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

Yohann Couté,[†,‡] Christophe Bruley,[†,‡] and Thomas Burger[*,†]

†*Univ. Grenoble Alpes, CNRS, CEA, INSERM, IRIG, BGE, F-38000 Grenoble, France*

‡*Equal contribution*

E-mail: thomas.burger@cea.fr

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.*

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[1]. The greater the resemblance between the experimental spectrum and the theoretical spectrum, the higher the matching score. This methodology has been widely adopted, and it has soon been accepted that it could lead to false positive identifications[2]. 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[3] 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[4]. 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 (present in the "target" database) and to organize a competition between target and decoy assignments. Under the so-called Equal Chance Assumption (or ECA, stating that target mismatches and decoy matches are equally likely[4]), it is possi-

ble, 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 ECA validity, *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[5]. 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[6]; *(ii)* In theory, the larger the decoy database, the more precise the mismatch score distribution[7,8]; *(iii)* The decoys must respect the cleavage sites[9] 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[10]; *(v)* Decoy counting also has an influence[8]. In addition to these restrictions, numerous parameters have been reported and discussed to control their relative importance[11]. This extensive body of literature has notably contributed to installing the competition step of TDC as essential, and today, target-decoy searches without competition[12,13] are scarcely ever reported. Despite TDC wide acceptance, a series of letters from Bret Cooper[14,15] initiated a controversy regarding the observed downfall of TDC validation levels with data produced by high resolution mass spectrometers. 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 the target database could be identified). As this stance contradicted both empirical and theoretical evidence, few articles were published arguing

against this view[16,17] while others confirmed[18,19], 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[7,8] or by averaging a growing number of TDC-FDRs resulting from randomly generated decoy databases, in a boosting-like strategy[10]), 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). In this context, even if *asymptotically unbiased* (*i.e.*, no systematic difference between the FDP and the FDR) TDC could sometimes lead to inaccurate FDRs (*i.e.*, over-conservative or anti-conservative estimates) due to too high *variance* (*i.e.*, important stochastic fluctuations).

In this article, we 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[20] and its numerous variants[21] can be applied at spectrum, peptide and protein levels. This leads to an original and powerful framework that demonstrably controls the FDR without 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 identification results from MS-based proteomics can be simply and accurately validated by applying alternative strategies.

Experimental Section

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[22].

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).

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 below). 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). Proline software[23] 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 section. Precisions regarding the choice of the score (individualized vs. contextualized) as well as of alternative search engines can be found in Supporting Information 2.3 and 4.1.

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 prior knowledge[11].

To compute the EN FDRs[24] of Figure 3, we have relied on the empirical null model provided by a shuffled database that was not in competition with the target one[12,13] coupled to a boosting strategy[10], *i.e.* a procedure averaging the FDR estimates of multiple shuffle decoy databases (see Supporting Information 4.2 for detailed justifications). Concretely, we used 10 shuffled databases, each being of length equal to that of the database, and resulting from a shuffling procedure which respected the cleavage site to maintain the precursor mass distribution[9] (R code available in Supporting Information 1.1).

Results and Discussion

TDC and Benjamini-Hochberg procedures yield different FDRs

To counteract the downfall of TDC validation levels observed with modern MS instruments[14,15], we had the intuition that Benjamini-Hochberg (BH) approach to FDR control[20] could be an interesting alternative. Discrepancies between BH and TDC have already been reported[25,26], but direct quantitative comparisons in terms of bias (*i.e.*, systematic error) and variance (*i.e.*, lack of stability) are unfortunately impossible:[27] It would require to precisely know the FDP, whereas it is not accessible, even using a controlled dataset. Thus, to better grasp the possible differences of behavior between TDC and BH FDRs, we reproduced Cooper's experiment with the following extensions: First, both FDRs were computed simultaneously on the same datasets; Second, many experimental parameter settings were varied:

we used multiple analytical replicates, different instruments and various combinations of precursor and fragment mass tolerance sets.

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[28]: 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 Supporting Information 4.1). 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.*[25,26]. In theory, p-values always distribute uniformly across the [0,1] interval[3]. However, in practice, it is not always the case and then, the p-values are termed *ill-calibrated* or *miscalibrated*. As miscalibration can hamper BH estimate correctness, tools exists to check or improve the p-value calibration (see Supporting Information 1.2).

We applied both TDC and BH methodologies to results acquired with a Q-Exactive Plus instrument on the ten analytical replicates of *E. coli* lysate. MS and MS/MS spectra were acquired at relatively high resolutions (70,000 and 17,500 at m/z 200, respectively). We used Mascot to run a TDC search (see Experimental Section) 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 analysis. In parallel, we ran target-only searches using the same parameters. Scores were then converted into p-values using Eq. 1 and the calibration of the resulting p-values was assessed as reported in Supporting Information 2.1. Finally, the classical BH step-up procedure for p-value adjustment was run (see Data and

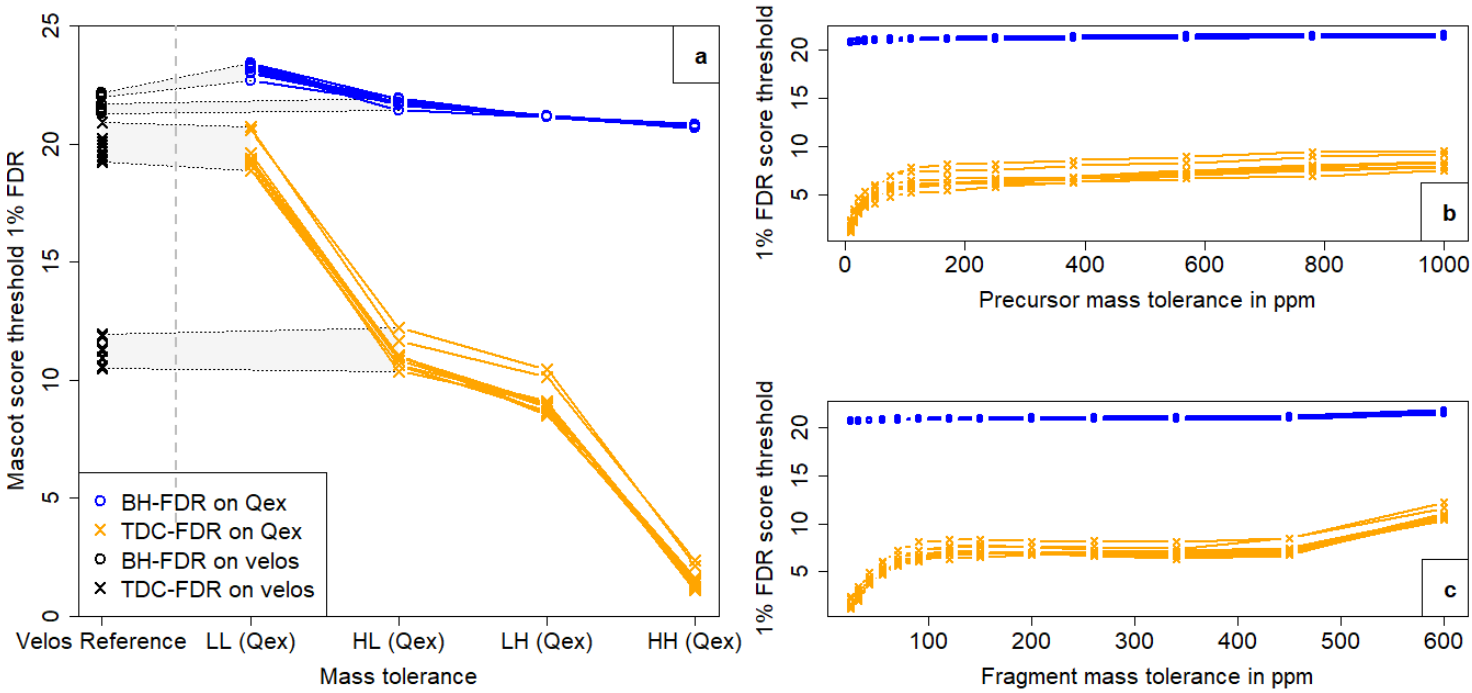

Figure 1: Score thresholds obtained when applying TDC (Orange) and BH (Blue) filtering at an FDR of 1%, as a function of the search engine mass tolerance parameters, for 10 samples analyzed with a Q-Exactive Plus (Qex) instrument. (**a**) Precursor and fragment mass tolerances are tuned to the LL, LH, 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).

code availability).

Figure 1A shows the score thresholds obtained at an FDR of 1% as a function of the mass tolerance combinations (see Supporting Information 2.2 for numerical values). At first glance, the downfall of validation score is obvious with TDC, while almost immaterial with BH. To avoid drawing sweeping conclusions, we made complementary experiments. First, to better capture the influence of mass tolerance tuning, 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 trends obtained support the one depicted on Figure 1A. Moreover, they confirm that tolerances at precursor and fragment levels have more influence on the threshold scores resulting from TDC than from BH.

Second, to confirm that results obtained by lowering mass tolerance in search parameters mimicks results obtained with lower resolution instruments, we analyzed another batch of ten analytical replicates of the *E. coli* lysate using a LTQ-Orbitrap Velos Pro with 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 results from Q-Exactive and LTQ-Orbitrap data obtained using the same thresholds is excellent. This 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 tolerances, produces roughly equivalent outputs.

Therefore, we can interpret the different database search sets as surrogates for the recent evolution of instrumental capabilities: when TDC was first applied to data from low-resolution instruments, TDC and BH led to roughly similar results in terms of score cut-off to reach 1% FDR. Since then, the resolution of mass spectrometry instruments has progressively increased, and now TDC and BH diverge considerably on the same lists of putative PSMs, so that at least one of them (but possibly both) yields incorrect FDR control.

BH is more stable than TDC

From Figure 1, BH appears more stable than TDC, both within and between experimental settings: First, for each tuning taken individually, the TDC threshold was less stable than its BH counterpart, as the set of ten cut-off scores was more dispersed with TDC. Second, depending on the mass tolerances applied during database searching, the TDC threshold on the Mascot score varied from 1.11 to 20.73 (respectively with HH and LL settings), whereas its BH counterpart was more stable (between 20.67 and 23.43).

A part of the TDC's instability (relatively to BH) can be explained by the random nature of decoy sequence generation[10], regardless of the search engine used. However, at first glance, there is no reason 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), as Cooper first did[14]. Unfortunately, he then reported[15] similar pitfalls (at least for the precursor tolerance parameter) with X!tandem and MS-GF+ (formerly referred to as MS-GFDB). In addition, we discovered similar effects with Andromeda (Maxquant environment, see Supporting Information 2.3). Finally, our experiments have proved that not only the set precursor mass tolerance, but also, the set fragment mass tolerance are involved. For all these reasons, we believe that it is legitimate to point the TDC intrinsic lack of stability.

TDC instability can lead to anti-conservative FDRs

Naturally, more stable FDRs should be preferred, yet less stable ones are not necessarily untrustworthy: differently parametrized searches yielding different putative PSM lists, different albeit correct FDRs would make sense. However, the spread between BH and TDC cut-offs in the HH setting is alarming, as the PSMs identified in the target database before validation were the same for both. The very low cut-off scores obtained with TDC while using the HH setting led us to question TDC procedure: Even if TDC is assumed unbiased, can its instability result in FDRs that are sometimes conservative (*e.g.* LL cases) and sometimes

not (*e.g.* HH cases)? This context-dependent anti-conservativeness would clearly hamper a reliable use of TDC approach, as false discoveries would no longer be controlled.

To answer this question, we relied on the following rationale of FDR control: When one filters a PSM list at 1% FDR, it does not mean that we allow 1% of additional poor matches in the result. On the contrary, it means that, despite all the validated PSMs apparently depicting matches of sufficient quality, 1% of them are spurious (randomly distributed over the full range of scores, not necessarily those with the lowest scores). Although counter-intuitive, this property has already been empirically confirmed[29]. Therefore, examining the list of PSMs with low scores validated by TDC only using the HH setting is insightful to assess its conservativeness. Concretely, as in the results from the first analytical replicate, 12209 PSMs were validated after TDC, the number of expected mismatches at 1% FDR is 122. Thus, we randomly selected 150 of these PSMs with a Mascot score < 10. The matches between theoretical and experimental spectra were obviously of too poor quality to yield confident identifications for the vast majority of them (see Supporting Information 2.4). Therefore, the [1.61; 10[interval (containing only 635 PSMs, i.e. 5.53% of 12209) already leads to more mismatches than expected in the entire dataset. This is incompatible with 1% FDR validation, as random misidentification depicting apparently correct matches are also expectable at higher scores[29]. In other words, on this dataset using the HH setting, the TDC FDR did not conservatively estimate the FDP.

These observations confirm those reported by Cooper. They are not sufficient to conclude on an intrinsic bias of TDC. However, they confidently show the downfall of validation cut-off scores comes along with, at least, an increased risk of anti-conservative FDR.

Downfall mechanistic explanation

The above results contrast with the theoretical guarantees that have accompanied TDC[7,30,31]. However, none of them account for the application of preliminary filters which reduce the number of decoy competitors, while Cooper's controversy is rooted in such filters. This

may explain the discrepancy: 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. Therefore, the ECA may no longer apply and the FDRs may be underestimated. This can be experimentally observed while looking at the distributions of the PSMs identified in the decoy database using LL and HH settings. Indeed, the number of decoy identifications as well as their corresponding scores are strongly decreased in HH compared to LL settings (see Figure 2). This very low number of identified decoy PSMs in HH settings cannot allow for an accurate FDR estimation.

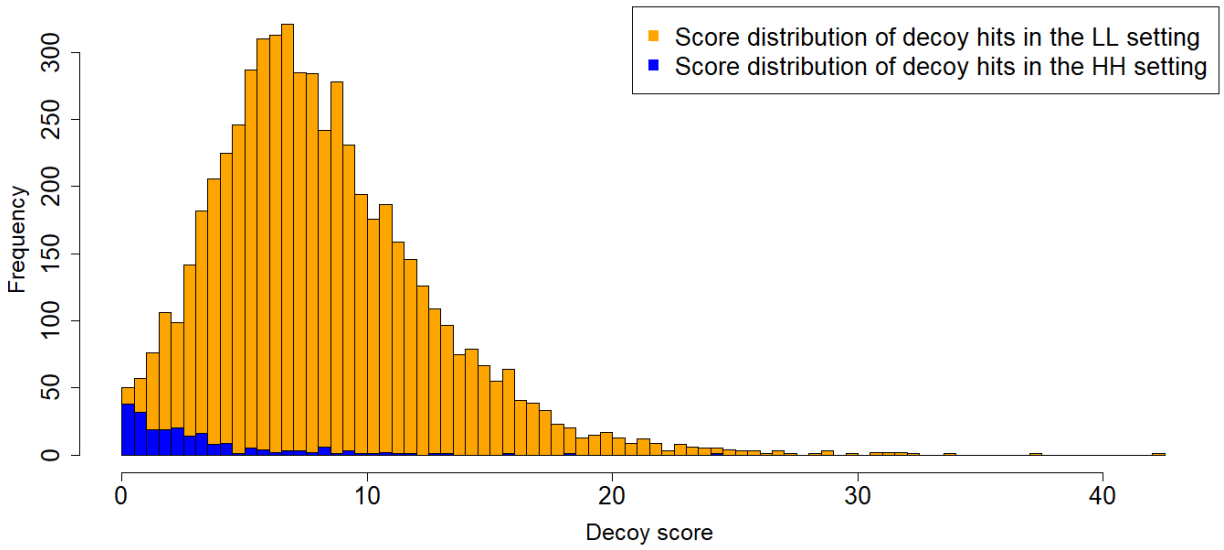

Figure 2: Distributions of scores of PSMs identified in the decoy database using the LL (orange) and HH (blue) settings. Presented data correspond to replicate 1, Qex analysis. After concomitant searches in target and decoy databases using LL and HH parameters, the scores of the PSMs identified only in the decoy database, and without FDR filtering, were collected and represented as histograms (x-axis: score, bin 0.5; y-axis: frequency).

As TDC seems to be inaccurate when the number of decoy challengers is too small, one straightforward solution would be to enlarge the database in proportion. Although the link

between mass tolerance and decoy database size can be efficiently exploited to limit the computational cost of repeated searches[32], it was, unfortunately, inefficient in our case (see Supporting Information 4.3).

Decoy-based empirical null estimation and BH FDRs are consistent

An alternative way to investigate the above mechanistic intuition is to build a strategy in which the quality of decoy matches is preserved, regardless of the stringency of the database search preliminary filters; and then to evaluate how it estimates the FDR. Making false discoveries look alike true ones is at the heart of the statistical theory of *empirical null estimation*[24]. Even though this theory has already been applied in various proteomics contexts (where it was referred to as *entrapment methods*[27,33–37]), it provided us a nice framework to elaborate on. Surprisingly, we found that a fair (but still unstable) empirical estimate of the null distribution could be obtained by performing a separate decoy search, without competition[12,13]. However, to cope for the lack of stability, we averaged 10 repeated estimations, each based on a different shuffled database[10] (see Experimental Section and Supporting Information 4.2). For this reason, for each of the 10 sample replicates of *E. coli* lysate analysed with the Q-Exactive Plus, Figure 3 shows the 10 empirical null (EN) FDRs corresponding to the cut-off scores obtained using BH and TDC 1% FDR filtering (HH setting).

The difference is striking: although BH thresholds (Mascot scores between 20.67 and 20.84) produced EN 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) led to 10-fold larger EN FDRs (between 7.72% and 11.71%, with an average ≈9.1%). This result is insightful for three reasons: *(i)* The fact that two orthogonal methods to compute an FDR (namely BH and EN) provided concurring results is an evidence supporting their correctness; *(ii)* it confirms that TDC can lead to considerable FDR under-estimations if used inappropriately; *(iii)* 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

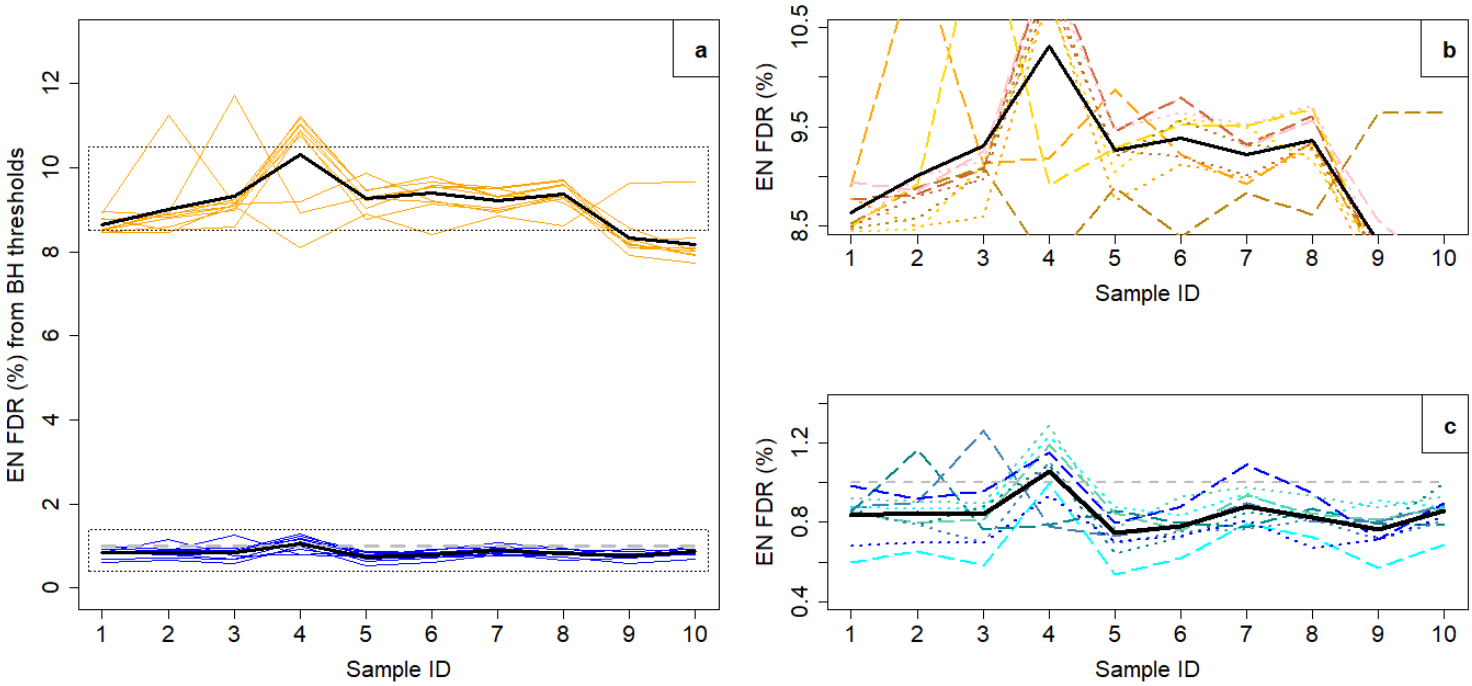

Figure 3: (**a**) Empirical Null (EN) FDRs computed for each replicate matched against 10 randomly generated (shuffled) decoy databases, according to its 1% FDR validation score threshold (gray dash line), computed by applying BH (blue) and TDC (orange) methodologies; The black continuous lines depict the average of the 10 EN 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.

to cope for the cut-off downfall and to recover coherent FDRs.

Practical comparison of EN and BH approaches

To summarize, the implemented EN approach essentially amounts to averaging multiple target-decoy searches without competition, and can be viewed as an improvement of a 12 years old method[12,13] well-fitted in the proteomics landscape, yet outmoded. On the other hand, the BH approach is mainly theoretically motivated, and albeit twice older, is scarcely used in proteomics. As both provide concurring estimates, the one to promote mainly depends on their respective conditions of applicability and ease of use.

Concerning BH: If one compares on Figure 3 the EN FDRs derived from the BH-

thresholds with the expectation (i.e. 1%), BH appears to slightly overestimate. This result is probably due to the already described over-conservative property of the BH estimator[38]. Moreover, as previously mentioned, BH requires checking first that the p-values are not ill-calibrated. Fortunately, many methods can be used to limit BH over-conservativeness and applying the best one can be done concomitantly with the calibration assessment, by means of a simple visual tool[21] (see CP4P description in Supporting Information 1.2). Finally, computing BH FDR is extremely fast and does not require any decoy database.

Concerning EN: Figure 3 shows a high dependency on the different decoy 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% around 1% EN FDR, versus 5.82% around 10% EN 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 explains why averaging multiple FDRs with different randomly generated decoy searches is necessary, despite its additional computational and practical complexity. However, depending on the complexity of the experiment, the precise number of searches needed to stabilize the FDR cannot be estimated and will require manual trials. Moreover, if the experimental design requires iterative filtering of the database search (*e.g.* multiple-pass identifications[39]), one risks to end-up with the same anti-conservativeness issue as with the competition step, so that additional cautiousness is mandatory. Therefore, compared to the EN strategy, the BH procedure is appealing for its simplicity and stability.

FDR control at peptide level 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[40,41]. 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 Supporting Information 3.1). 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[36,37,42], 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[35–37,43]. In this setting, it makes sense to define the peptide score as equal to the best PSM score[35]. 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, see Supporting Figure 1). Fortunately, according to the following proposition, it is possible to modify the best-PSM score by applying a formula akin to Šidák correction[44] 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 Supporting Information 3.2.

Therefore, (See Supporting Information 3.2 and 3.3 for the full derivations), the **peptide**

p-value $p_i^\diamond$ and **peptide score** $S_i^\diamond$ of peptide sequence $\mathbf{seq}_i$ can be defined as:

$$p_i^\diamond = 1 - \left(1 - 10^{-\frac{\max_q (S_{iq})}{10}}\right)^Q \quad \text{and} \quad S_i^\diamond = -10 \cdot \log_{10}(p_i^\diamond) \tag{2}$$

FDR control at protein level using the BH procedure

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[32,36,37,42] and the full derivation is presented in Supporting Information 3.4. 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 Supporting Information 3.5); (ii) the impact of the poorly conclusive peptide-level evidence in an open-world assumption context (see Supporting Information 3.6). 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:

$$p_\pi^\star = \min_{A \in 2^{\{1,\ldots,K\}}} \left[\int_{0.2\ln(10)\cdot\sum_{k\in A} S_k^\diamond}^{\infty} f_{2\cdot|A|}(x)dx \right] \tag{3}$$

and the **protein score** as $S_\pi^\star = -10\log_{10}(p_\pi^\star)$, 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. Although possibly obscure at first glance, $p_\pi^\star$ simply interprets as the p-value resulting from Fisher combined probability test applied to the subset of peptides which best explains the protein (see Supporting Information 3.4).

Experimental assessment of BH FDR at peptide and protein levels

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 Experimental Section). 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 Supporting Table 2). Although the corresponding proportions 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.

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).

	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

Conclusions

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[12,13], the instability induced by the random generation of decoy sequences remains[10]. 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; or with pipeline modifications that change the number of target and decoy candidates).

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 Data & code availability) 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 calibration necessary to 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[4] has stimulated the field of theoretical biostatistics and led to the idea that simulating null tests from the data (termed *knockoffs* instead of decoys) could produce efficient alternatives to BH procedures, which demonstrably control the FDR[30,31,45]. Transferring these theoretical results into biostatistics routines that can be applied on a daily basis still requires some investigation[46,47]. However, they will hopefully contribute to computational proteomics in the future, as an example of an interdisciplinary virtuous circle.

Data and 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 also implemented in Proline software[23], written in Java/Scala, so that any proteomics data analyst can use them whatever their coding skills. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[48] partner repository with the dataset identifier PXD016669 and 10.6019/PXD016669 (https://www.ebi.ac.uk/pride/, Username: reviewer43825@ebi.ac.uk, Password: eA7lMPa7). A portable version of Proline software can be downloaded at ftp://ftp.cea.fr/pub/edyp/Proline/BH-FDR/. For each of the ten *E. Coli* replicates, it contains the search results from Mascot after the database search in the target and the fivefold decoy databases.

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 and SYMER projects (ANR-15-IDEX-02) as well as MIAI @ Grenoble 718 Alpes (ANR-19-P3IA-0003). 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.

Author contributions and competing interests

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 to the manuscript and approved its final version. The authors declare no competing interests.

References

(1) Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods* **2004**, *1*, 195–202.

(2) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L. Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of proteome research* **2004**, *3*, 1082–1085.

(3) Burger, T. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *Journal of Proteome Research* **2018**, *17*, 12–22.

(4) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **2007**, *4*, 207.

(5) Danilova, Y.; Voronkova, A.; Sulimov, P.; Kertész-Farkas, A. Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification. *Journal of Proteome Research* **2019**, *18*, 2354–2358.

(6) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry* **2011**, *22*, 1111–1120.

(7) He, K.; Fu, Y.; Zeng, W.-F.; Luo, L.; Chi, H.; Liu, C.; Qing, L.-Y.; Sun, R.-X.; He, S.-M. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv preprint arXiv:1501.00537* **2015**,

(8) Levitsky, L. I.; Ivanov, M. V.; Lobas, A. A.; Gorshkov, M. V. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *Journal of proteome research* **2016**, *16*, 393–397.

(9) Bianco, L.; Mead, J. A.; Bessant, C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *Journal of proteome research* **2009**, *8*, 1782–1791.

(10) Keich, U.; Tamura, K.; Noble, W. S. An averaging strategy to reduce variability in target-decoy estimates of false discovery rate. *Journal of proteome research* **2019**, *18*, 585–593.

(11) Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics* **2012**, *13*, S2.

(12) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research* **2007**, *7*, 29–34.

(13) Martínez-Bartolomé, S.; Navarro, P.; Martín-Maroto, F.; López-Ferrer, D.; Ramos-Fernández, A.; Villar, M.; García-Ruiz, J. P.; Vázquez, J. Properties of average score distributions of SEQUEST: the probability ratio method. *Molecular & Cellular Proteomics* **2008**, *7*, 1135–1145.

(14) Cooper, B. The problem with peptide presumption and low Mascot scoring. *Journal of proteome research* **2011**, *10*, 1432–1435.

(15) Cooper, B. The problem with peptide presumption and the downfall of target–decoy false discovery rates. *Analytical chemistry* **2012**, *84*, 9663–9667.

(16) Chalkley, R. J. When target–decoy false discovery rate estimations are inaccurate and how to spot instances. *Journal of proteome research* **2013**, *12*, 1062–1064.

(17) Bonzon-Kulichenko, E.; Garcia-Marques, F.; Trevisan-Herraz, M.; Vázquez, J. Revisiting peptide identification by high-accuracy mass spectrometry: problems associated with the use of narrow mass precursor windows. *Journal of proteome research* **2014**, *14*, 700–710.

(18) Ezkurdia, I.; Vázquez, J.; Valencia, A.; Tress, M. Analyzing the first drafts of the human proteome. *Journal of proteome research* **2014**, *13*, 3854–3855.

(19) Ezkurdia, I.; Calvo, E.; Del Pozo, A.; Vázquez, J.; Valencia, A.; Tress, M. L. The potential clinical impact of the release of two drafts of the human proteome. *Expert review of proteomics* **2015**, *12*, 579–593.

(20) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* **1995**, 289–300.

(21) Giai Gianetto, Q.; Combes, F.; Ramus, C.; Bruley, C.; Couté, Y.; Burger, T. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. *Proteomics* **2016**, *16*, 29–32.

(22) Salvetti, A.; Couté, Y.; Epstein, A.; Arata, L.; Kraut, A.; Navratil, V.; Bouvet, P.; Greco, A. Nuclear functions of nucleolin through global proteomics and interactomic approaches. *Journal of proteome research* **2016**, *15*, 1659–1669.

(23) Bouyssié, D. et al. Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics* **2020**, *36*, 3148–3155.

(24) Efron, B. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*; Cambridge University Press, 2012; Vol. 1.

(25) Klammer, A. A.; Park, C. Y.; Noble, W. S. Statistical calibration of the SEQUEST XCorr function. *Journal of proteome research* **2009**, *8*, 2106–2113.

(26) Howbert, J. J.; Noble, W. S. Computing Exact p-values for a Cross-correlation Shotgun Proteomics Score Function. *Molecular & Cellular Proteomics* **2014**, *13*, 2467–2479.

(27) Granholm, V.; Noble, W. S.; Käll, L. 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* **2011**, *10*, 2671–2678.

(28) Verheggen, K.; Ræder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and evolution of database search engines – a central component of mass spectrometry based proteomic workflows. *Mass spectrometry reviews* **2017**,

(29) Bogdanow, B.; Zauber, H.; Selbach, M. Systematic errors in peptide and protein identification and quantification by modified peptides. *Molecular & Cellular Proteomics* **2016**, *15*, 2791–2801.

(30) Barber, R. F.; Candès, E. J., et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **2015**, *43*, 2055–2085.

(31) He, K.; Li, M.; Fu, Y.; Gong, F.; Sun, X. A direct approach to false discovery rates by decoy permutations. *arXiv preprint arXiv:1804.08222* **2018**,

(32) Spirin, V.; Shpunt, A.; Seebacher, J.; Gentzel, M.; Shevchenko, A.; Gygi, S.; Sunyaev, S. Assigning spectrum-specific p-values to protein identifications by mass spectrometry. *Bioinformatics* **2011**, *27*, 1128–1134.

(33) Colaert, N.; Degroeve, S.; Helsens, K.; Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *Journal of proteome research* **2011**, *10*, 5555–5561.

(34) Vaudel, M.; Burkhart, J. M.; Breiter, D.; Zahedi, R. P.; Sickmann, A.; Martens, L. A complex standard for protein identification, designed by evolution. *Journal of proteome research* **2012**, *11*, 5065–5071.

(35) Granholm, V.; Navarro, J. F.; Noble, W. S.; Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of proteomics* **2013**, *80*, 123–131.

(36) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. 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* **2016**, *27*, 1719–1727.

(37) The, M.; Edfors, F.; Perez-Riverol, Y.; Payne, S. H.; Hoopmann, M. R.; Palmblad, M.; Forsström, B.; Käll, L. A protein standard that emulates homology for the characterization of protein inference algorithms. *Journal of proteome research* **2018**, *17*, 1879–1886.

(38) Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2002**, *64*, 479–498.

(39) Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **2013**, *13*, 1352–1357.

(40) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Briefings in bioinformatics* **2012**, *13*, 586–614.

(41) Serang, O.; Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface* **2012**, *5*, 3.

(42) Alves, G.; Yu, Y.-K. Mass spectrometry-based protein identification with accurate statistical significance assignment. *Bioinformatics* **2014**, *31*, 699–706.

(43) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular & Cellular Proteomics* **2015**, *14*, 2394–2404.

(44) Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **1967**, *62*, 626–633.

(45) Candes, E.; Fan, Y.; Janson, L.; Lv, J. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2018**, *80*, 551–577.

(46) Emery, K.; Hasam, S.; Noble, W. S.; Keich, U. Multiple competition based FDR control. *arXiv preprint arXiv:1907.01458* **2019**,

(47) Lu, Y.; Fan, Y.; Lv, J.; Noble, W. S. DeepPINK: reproducible feature selection in deep neural networks. Advances in Neural Information Processing Systems. 2018; pp 8676–8686.

(48) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M., et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research* **2018**, *47*, D442–D450.