**Supplementary Note**

Here we discuss the target-decoy approach (TDA) to estimate false discovery rates (FDRs) for mass spectrometry (MS) experiments, and show how it can be employed in conjunction with *de novo* peptide sequencing. We then describe consequences of its use in stratified databases (e.g. composed of conventional and cryptic peptides). Finally, we discuss an example from the literature, where non-compliance with the fundamental conditions of TDA has led to contradictory results.

**1. Target-decoy approach**

Here, we briefly expand on the main messages from Ref. (25). TDA is the most widely (and virtually exclusively) used way to estimate FDRs for peptide identifications from MS data (26).

The basic components for TDA are (i) a target sequence database (ii) a decoy sequence database, and (iii) a program that, based on an input sequence database D, assigns a peptide from D and an associated score to each tandem mass spectrum (peptide-spectrum matches, PSMs).

Many PSMs refer to false peptides (i.e. peptides that were actually not in the sample analyzed by MS). Therefore, a threshold *t* is used, and all PSMs exceeding *t* are called discoveries. The purpose of TDA is to estimate the rate of false discoveries, i.e. the percentage of false PSMs in the set of all PSMs exceeding *t*. In addition to these so-called PSM FDRs, peptide FDRs can be defined as the percentage of false peptides in the set of all peptides from PSMs exceeding *t*.

The TDA is as follows: If all decoy peptides are false (**condition 1**), and the probability for a discovery from the decoy database is equal to the probability for a false discovery from the target database (**condition 2**), then the expected value of the number of false discoveries from the target database is the number of decoy discoveries. The FDR can then be computed by dividing the expected value by the total number of target discoveries.

Violations of condition 1 lead to conservative FDR estimates (i.e. the actual FDR is less than the estimated FDR). Condition 1 is usually ensured by filling the decoy database with peptides that are not in the known proteome.

More problematically, violations of condition 2 can lead to underestimated FDRs (i.e. the actual FDR is greater than the estimated FDR). This is the case if the probability of decoy discoveries is less than the probability of false target discoveries. Condition 2 holds if (i) target and decoy databases have the same size and (ii) the properties of target and decoy peptides are similar. This is usually achieved by reversing or shuffling the target database.

Obviously, condition 2 is crucial. For the sake of the following arguments, we will reformulate it:

**Master condition:**
The score histogram of decoys is the same as the score histogram of false targets.

There are two important remarks here: First, the “decoys” and “targets” might both refer either to PSMs or peptides. Second, if the master condition as stated is fulfilled, condition 2 above is fulfilled (as we defined “discoveries” by a threshold on the score). The converse is not necessarily true, as condition 2 demands equality only for the chosen threshold, not for all thresholds.

Moreover, the FDR as defined above is the expected value of a random variable and depends on the choice of the decoy database. It is therefore pertinent to make sure the estimate of the FDR is precise by ensuring a large enough decoy database.

We can categorize peptides in the databases into three classes:

1. Decoy peptides
2. False target peptides
3. True target peptides

Classes 1 und 2 have the same score distribution with density $p\_{false}(x)$, and class 3 has a distinct score distribution with density $p\_{true}(x)$ (see Fig. 1a-b). We cannot observe class 2 and 3 separately, but the mixture of both (i.e. the scores of the target database; see Fig. 1c). However, we can observe the score histogram of decoys, which is the same as for false targets based on the master condition (see Fig. 1d). For any threshold *t*, the number of decoys exceeding *t* equals the number of false targets exceeding *t*. Thus, we can choose a threshold such that the percentage of false targets is 1% of all targets exceeding *t* (see Fig. 1d).

**a**,

**b**

**Fig. 1** Score distributions. **a,** The score distribution of false peptides $p\_{false}(x)$, and **b,** the score distribution of true peptides $p\_{true}(x)$. **c,** Score histogram of 500 false targets and 2000 true targets. False targets are stacked on top of the true targets. **d,** Score histogram of 500 decoys, 500 false targets and 2000 true targets. False targets and decoys are stacked on top of the true targets. The 1% FDR threshold is indicated.

**c**,

**d**,

**2. TDA for de novo peptide sequencing**

As defined above, the third component for TDA is a program that, based on an input database D, assigns a peptide from D and an associated score to each tandem mass spectrum. Usually this program is a sequence database search tool such as Mascot, Sequest, Andromeda or X!Tandem. However, given any de novo peptide sequencing tool (such as implemented in Peaks), the following pseudocode also constitutes such a program:

1. Perform de novo peptide sequencing
2. Output the top *n* candidates for each spectrum
3. Find the top-most candidate that occurs in the sequence database

Such an approach has been proposed earlier (27). There, the fundamental problem was to identify the “*n*” for a given spectrum, i.e. how many candidates must be considered to identify one that occurs in the sequence database. We noticed that due to (i) high mass accuracy tandem mass spectra and (ii) the length restriction (8-11 amino acids) of HLA-I peptides, the top candidate is correct for a large fraction of spectra and among the top 10 for virtually all spectra where otherwise a peptide could be identified by standard database search.

**3. Stratified databases**

The theory behind TDA also applies if the target database actually consists of multiple strata, where each stratum has a different size and likelihood of identifying a peptide. However, the estimated FDR is global and does not allow drawing conclusions for individual strata (28).

**Example:** The target database consists of strata A and B. 9900 peptides from A and 100 peptides from B have been found at a global FDR of 1%. Clearly, we expect 100 peptides to be false discoveries. The TDA does not provide arguments to assume a 1% FDR for both A and B. Thus, all peptides from B could be false. This case can indeed happen, e.g. if B completely consists of false peptides, and is large enough (see Fig. 2).

A straightforward solution is to employ TDA per stratum. However, as individual strata can be small, which would lead to highly imprecise FDR estimates, we follow a different path in Peptide-PRISM (inspired by the popular Peptide Prophet (29); this is done for each peptide length *l*): We use the histogram of all decoy peptides to estimate the score distribution of false hits with density $\_{l}(x)$ (by using a non-parametric fit based on a uni-modal regression model). Then, based on the most promising stratum (that has the greatest number of true hits, i.e. CDS), we estimate the score distribution with density $p\_{l}(x)$ of true hits (by first subtracting the CDS decoy histogram from the CDS target histogram, and then applying the same fitting procedure as for decoys). We assume that for any stratum/category *c*, the actual score distribution of targets is a mixture distribution of these two components with density $f\_{l,c}\left(x\right)=w\_{l,c}⋅p\_{l}\left(x\right)+\left(1-w\_{l,c}\right)⋅n\_{l}(x)$ . Here, $w\_{l,c}$ is the total fraction of true targets of *c* with length *l* and can be used to estimate the total number of true targets from *c*. In Peptide-PRISM, $w\_{l,c}$ is estimated numerically using maximum likelihood (30). Then, the mixture distribution is fully specified and allows to compute FDRs (and other statistics).

**a**,

**b**

**Fig. 2** Stratified databases. **a,** The distribution of false (red) and true (black) targets in the database stratum A from the example (see Text). The 1% FDR threshold is indicated. **b,** Same as a, but with the targets from stratum B (that are all false) added. The now shifted 1% FDR threshold is indicated.

**4. Rationale of Peptide-PRISM**

**Overal model design:** The mixture model for all peptides of length *l* consists of two component distributions: One for false hits, and one for true hits. It is important to emphasize that we estimate a single distribution of false hits and a single distribution of true hits for each length *l*. This is based on our observation that the score distribution of decoys originating from CDS sequences is not different from the distribution of decoys from any other stratum (Peptide-PRISM generates diagnostic plots in each run to check this). This means that the potentially distinct amino acid composition in different strata has no significant effect on the score distribution. Consequently, we also use a single distribution for true hits.

Thus, for each stratum, the observed score distribution of targets (for peptides of length *l*) is a mixture of these two distributions, and the only difference is the relative contribution of the two mixture components, respectively (e.g. for the CDS stratum, the true hit distribution generally has more weight than for the intergenic stratum). These contributions can be estimated using maximum likelihood, and FDRs can be directly extracted from the mixture model as proposed by Peptide Prophet (29). Importantly, Peptide Prophet is designed for standard proteomics experiments, i.e. without stratification.

**Estimating the false hit distribution:** For estimating the false hit distribution, we use all decoy hits of length *l* from all strata. As pointed out above, there is no significant difference for decoys generated from different strata. Thus, we can use all available decoys to obtain estimates as precise as possible.

**Estimating the true hit distribution:** Decoys are by definition false hits. For true hits, it is not possible to generate such a database. To resolve this problem, Peptide Prophet (29) estimates all parameters of the mixture model simultaneously using all observed identification scores (i.e. the mixture weight as well as the parameters of the component distributions). It can do this with or without incorporating information of decoy hits for the false hit component.

As Peptide-PRISM always generates decoys, we use a more straight-forward way to estimate the true hit distribution: In any stratum and for any peptide length, we have *ds* hits in the decoy database and *ts* hits in the target database for some (integer) identification score *s*. Extending the argument of the target decoy approach, the expected number of (false) decoy hits with score *s* is the same as the expected number of false target hits with score *s*. Thus, we can obtain the expected number of true target hits with score *s* as *ts* - *ds*. In Peptide-PRISM, we always use only use the histogram from the most promising stratum, i.e. the one that induces the highest total number of expected true targets.

**Non-parametric fit:** In principle, we could use the true hit and false hit histograms directly in place of the two component distributions of the mixture model. However, due to the limited sample size, both are, to some extent, noisy (again, this can be inspected in the diagnostic plots produced by Peptide-PRISM). To smooth these rugged histograms, we tested several alternatives and obtained the visually best results using a non-parametric method (uni-modal regression). This is based on much weaker assumptions than the Peptide Prophet, which fits a gamma distribution for the false hit distribution, and a normal distribution for the true hit distribution. It alternatively allows to use kernel density estimates as a semi-parametric method, which is not appropriate for fitting the ALC scores produced by the *de-novo* search of Peaks, as they are bounded between 0 and 100, which leads to considerable boundary artifacts when kernel density estimators are used.

**Overlap of target and decoy sequences:** For the standard TDA, peptides that are both in the target and the decoy database pose a problem: They could be true hits, but if they are treated as targets, the decoy database is effectively smaller than the target database. This can be handled by using a corresponding factor when computing the FDR. For our mixture modeling approach, overlaps of target and decoy sequences do not represent such a problem: The decoys are only used to construct the false hit score distributions. FDRs are based on the mixture model that is estimated from targets alone (i.e. we can safely treat all ambiguous peptides as targets).

**5. Violations of the master condition in previous HLA-I immunopeptidome studies on PCPS**

The contribution of proteasome catalyzed peptides splicing (PCPS) in immunopeptidomes has raised substantial controversy in the past. Results ranged from more than 33% (20) to likely less than 2-6% (3) and back to 16% (2) for the same dataset (GR-LCL). Of note, these numbers are based on TDA estimates and their validity is therefore fully dependent on compliance with the master condition.

We show here that these claims on these percentages were based on non-compliance with TDA. Based on our analyses the actual percentage of GR-LCL is < 0.3%, of which most appear to undermine the TDA, even if the master condition is not violated (this is because a large part of the identified sequence actually is correct, see main text).

In both Refs. (20) and (2) , the workflow as reported was as follows: All cis-spliced (i.e. from the same protein) peptides with a maximal intervening distance of 25 amino acids were generated and sequences occurring consecutively (i.e. non PCPS) in the proteome were removed. To make sequence database search using Mascot feasible, the database was then reduced to only contain peptides that match the mass of a precursor peak on the MS data (with a very stringent mass tolerance). Peptides were then concatenated into protein-like sequences. Finally, the sequences of these pseudo-proteins were shuffled to produce a decoy database. The same filtering, concatenation and shuffling was repeated for consecutive peptides, and the PCPS or non-PCPS status was indicated in the pseudo-protein name to make this information directly accessible in the Mascot output.

At the first glance, there is no obvious reason why this is non-compliant with TDA. However, sequence database search tools such as Mascot only consider peptides that are within a user-defined mass window around the precursor mass to match against a tandem mass spectrum. By construction, all target peptides matched at least one experimental precursor mass. In contrast, by shuffling concatenated protein sequences, the large majority of peptides in the decoy database will not match any precursor mass. This generates a decoy database that is effectively much smaller than the target database, violating the master condition.

Of note, the fact that about 50% of the peptides reported in the newer publication (2) (the initial publication (20) did not contain tables with spectrum scan numbers to investigate that) as being PCPS peptides also match conventional peptides in the proteome equally well is an issue separate from non-compliance with TDA. For the majority of those peptides, the reported spliced sequence itself or sequences with L/I exchanges actually occur consecutively in the proteome (Extended Data Table 1), indicating an issue with the assignment of spliced vs. non-spliced peptides.