data sets. *Journal of proteome research*, 8(4):1782–1791, 2009.

[8] Uri Keich, Kaipo Tamura, and William Stafford Noble. An averaging strategy to reduce variability in target-decoy estimates of false discovery rate. *Journal of proteome research*, 18(2):585–593, 2019.

[9] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. False discovery rates in spectral identification. *BMC Bioinformatics*, 13(16):S2, Nov 2012.

[10] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(01):29–34, 2007.

[11] Bret Cooper. The problem with peptide presumption and low mascot scoring. *Journal of proteome research*, 10(3):1432–1435, 2011.

[12] Bret Cooper. The problem with peptide presumption and the downfall of target–decoy false discovery rates. *Analytical chemistry*, 84(22):9663–9667, 2012.

[13] Robert J Chalkley. When target–decoy false discovery rate estimations are inaccurate and how to spot instances. *Journal of proteome research*, 12(2):1062–1064, 2013.

[14] Elena Bonzon-Kulichenko, Fernando Garcia-Marques, Marco Trevisan-Herraz, and Jesús Vázquez. Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows. *Journal of proteome research*, 14(2):700–710, 2014.

[15] Iakes Ezkurdia, Jesús Vázquez, Alfonso Valencia, and Michael Tress. Analyzing the first drafts of the human proteome. *Journal of proteome research*, 13(8):3854–3855, 2014.

[16] Iakes Ezkurdia, Enrique Calvo, Angela Del Pozo, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. The potential clinical impact of the release of two drafts of the human proteome. *Expert review of proteomics*, 12(6):579–593, 2015.

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

[18] Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, and Thomas Burger. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying fdr control in quantitative experiments. *Proteomics*, 16(1):29–32, 2016.

[19] Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[20] Kun He, Mengjie Li, Yan Fu, Fuzhou Gong, and Xiaoming Sun. A direct approach to false discovery rates by decoy permutations. *arXiv preprint arXiv:1804.08222*, 2018.

[21] Aaron A Klammer, Christopher Y Park, and William Stafford Noble. Statistical calibration of the sequest xcorr function. *Journal of proteome research*, 8(4):2106–2113, 2009.

[22] J. Jeffry Howbert and William Stafford Noble. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular & Cellular Proteomics*, 13(9):2467–2479, 2014.

[23] Kenneth Verheggen, Helge Ræder, Frode S Berven, Lennart Martens, Harald Barsnes, and Marc Vaudel. Anatomy and evolution of database search engines – a central component of mass spectrometry based proteomic workflows. *Mass spectrometry reviews*, 2017.

[24] Viktor Granholm, William Stafford Noble, and Lukas Käll. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of proteome research*, 10(5):2671–2678, 2011.

[25] Viktor Granholm, José Fernández Navarro, William Stafford Noble, and Lukas Käll. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of proteomics*, 80:123–131, 2013.

[26] Victor Spirin, Alexander Shpunt, Jan Seebacher, Marc Gentzel, Andrej Shevchenko, Steven Gygi, and Shamil Sunyaev. Assigning spectrum-specific p-values to protein identifications by mass spectrometry. *Bioinformatics*, 27(8):1128–1134, 2011.

[27] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

[28] Matthew The, Michael J. MacCoss, William S. Noble, and Lukas Käll. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *Journal of The American Society for Mass Spectrometry*, 27(11):1719–1727, Nov 2016.

[29] Matthew The, Fredrik Edfors, Yasset Perez-Riverol, Samuel H Payne, Michael R Hoopmann, Magnus Palmblad, Björn Forsström, and Lukas Käll. A protein standard that emulates homology for the characterization of protein inference algorithms. *Journal of proteome research*, 17(5):1879–1886, 2018.

[30] Ting Huang, Jingjing Wang, Weichuan Yu, and Zengyou He. Protein inference: a review. *Briefings in bioinformatics*, 13(5):586–614, 2012.

[31] Oliver Serang and William Noble. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface*, 5(1):3, 2012.

[32] Gelio Alves and Yi-Kuo Yu. Mass spectrometry-based protein identification with accurate statistical significance assignment. *Bioinformatics*, 31(5):699–706, 2014.

[33] Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular & Cellular Proteomics*, 14(9):2394–2404, 2015.

[34] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[35] Pratik Jagtap, Jill Goslinga, Joel A Kooren, Thomas McGowan, Matthew S Wroblewski, Sean L Seymour, and Timothy J Griffin. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13(8):1352–1357, 2013.

[36] Kristen Emery, Syamand Hasam, William Stafford Noble, and Uri Keich. Multiple competition based fdr control. *arXiv preprint arXiv:1907.01458*, 2019.

[37] Anna Salvetti, Yohann Couté, Alberto Epstein, Loredana Arata, Alexandra Kraut, Vincent Navratil, Philippe Bouvet, and Anna Greco. Nuclear functions of nucleolin through global proteomics and interactomic approaches. *Journal of proteome research*, 15(5):1659–1669, 2016.

[38] Sangtae Kim and Pavel A Pevzner. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[39] David Fenyö and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003.

[40] David Shteynberg, Eric W Deutsch, Henry Lam, Jimmy K Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L Moritz, Ruedi Aebersold, and Alexey I Nesvizhskii. iprophet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & Cellular Proteomics*, 10(12):M111–007690, 2011.

[41] John S Cottrell and David M Creasy. Response to: the problem with peptide presumption and low mascot scoring. *Journal of proteome research*, 10(11):5272–5273, 2011.

[42] David L Tabb. The sequest family tree. *Journal of The American Society for Mass Spectrometry*, 26(11):1814–1819, 2015.

[43] Niklaas Colaert, Sven Degroeve, Kenny Helsens, and Lennart Martens. Analysis of the resolution limitations of peptide identification algorithms. *Journal of proteome research*, 10(12):5555–5561, 2011.

[44] Marc Vaudel, Julia M Burkhart, Daniela Breiter, René P Zahedi, Albert Sickmann, and Lennart Martens. A complex standard for protein identification, designed by evolution. *Journal of proteome research*, 11(10):5065–5071, 2012.

[45] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of proteome research*, 7(01):40–44, 2007.

[46] Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[47] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[48] Joachim Hartung. A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(7):849–855, 1999.

[49] James T Kost and Michael P McDermott. Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190, 2002.

[50] C Maria Keet. Open world assumption. In *Encyclopedia of Systems Biology*, pages 1567–1567. Springer, 2013.

[51] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life. *Studies in social psychology in World War II*, 1, 1949.

[52] AM Mathai. On linear combinations of independent exponential variables. *Communications in Statistics-Theory and Methods*, 12(6):625–632, 1983.

[53] Yasset Perez-Riverol, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J Kundu, Avinash Inuganti, Johannes Griss, Gerhard Mayer, Martin Eisenacher, et al. The pride database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*, 47(D1):D442–D450, 2018.

[54] Jurgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of proteome research*, 10(4):1794–1805, 2011.

Supplementary Information

A Supplementary figures and tables

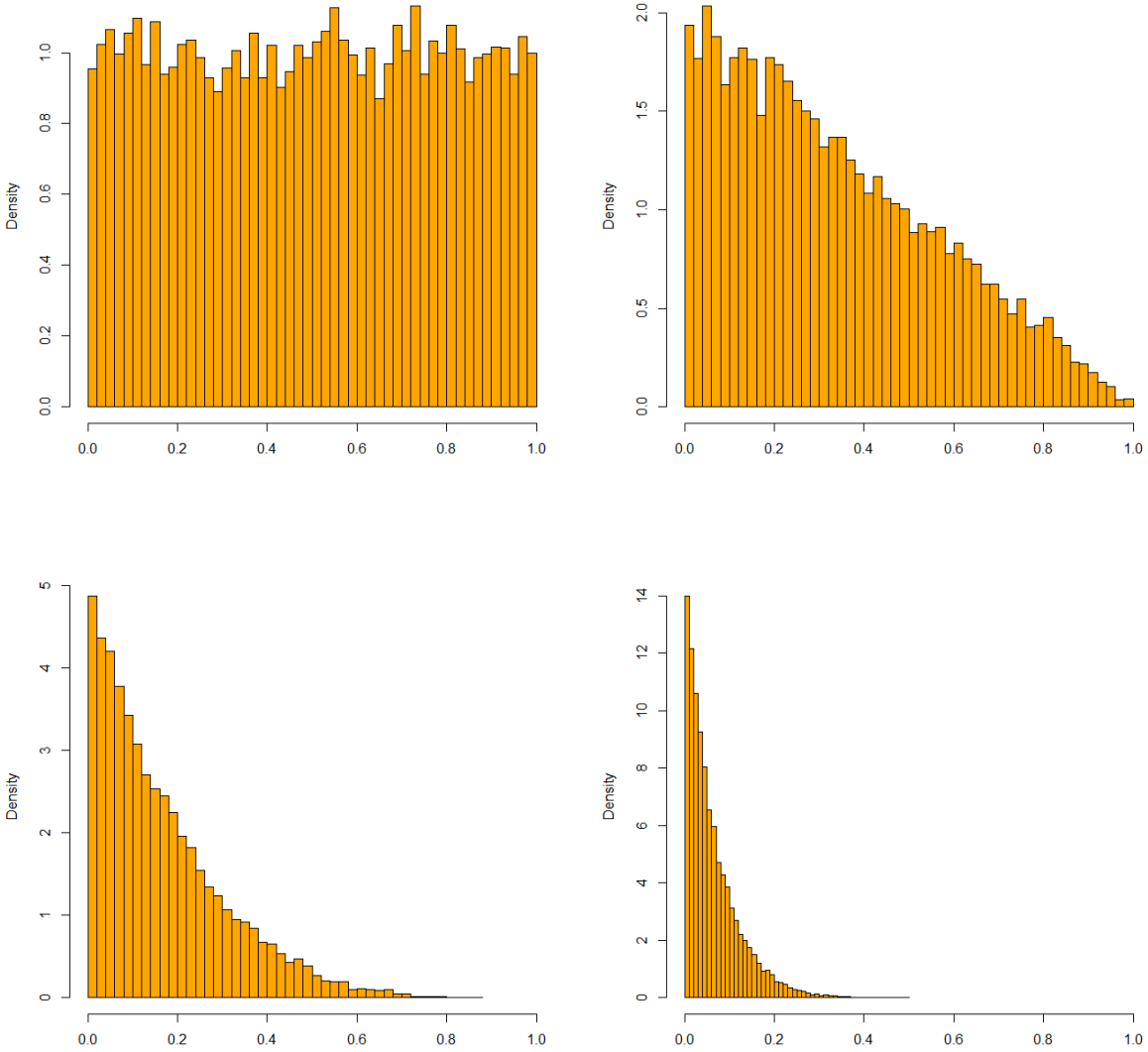
Supplementary Table 1: Quantitative summary of the results depicted in Figure 1, completed with the number of validated PSMs.

QEx

	Queries	BH Low-Low PSM target	BH Low-Low Score	BH High-Low PSM target	BH High-Low Score	BH Low-High PSM target	BH Low-High Score	BH High-High PSM target	BH High-High Score	TDC Low-Low PSM target	TDC Low-Low Score	TDC High-Low PSM target	TDC High-Low Score	TDC Low-High PSM target	TDC Low-High Score	TDC High-High PSM target	TDC High-High Score
R1	26604	12135	23,21	10954	21,69	11623	21,42	10551	20,71	12758	19,23	11951	11,01	13363	8,83	12209	1,61
R2	26520	11880	23,28	10863	21,78	11368	21,52	10446	20,77	12546	19,41	11948	11,04	13251	8,59	12224	1,58
R3	26103	12364	23,07	10830	21,74	11863	21,41	10463	20,71	12988	19,15	11802	11,04	13629	8,48	12021	1,52
R4	24542	12680	22,68	11543	21,44	12029	21,28	11045	20,71	13284	18,88	12589	10,62	13828	8,69	12819	1,26
R5	25759	11756	23,24	10696	21,76	11224	21,53	10275	20,75	12309	19,64	11729	10,96	12977	9,01	12039	1,2
R6	25900	12379	23,01	11043	21,66	11853	21,39	10658	20,67	12994	19,09	12049	10,64	13519	9,09	12309	1,11
R7	25579	12121	23,07	10867	21,69	11592	21,47	10488	20,70	12703	19,32	11941	10,37	13276	9,05	12134	1,42
R8	25500	12095	23,05	10825	21,65	11624	21,41	10468	20,67	12659	19,26	11816	10,84	13274	8,91	12019	1,39
R9	23726	10415	23,38	9509	21,89	9975	21,69	9128	20,84	10852	20,73	10439	11,66	11510	10,14	10705	2,37
R10	23819	10428	23,43	9462	21,94	9968	21,71	9145	20,82	10856	20,61	10389	12,21	11507	10,45	10762	2,13

Velos

	Queries	BH Low-Low PSM target	BH Low-Low Score	BH High-Low PSM target	BH High-Low Score	BH Low-High PSM target	BH Low-High Score	BH High-High PSM target	BH High-High Score	TDC Low-Low PSM target	TDC Low-Low Score	TDC High-Low PSM target	TDC High-Low Score	TDC Low-High PSM target	TDC Low-High Score	TDC High-High PSM target	TDC High-High Score
R1	18637	10784	22,17	10139	21,15	·	·	·	·	11202	19,65	11169	11,32	·	·	·	·
R2	18722	10835	22,18	10210	21,22	·	·	·	·	11179	19,96	11244	11,26	·	·	·	·
R3	18511	10882	22,1	10167	21,15	·	·	·	·	11226	19,48	11127	11,24	·	·	·	·
R4	18242	10848	22,06	10148	21,15	·	·	·	·	11121	19,35	11206	10,52	·	·	·	·
R5	18302	10715	22,13	10048	21,19	·	·	·	·	11166	19,25	11069	11,28	·	·	·	·
R6	18352	10746	22,13	10021	21,2	·	·	·	·	11182	19,54	11084	10,93	·	·	·	·
R7	18402	10794	22,13	10086	21,21	·	·	·	·	10944	20,23	11149	10,98	·	·	·	·
R8	17799	10665	22,04	10008	21,17	·	·	·	·	11725	20,91	11020	10,58	·	·	·	·
R9	19041	11539	22,05	10727	21,17	·	·	·	·	11790	20,07	11658	11,96	·	·	·	·
R10	19035	11499	22	10676	21,15	·	·	·	·	11084	20,22	11619	11,83	·	·	·	·



Supplementary Figure 1: Histogram of 15,000 simulated p-values, after taking the minimum out of N uniformly distributed samples, with $N = 1$ (upper left), $N = 2$ (upper right), $N = 5$ (lower left) and $N = 15$ (lower right). Clearly, the uniform distribution is lost because of the minimum operator. The rationale behind is intuitive: by taking the minimum p-values, one promotes small p-values with respect to large p-values, so that from a distribution with used to be uniform, one ends up with another distribution with is shifted to the left.

Supplementary Table 2: Replicate-wise details of the validation results summarized in Table 1.

		No validation rule								
		PSMs			Peptides			Proteins		
		#Targets	#Decoys	min Score	#Targets	#Decoys	min Score	#Targets	#Decoys	min Score
Fivefold shuffle	R1	12319	744	0	9293	734	0	1482	712	0
	R2	12347	840	0	9286	821	0	1462	795	0
	R3	12149	849	0.01	9254	829	0.01	1451	803	0.01
	R4	12843	714	0.01	9898	684	0.01	1427	667	0.01
	R5	12092	826	0	9465	806	0	1470	775	0
	R6	12332	837	0	9422	821	0	1435	784	0
	R7	12201	852	0	9408	833	0	1476	807	0
	R8	12095	830	0	9418	808	0	1477	780	0
	R9	10908	935	0	9026	914	0	1503	869	0
	R10	10916	946	0	9043	922	0	1482	868	0
	Average	12020.2	837.3	0.002	9351.3	817.2	0.002	1466.5	786	0.002

		1% FDR at PSM level (alone)								
		PSMs			Peptides			Proteins		
		#Targets	#Decoys	min Score	#Targets	#Decoys	min Score	#Targets	#Decoys	min Score
Fivefold shuffle	R1	10523	9	20.94	8175	9	18.9	1305	9	21.04
	R2	10420	8	21.02	8115	8	20.17	1293	8	21.78
	R3	10424	11	20.96	8151	11	19.84	1286	11	20.96
	R4	11012	29	20.9	8723	23	19.61	1292	22	21.01
	R5	10239	11	21.02	8250	11	19.96	1304	11	21.23
	R6	10624	10	20.93	8309	10	19.55	1288	10	20.95
	R7	10458	10	20.98	8272	10	20.23	1305	9	21.08
	R8	10440	10	20.95	8327	10	19.73	1318	10	20.95
	R9	9094	9	21.15	7723	9	18.53	1290	9	21.54
	R10	9102	8	21.15	7758	8	20.24	1296	8	21.33
	Average	10233.6	11.5	21	8180.3	10.9	19.676	1297.7	10.7	21.187

		1% FDR at peptide level (alone)								
		PSMs			Peptides			Proteins		
		#Targets	#Decoys	min Score	#Targets	#Decoys	min Score	#Targets	#Decoys	min Score
Fivefold shuffle	R1	11073	8	0.04	8151	8	20.92	1304	8	21.04
	R2	11032	8	0.02	8089	8	20.97	1293	8	21.78
	R3	10926	11	0.02	8126	11	20.95	1286	11	20.96
	R4	11537	32	0.02	8705	23	20.85	1292	22	21.01
	R5	10755	10	0	8232	10	20.96	1304	10	21.23
	R6	11117	9	0.01	8286	9	20.92	1287	9	20.95
	R7	10975	9	0.02	8251	9	20.95	1305	8	21.08
	R8	10909	9	0.01	8307	9	20.9	1318	9	20.95
	R9	9484	8	0.05	7698	8	21.12	1290	8	21.54
	R10	9554	7	0.04	7750	7	21.09	1300	7	21.09
	Average	10736.2	11.1	0.023	8159.5	10.2	20.963	1297.9	10	21.163

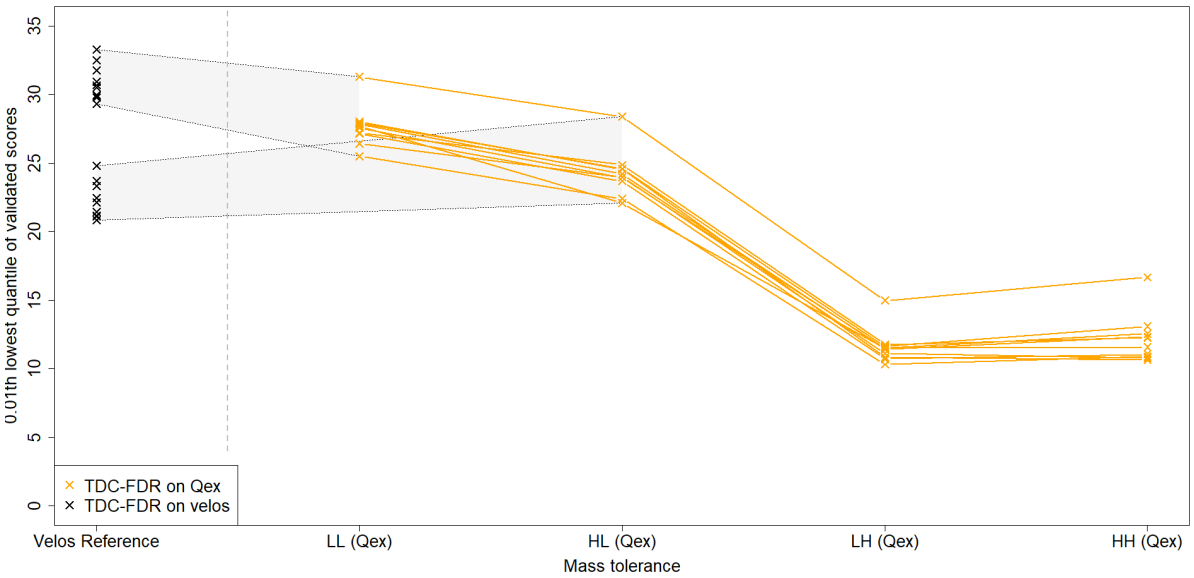
		1% FDR at protein level (alone)								
		PSMs			Peptides			Proteins		
		#Targets	#Decoys	min Score	#Targets	#Decoys	min Score	#Targets	#Decoys	min Score
Fivefold shuffle	R1	12123	7	0.04	9099	7	0.08	1300	7	22.54
	R2	12163	8	0.02	9103	8	0.03	1293	7	22.7
	R3	11964	14	0.01	9070	14	0.01	1282	11	22.45
	R4	12702	32	0.02	9760	21	0.04	1293	18	22.25
	R5	11912	13	0	9288	12	0.01	1299	9	22.35
	R6	12170	16	0.01	9259	14	0.01	1284	9	22.44
	R7	12019	13	0	9227	12	0	1306	8	22.48
	R8	11924	12	0	9247	11	0	1316	8	22.42
	R9	10679	12	0.02	8801	10	0.02	1293	7	22.99
	R10	10710	11	0.01	8840	11	0.01	1290	8	22.72
	Average	11836.6	13.8	0.013	9169.4	12	0.021	1295.6	9.2	22.534

B Downfall of Andromeda scores

The TDC lack of stability already observed with Mascot (see Figure 1) can also be illustrated with Andromeda, even though the conclusions should be cautiously interpreted for the following reasons: (1) Andromeda code is not accessible, so that it is not possible to check whether the provided scores are individual or if they should be considered as contextualized scores, because of the "*fixed additive component*" which accounts for peptide dependences, as described in [54]; (2) Depending on the version, the TDC procedure is by default applied on posterior probabilities or on delta scores, both of us being contextualized scores; (3) Due to the specific relationship between the posterior probabilities and the scores of long peptides (see Methods 4.3) the minimum observed Andromeda score within the validated list is almost always near zero, so that focusing on the variations of cut-off score is not informative.

Concerning the first point, it is more a state of affair than an issue to solve, which will only make Andromeda more or less adapted to evaluate the TDC procedure. This is notably the reason why we do not compute and display the BH cut-off scores: it is impossible to make sure that Equation 1 is applied on the correct score. As for the second one, we have used the posterior probability, because it is the published method. Finally, concerning the last one, we simply have to find a statistics other than the minimum score which depicts the quality of the borderline validated peptides. We have decided to rely on the lowest percentile of the score distribution. This interprets as following: a value of x indicates that the 1% lowest PSM scores are distributed between 0 and x. As detailed in Methods 4.3, as the lowest PSM scores are expected to remain of constant quality, it makes an interesting statistics to illustrate a potential lack of stability in the TDC.

The results are depicted on Supplementary Figure 2. Although the first percentile can be expected to be more stable than the minimum value, one observes an important instability, both for each tolerance tuning taken individually, and across the tolerance tunings. However, it appears that contrarily to Mascot, X!tandem and MS-GF+, the effect of the fragment mass tolerance tuning is much more important than that of the precursor one. Moreover, the mapping between the Velos and Qex analyses is not as good as with Mascot. However, in the LL setting, switching from a Velos to a Qex leads to a stringency loss, while on the contrary, it leads to a stringency increment in the HL setting. Considered together, these observations confirms the lack of of stability of TDC procedure.



Supplementary Figure 2: Same figure as Figure 1, yet with Andromeda search engine instead of Mascot.