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

September 20, 2019

Abstract

Target-decoy competition (TDC) is the most popular method for false discovery rate (FDR) control in bottom-up discovery proteomics. Despite unquestionable statistical foundations, we unveil a so far unknown weakness of TDC: its 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 work pinpoints evidence that TDC will become less reliable along with improved accuracies of modern mass spectrometers. 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 Benjamini-Hochberg framework. This method being simpler to use, faster to compute and more stable than TDC, we argue that it is more adapted to the standardization and throughput constraints of nowadays proteomic platforms.

1 Introduction

The majority of 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 these experimental spectra with theoretical ones derived from a list of protein sequences. Since this methodology has been widely adopted, it has been accepted that it could lead to false positive identifications. Indeed, among the tremendous number of spectra generated by the peptide mixture prepared from a complex biological sample, by chance, at least a few of them are expected to match an erroneous sequence. To avoid corrupting the biological conclusions of the analysis, it has become standard 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. This quality control problem being ubiquitous in science, statisticians have extensively studied it. Their main conclusions (See [1] for a proteomic-oriented summary) are the following: *(i)* Due to the random nature of the mismatches, it is impossible to precisely compute the FDP; *(ii)* However, various methods exist to estimate it, and they are termed FDRs (False Discovery Rates); *(iii)* Depending on the experiment, any FDR will provide a more or less accurate estimate; *(iv)* Therefore, practitioners should carefully choose the FDR methodology, and interpret its result as cautiously as an educated guess (*e.g.*, like a political poll before an election).

In parallel, target-decoy competition (TDC) has emerged as the most popular method to estimate the FDP in MS-based discovery proteomics [2]. Its success comes from both its conceptual simplicity and its broad scope of application. Its principle is to create artificial mismatches by means of a specific database (termed "decoy") containing random sequences which differ from those of interest (the "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 becomes possible, for any cut-off score, to estimate the number of target mismatches that will be validated. As any other estimator, TDC-FDR can lead to inconsistent estimates if the theoretical assumptions it is based on are not verified in practice. Notably, the quality of TDC-FDR intrinsically relies on the validity of the Equal

Chance Assumption, *i.e.* the decoy capability to adequately fool the database search engine. If they fool it too much, the TDC-FDR will overestimate the FDP; while if they are too unrealistic to fool the search engine, the FDP will be underestimated [3]. This is the very reason why decoy database construction and conditions of application have been extensively studied. Among them, *(i)* the search engine must be compliant with TDC [4]; *(ii)* In theory, the longer the decoy database, the more precise the mismatch score distribution [5, 6] (similarly to political polls which accuracy depends on the number of surveyed citizens); *(iii)* The decoys must respect the cleavage sites [7] to avoid systematic target matching regardless the 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)* The way the decoys are counted has an influence [6]. Beyond, numerous parameters have been reported for their relative importance and have been discussed [9]. This has notably participated to install the competition step of TDC as essential, and to date, target-decoy searches without competition [10] is scarcely reported. Besides, a series of letters from Bret Cooper [11, 12] initiated a controversy regarding the observed downfall of the TDC validation levels with higher resolution instruments. He provided experimental arguments to reject the idea that such downfall was only a positive consequence of instrument evolution, leading to an increment of the numbers of identified peptides. Notably, he pointed that very low quality spectra incompatible with confident peptide identifications could be validated despite stringent FDR cut-off, and concluded on the "*inherent bias*" of "*peptide presumption*" (i.e. the identification of peptides already listed in a target database, see [12]). As this contradicted with empirical and theoretical knowledge, few articles positioned against this view [13, 14] while others confirmed [15, 16], maintaining the *statu quo*.

However, it is possible to reconcile Cooper's observations and 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 converges theoretically towards the FDP. Although essential, this asymptotic property is unfortunately not sufficient to 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 one). Thus, the convergence speed and the stability of the TDC estimator should be verified: If the convergence is very slow or if the TDC provides volatile estimates (i.e. two slightly different conditions will provide estimates of largely different quality), it is possible, in given conditions of application, to practically obtain inaccurate FDRs.

In this article, we first present results that seriously question how MS/MS identification results are classically validated using TDC in discovery proteomics. We notably shed a new light on Cooper's observations, which reconciles opposing opinions: While in our view target and decoy searches can be used to accurately compute FDRs, we legitimate his concerns by showing that, with state-of-the-art highly resolutive instruments, the risk that TDC strongly underestimates the FDP increases. We then describe a series of mathematical transformations of classical identification scores, which authorize the application of the well-known Benjamini-Hochberg (BH) procedure [17] as well as its numerous variants [18] at spectrum, peptide and protein levels. This leads to an original and powerful framework that provably control the FDR without the need for decoy databases. Altogether, the presented results demonstrate that making TDC-FDR compliant with instrument improvements requires unexpected efforts (careful implementation, fine-tuning, manual checks and computational time), while it is possible to simply and accurately validate proteomics results with an alternative strategies.

2 Results

2.1 TDC accuracy depends on mass tolerances set during database searching

The mechanics underlying the mass tolerance issue of [12] is the following: Depending on the instrument having a lower or higher accuracy on the precursor mass measurement due to higher or lower resolving powers, and assuming the search engine is tuned accordingly, more or less decoys are considered as possible competitors for a given spectrum. Thus, MS data acquired with high-resolution mass spectrometers and analyzed with search engines set with narrow mass tolerances *de facto* leads to smaller numbers of decoy matches, and therefore, to lower FDR thresholds. Cooper pinpointed this mechanism, and concluded that "*there may be no remedy*". In contrast, it is now well-established that from a theoretical viewpoint, TDC provide good estimates of the FDP [19, 5, 20].

However, to the best of our knowledge, TDC stability with respect to preliminary filters which non-uniformly reduce the number of decoy competitors, has never been demonstrated so far. This is what

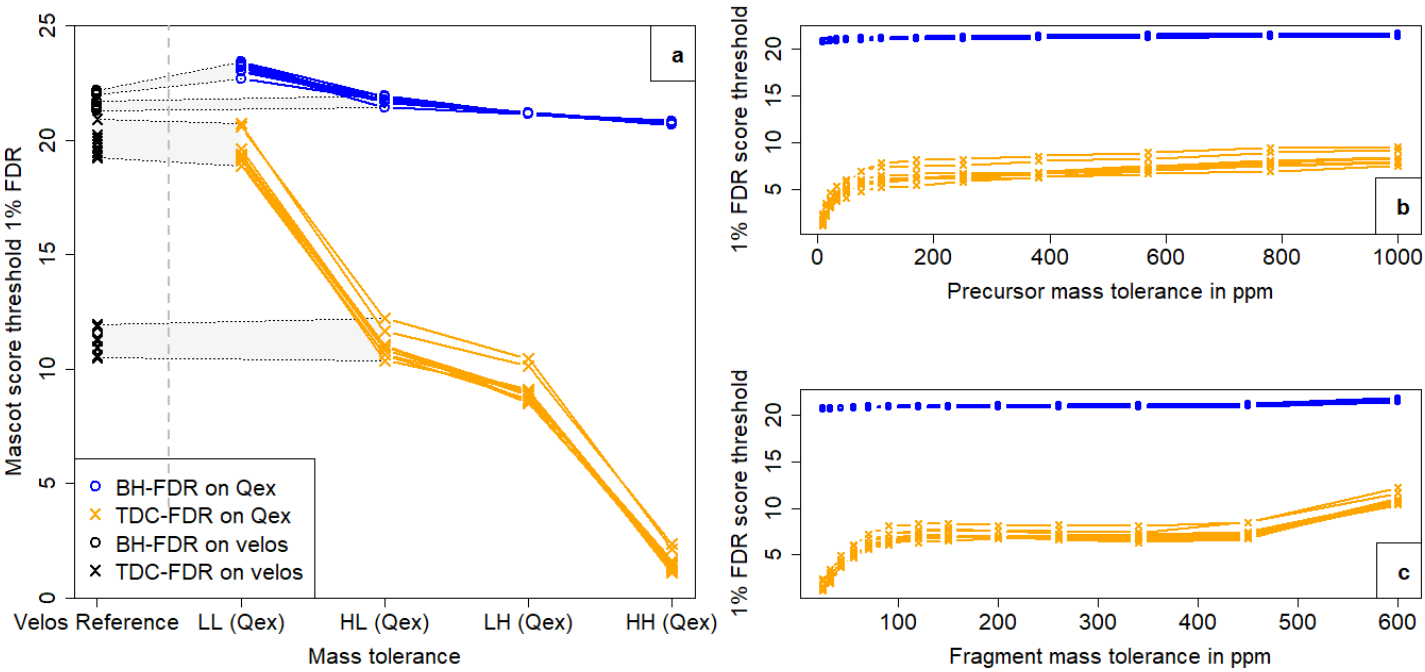


Figure 1: Score thresholds resulting from TDC (Orange) and BH (Blue) filtering at an FDR of 1%, in function of the search engine mass tolerance parameters, for 10 samples analyzed with a Q-Exactive Plus (Qex) instrument. (**a**) the precursor and fragment mass tolerances are tuned to the LL, LH, HL and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

allows us to defend an intermediary viewpoint: If adequately used, it could lead to accurate estimates, even with highly resolutive instruments; Yet, its lack of demonstrated stability is consistent with possible inaccurate FDRs, as reducing the mass tolerance during database searching would increase the risks that the Equal Chance Assumption [2] is no longer valid.

To illustrate the TDC instability, we compared the behavior of the TDC estimate with another estimate whose stability has been established: Benjamini-Hochberg (BH) FDR [17]. Such comparison has already been performed (*e.g.* [21, 22]), yet, the observed discrepancies were not studied in terms of estimate stability. To compute a BH-FDR, one needs a search engine providing PSM scores which can be related to p-values. Fortunately, numerous state-of-the-art search engines do so [23]: For instance, Mascot and Andromeda provide scores of 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. Besides, PepProbe, InsPecT and MyriMatch directly provides p-values as scores. Finally, the scores from X!tandem and Sequest can be transformed into p-values thanks to dedicated wrappers, *e.g.* [21, 22]. Moreover, for the BH-FDR to be accurate, the p-values must be well-calibrated – that is mismatch p-values should uniformly distribute in the [0,1] interval [1]. According to [24, 25], it is possible to assess the well-calibration of a search engine thanks to so-called *entrapment databases*, with which one can empirically estimate the quantiles of the mismatch p-value distribution thanks to artefactual amino acid sequences.

We applied TDC and BH methodologies on results acquired with a Q-Exactive Plus instrument on ten analytical replicates of an *E. coli* lysate, classically used as quality control samples on our platform (see Methods 4.1 and 4.2). PSM were then identified using Mascot using a reverse database allowing

to perform TDC. Figure 1A depicts the score thresholds obtained at a given FDR of 1% as function of the various mass tolerances set during database search (see Supplementary Table 1 for numeric values). While MS and MS/MS spectra were acquired at rather high resolutions (70'000 and 17'500 at m/z 200, respectively), we considered four combinations of mass tolerances tuning at the precursor and fragment levels: LL, HL, LH and HH (L standing for low precision or large tolerance, and H for high precision or narrow tolerance), the latter one (HH) corresponding to tolerances classically used on our platform for Q-Exactive data. Several conclusions can be drawn from these results: First, for each tuning taken individually, the TDC threshold is less stable than its BH counterpart (each set of ten cut-off scores is 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 varies from 1.11 to 20.73 while its BH counterpart is more stable (between 20.67 and 23.43). Fourth, if the different database search sets are interpreted as surrogates for the recent evolution of the instrumentation capabilities, one observes that the discrepancy between TDC and BH increases. From a low resolution instrument situation when TDC was first principled and for which TDC and BH led to roughly similar results, one has progressively shifted to a higher resolution instrument situation where TDC and BH largely diverge. To guarantee that modifying the mass tolerance tuning allows for approximating results obtained with lower resolution instruments, we have analyzed another batch of ten analytical replicates of the *E. coli* lysate using a LTQ-Orbitrap Velos Pro allowing the use of 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 the FDR thresholds were computed as aforementioned. Interestingly, the mapping between the thresholds is excellent, which justifies our methodology: from an FDR viewpoint, switching to a lower resolution dataset analyzed using a search engine tuned accordingly, or keeping the higher resolution data while substantially increasing the mass tolerance tuning, end up with roughly equivalent outputs.

To better capture the influence of the mass tolerance tuning, we conducted more detailed experiments: Starting from the HH setting, we progressively loosened the mass tolerance range, either for the precursor (Figure 1B) or fragment masses (Figure 1C). The obtained results support the four conclusions derived from Figure 1A. In addition, they help understanding the interplays between the mass tolerances and the FDR computation. First, a thinner mass tolerance at precursor (respectively, fragment) level cannot compensate for a too wide tolerance at fragment (respectively, precursor) level: On the right parts of the plots, the TDC scores have not reached their BH counterparts. Second, the fragment mass tolerance tuning influence is not as clear as the precursor one (observe the plateau effect on Figure 1C), yet, precursor and fragment mass tolerance sets have broadly similar effects.

To conclude, our benchmark confirms the downfall of validation thresholds described by Cooper. Moreover, it proves that not only the precursor mass tolerance is concerned, but also, the fragment mass tolerance. More generally, any instrumental or software tuning that may reduce the number of fair decoy competitors in the matching process should be investigated for its potential spurious consequences. In addition, TDC-FDR appears to be more sensitive than BH-FDR, in the sense that reducing the mass tolerances increases its anti-conservativeness faster. Last, the relative instability of TDC-FDR with respect to BH-FDR is blatant.

2.2 Recovering accurate target decoy FDR is possible yet 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, even though wrong. Then, in any random case where the single decoy sequence generated within the corresponding precursor mass window is too unrealistic, one may end up with an almost systematic target assignment relying only on a poor recovery of fragment matches, just because the decoy alternative leads to even less fragment matches In this situation, the Equal Chance Assumption is clearly violated. Thus, to increase the chances of having decent decoys despite too thin mass tolerance parameters, it intuitively makes sense to enlarge the decoy database. Although the link between mass tolerance and decoy database length can be efficiently exploit to limit the computational cost of repeated seaches [26], it is unfortunately 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 for the ratio between the sizes of the entire decoy and target databases, the FDR should read $\frac{d+1}{r \times t}$. However, as demonstrated in the supplemental materials of [6], this is an approximation which only holds when $t \to \infty$ and $d/t \leq 5\%$. Consequently, with a validation threshold corresponding to an FDR of 1%, r should be smaller than 5. Therefore, to increase r in proportion to the reduction of the precursor mass tolerance, the FDR should only be controlled at immaterial levels. An alternative is to refine TDC by a

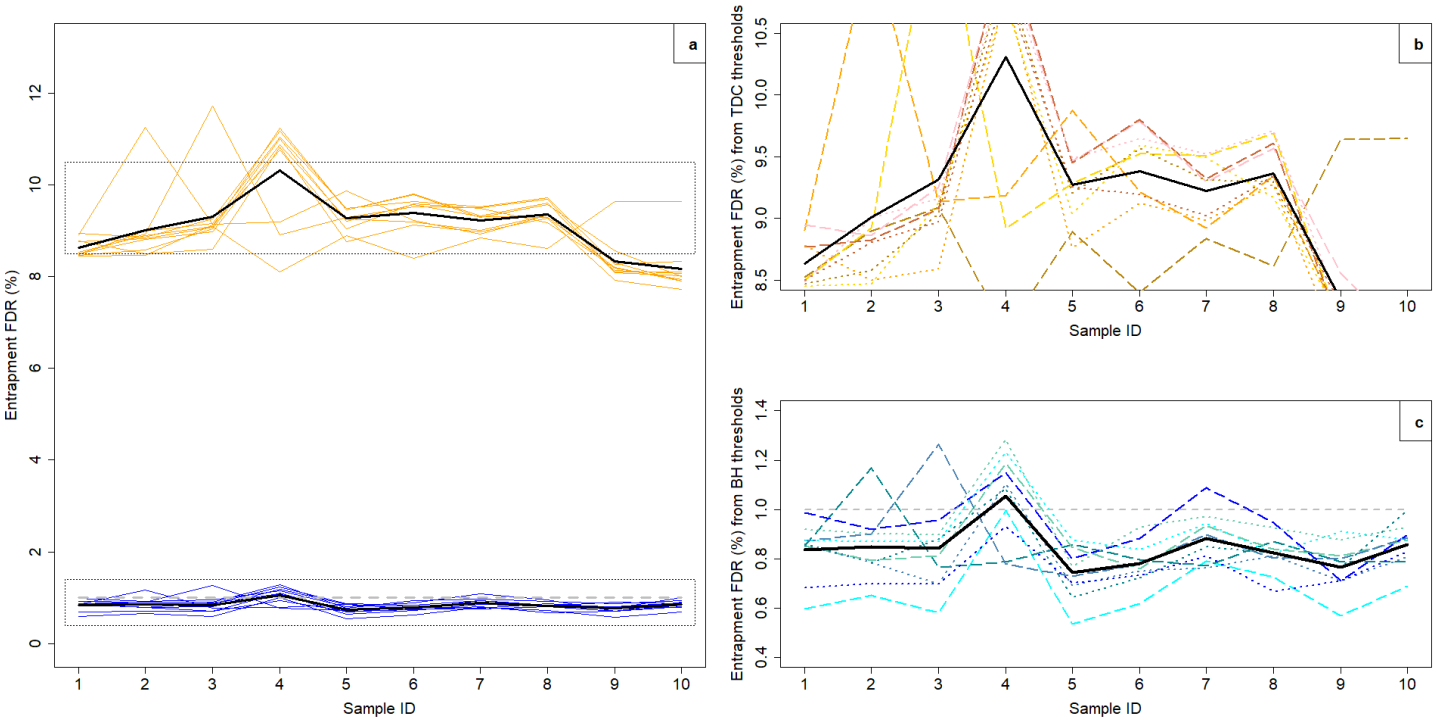


Figure 2: (**a**) Entrapment FDRs computed for each replicate with 10 randomly generated (shuffling) entrapment databases, according to its 1% FDR validation score threshold (gray dash line), computed with BH (blue) and TDC (orange) methodologies; The black plain lines depict the average of the 10 entrapment 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.

procedure akin to *empirical null estimation* [27]. Such approaches are 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 have therefore implemented such a strategy (see Methods 4.3 for details). In a nutshell, we have observed that a fair (yet still unstable) empirical estimate of the null distribution could be achieved with a separated decoy search, without competition. To stabilize the estimations, we average 10 repetitions, each one being based on a different shuffled database, leading to a mix strategy between [10] and [8]. This is why, for each of the 10 sample replicates of *E. coli* lysate analysed with a Q-Exactive Plus, we have reported on Figure 2 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) of the HH setting.

The difference is striking: while BH thresholds (Mascot scores between 20.67 and 20.84) lead to 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 entrapment FDRs larger by a 10-fold (between 7.72% and 11.71%, with an average ≈9.1%). This is insightful for three reasons: *(i)*, it confirms that TDC can lead to largely under-estimated FDRs if inadequately used; *(ii)*, it shows that contrarily to Cooper's concerns, the concept of "peptide presumption" is not inherently biased, as with an adequate decoy search strategy, it is possible to recover coherent FDRs. *(iii)*, Using such an adequate strategy is more complex than a simple TDC with a reverse or shuffle database, or than 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 globally slightly larger than the entrapment ones. This is most probably due to the conservative property of the BH estimator. Note that many methods are available to limit its over-conservativeness (a handful of them are directly accessible in the CP4P R package [18]). Finally, let us observe that the entrapment FDRs highly depends on the entrapment databases: From one randomly generated version to another, the estimated FDR significantly varies. At first look, the FDRs around 1% seem slightly more stable than those around 10%. However, after normalization by the mean FDR value, it indicates that it is finally the opposite that occurs (mean coefficient of variation of 13.90% around 1% entrapment FDR, versus 5.82% around 10% entrapment FDR). This can be easily explained: With lower

FDR thresholds, less decoys passed the threshold so the statistics was computed on smaller sample sizes, which are more sensitive to randomization. This is notably why it is necessary to rely on a boosting strategy (i.e. averaging multiple runs, see above), despite its additional computational complexity.

To conclude, these experiments show that target and decoy searches can be orchestrated to deliver accurate FDRs. However, this requires elaborating different strategies (which notably omit the competition step of the TDC while stabilizing the FDR by averaging multiple independent estimates) which are in fact less intuitive, more hand-crafted and more computationally demanding. On the other hand, the BH procedure is appealing for its simplicity and stability.

2.3 FDR control at peptide and protein levels with BH procedure

The difficult problem of inferring a series of peptide and protein level knowledge from spectrum-level information, all endowed with quality control criteria, has been widely addressed in the literature (see *e.g.* [30, 31] for surveys). However, we are not aware of any available inference system which proposes a peptide or protein level FDR, without a preliminary decoy search. As nowadays, combining multiple levels of FDR control has become a standard good practice, we propose a generic procedure to extend BH-FDR at peptide and protein levels. Moreover, the proposed method is independent of the chosen inference rules (see Methods 4.4). In the following, we only assume the chosen inference rules unambiguously define which PSMs should be accounted for peptide scoring, as well as which peptides are accounted for protein group scoring [32, 28, 29], and we focus on the scoring methodologies.

The most classical peptide scoring methods assume that each peptide is identified by the spectrum with the highest PSM score amongst the Q ones that matched on it [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 one refers to the **PSM score** between peptide sequence $\mathbf{seq}_i$ and spectrum q as S_{iq}, then, one defines the **best-PSM score** as $\max_{q\in[1,Q]} S_{iq}$. This score can possibly 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 on Supplementary Figure 1). Fortunately, the following proposition indicates that it is possible to modify the best-PSM score by applying a formula conceptually close to that of Šidák correction [34] (despite relying on different mathematical hypotheses), and to recover correct calibration.

Proposition 1 *Let $S_1, \dots, S_n$ 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, $X_1, \dots, X_n$. 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 in $[0, 1]$.

Proof: See Methods 4.5.

Therefore, (See Methods 4.5 for the full derivations), one defines the **peptide p-value** of peptide sequence $\mathbf{seq}_i$ as:

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

and the **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, we leveraged the intuition that fragment matches are for PSM scores what peptide matches are for protein scores. This suggests to rely on Fisher's test to define protein scores/p-values from the ones of the best subset of peptides. As similar intuitions have frequently been investigated in the literature (see [26, 32, 28, 29]), the full derivation is presented in Methods 4.6. However, to the best of our knowledge, our approach is 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 4.7); (*ii*) the impact of the poorly conclusive peptide-level evidence in an open-world assumption context (see Methods 4.8). Finally, for a

protein sequence $\mathbf{seq}_\pi$ identified by k specific peptides with scores $S_1^\diamond, \ldots, S_K^\diamond$, one defines the **protein p-value** 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 K identified peptide set; A is a peptide set of 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.

2.4 Application to multiple-level validation

Thanks to these provably well-calibrated PSM, peptide and protein level scores, it is possible to design a pletora of elaborated validation strategies based on BH framework, so as to minimize false identifications without needs for decoy database. Discussing and comparing these strategies in various applicative contexts (discovery proteomics, relative quantification, proteoform analysis, etc.) stands beyond the scope of this article.

However, to quickly illustrate their potential, we have validated the first *E. coli* lysate replicate, using the following BH-FDR thresholds: 10% at spectrum level, no filter at peptide-level and 1% at protein-level. To do so, one obviously did not need any decoy or entrapment database. In spite of this, and because it delivered a striking illustration of the false identification discrimination capabilities of the proposed framework, we have introduced shuffled sequences in the searched database (see Methods 4.3). We considered a challenging scenario where the number of decoys was set to five times the number of target sequences.

The final statistics are the following: Only 9 decoys sequence matches (mapping on 9 distinct peptides from 9 distinct proteins) appeared within the 1322 validated proteins. These 1322 proteins are composed of 8764 validated peptides, which in turn, are supported by 11480 validated PSMs. and of (0.68%). The percentage of remaining decoy sequences is thus 0.68% at protein level, 0.1% at peptide level and 0.08% at PSM level. Although these percentages must certainly not be interpreted as FDRs, it is interesting to qualitatively compare them with the results delivered by classical TDC: For a marginal gain of 6% in the number of validated PSMs (12176 vs. 11480 - see Supplementary Table 1) classical TDC provided an FDR which underestimated by almost a tenfold the FDP, while on the contrary, our approach was conservative and led to an immaterial proportion of decoy mappings, despite the fivefold decoy over-representation.

3 Discussion

As a whole, this work sheds a new light on a crucial step of bottom-up proteomic experiments, the validation of identification results.

First, it illustrates that the TDC and the BH estimates of the FDP have progressively diverged along with the improvement of the accuracy of MS instruments. We have demonstrated this was due to the lack of stability of TDC with respect to the precursor and fragment mass tolerances set during database search. Although this can be partially balanced by suppressing the competition step of TDC [10], this instability adds up to one (well-known) induced by the random generation of decoy sequences. Therefore, even though it is possible to refined target-decoy strategies to partially cope for this instability, it is not satisfying: *(i)* These strategies are not as stable as the BH one, specially at lower FDRs; *(ii)* They are more complex to organize (implementation, computational cost) than BH procedure and require additional manual checks, notably with regards to stability; *(iii)* They do not provide any guarantee for FDR estimates in the future (on datasets acquired with more resolutive next generation instruments for which ever thinner mass tolerance windows will be expected); *(iv)* They provide no guarantee with respect to any pipeline modification that freely changes the number of target and decoy candidates (e.g. [35]).

Second, this work provides new peptide and protein scores which provably respect the calibration condition of BH procedure. Then, it shows that it is possible to orchestrate multiple levels of BH validation, in a practically simple and statistically supported way. Concretely, implementing BH-FDR at PSM-, peptide- and protein-level is straightforward (see Code Availability 5) for any database search engine which PSM p-values can be accessed. Moreover, it does not require any precise tuning, making its introduction into any platform pipeline straightforward.

For all these reasons, we propose to modify in-depth how FDR is estimated in discovery proteomics using database search and suggest to replace TDC by BH-FDR. Nevertheless, as a theoretical research field, TDC remains a subject of interest. The original idea of Elias and Gygi [2], essentially based on empirical observations and intuition, has fertilized the field of theoretical biostatistics and led to the idea that simulating null tests from the data could lead to efficient alternative to BH procedures, which provably 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 benefit to computational proteomics in the future, exemplifying 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 previously described [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 4.3 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 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, 38, 39, 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] and [26] which have first 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 [38].

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].

4.4 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 [40, 41]; or associated to the entire filtered list (typically, an FDR, but other multiple test corrections methods exist [42]).

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.

4.5 Peptide score definition

Let us have a peptide sequence $\mathbf{seq}_i$, a spectrum $\mathbf{spec}_j$ and a score $S^\circ_{ij} = \texttt{Score}(\mathbf{seq}_i, \mathbf{spec}_j)$ 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:

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

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:

$$S_i^\top = \max_{q\in[1,Q]} S^\circ_{iq} \tag{6}$$

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:

$$p_i^\top = \min_{q\in[1,Q]} p^\circ_{iq} \tag{7}$$

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:

$$\begin{aligned}
\mathbb{P}[Y \leq t] &= \mathbb{P}\left[1 - \left(1 - 10^{-\frac{1}{10}\cdot \max_\ell S_\ell}\right)^n \leq t\right] \\
&= \mathbb{P}\left[1 - \left(1 - \min_\ell \left[10^{-\frac{S_\ell}{10}}\right]\right)^n \leq t\right] \\
&= \mathbb{P}[1 - (1 - \min_\ell [p_\ell])^n \leq t] \\
&= \mathbb{P}[1 - (\max_\ell [1 - p_\ell])^n \leq t] \\
&= \mathbb{P}[(\max_\ell [1 - p_\ell])^n \geq 1 - t] \\
&= 1 - \mathbb{P}[(\max_\ell [1 - p_\ell])^n < 1 - t] \\
&= 1 - \mathbb{P}[\max_\ell [1 - p_\ell] < (1 - t)^{1/n}] \\
&= 1 - \mathbb{P}\left[\bigcup_\ell \left\{(1 - p_\ell) < (1 - t)^{1/n}\right\}\right] \\
&= 1 - \prod_{\ell=1}^{n} \mathbb{P}[1 - p_\ell < (1 - t)^{1/n}] \\
&= 1 - \prod_{\ell=1}^{n} \mathbb{P}[p_\ell \geq 1 - (1 - t)^{1/n}] \qquad (8)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}[p_\ell \geq 1 - (1 - t)^{1/n}] &= 1 - (1 - (1 - t)^{1/n}) \\
&= (1 - t)^{1/n} \qquad (9)
\end{aligned}$$

So that

$$\begin{aligned}
\mathbb{P}[Y \leq t] &= 1 - \prod_{\ell=1}^{n} (1 - t)^{1/n} \\
&= t \qquad (10)
\end{aligned}$$

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:

$$p_i^\diamond = 1 - (1 - p_i^\top)^Q \qquad (11)$$

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

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 [43], 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. Finally, 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].

4.6 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 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 [43] (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:

$$p^{\dagger}_{\pi} = \int_{0.2\ln(10)S^{\oplus}_{\pi}}^{\infty} f_{2K}(x)dx \tag{12}$$

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

4.7 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 [44, 45].

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 [40]. 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.

4.8 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, [46]): 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** $S^\star_\pi$:

$$S^\star_\pi = \max_{A \in 2^{\{1,\dots,K\}}} S^\dagger_A \tag{13}$$

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

$$p^\star_\pi = 10^{-\frac{S^\star_\pi}{10}} \tag{14}$$

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

$$p^\star_\pi = \min_{A \in 2^{\{1,\dots,K\}}} p^\dagger_A \tag{15}$$

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 [47] (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 [48] 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. Concretely, if the scores of all the PSMs pointing toward a same peptide sequence are stored are in a vector `psm.scores`, then, the peptide p-value `pep.pval` and peptide score `pep.score` are given by 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 following code can be used to compute the protein score `prot.score` and p-value `prot.pval` of a protein which peptide scores are stored in a vector `pep.scores`:

```
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 on top of the PSM scores provided by the search engine, one simply has to run a BH procedure with one of the numerous numerous R packages proposing a function to do so.

Implementing these scores with other languages is possible, even though a difficulty may appear: Not so many languages provide a function to directly access the logarithm of the tail of the χ^2 CDF (cumulative distribution function), so that it may be required to compute the CDF first, then subtract it to 1, and finally, take the log. However, for rather high protein scores, the resulting probabilities may be too small and thus rounded to 0, so that logarithmized them leads to infinite protein scores.

6 Acknowledgements

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.

7 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 to the final manuscript.

8 Competing Interests

The author declares 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] 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.

[39] 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.

[40] 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.

[41] 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.

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

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

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

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

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

[47] 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.

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

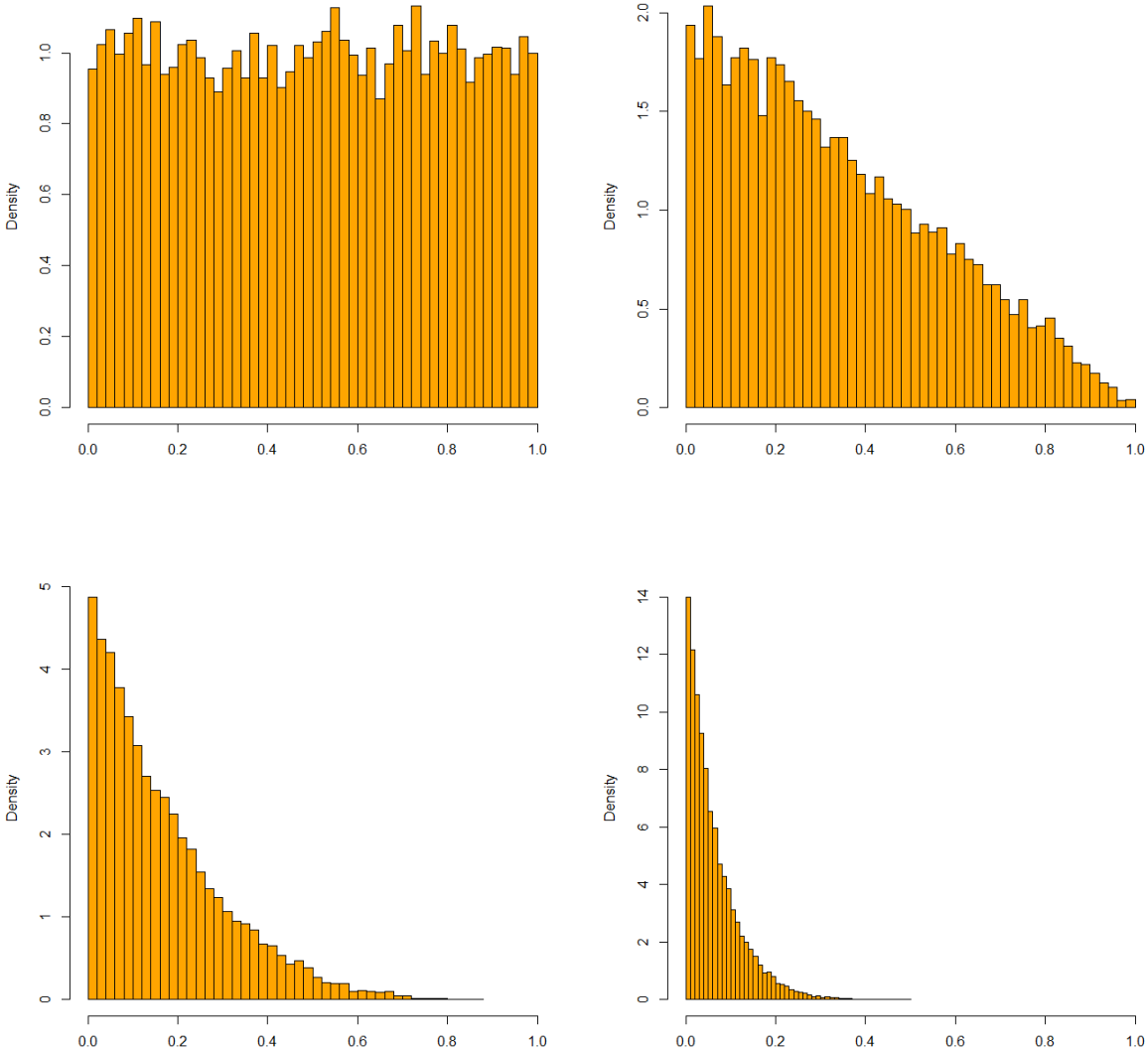
Supplementary Information

A Supplementary figures and tables

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

		BH								TDC							
		Low-Low		High-Low		Low-High		High-High		Low-Low		High-Low		Low-High		High-High	
	Queries	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score
QEx	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

		BH								TDC							
		Low-Low		High-Low		Low-High		High-High		Low-Low		High-Low		Low-High		High-High	
	Queries	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score	PSM target	Score
Velos	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.