A non-parametric cutout index (npCI) for robust evaluation of identified proteins

Oliver Serang[1,3], Joao Paulo[2,3], Hanno Steen[2,3*], and Judith Steen[1,3*]

[1]*Department of Neurobiology, Harvard Medical School and Boston Children's Hospital*

[2]*Department of Pathology, Harvard Medical School and Boston Children's Hospital*

[3]*Proteomics Center, Boston Children's Hospital*

[*]*Contributed equally*

Supplementary Methods, Figures, and Text

Supplementary Figure 1 Robustness of (k, n) parameters to decoy database design

Supplementary Figure 2 npCI and two-peptide estimates and the relationship between sensitivity and FDR

Supplementary Methods

Supplementary Figure 1: Robustness of (k, n) parameters to decoy database design

In the main text, a reversed decoy database is used. Here (**Supplementary Figure 1**) we demonstrate robustness to decoy database choice by achieving similar results using a shuffled decoy database, generated by shuffling proteins with Mascot. All searches are performed against a combined target-decoy database (*i.e.* using target-decoy competition).

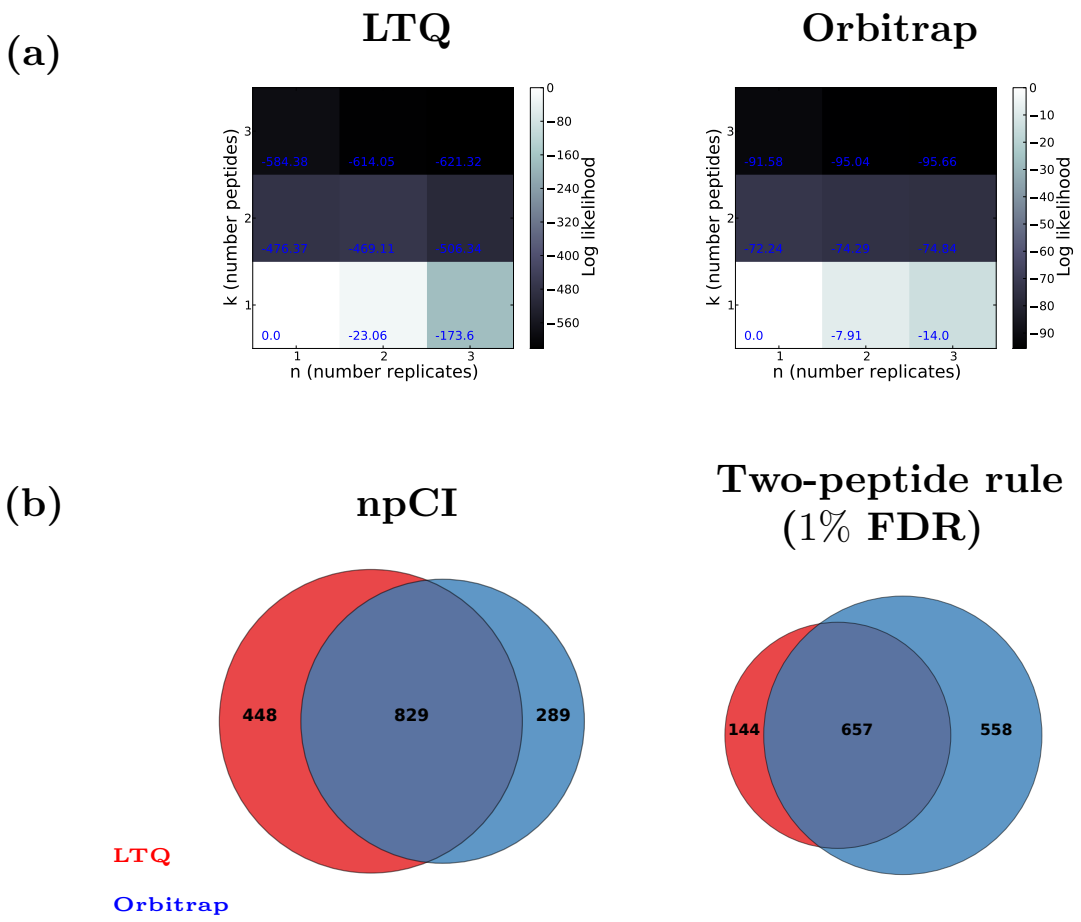

Figure 1: Further evaluation of strategies for protein identifications from HeLa cell replicates, using a shuffled decoy database rather than a reversed decoy database. (**a**) Heat map of log likelihoods (base e) from the protein rankings produced by matching at least k peptides in each of at least n replicate experiments on LTQ and Orbitrap instruments. To compare to the combined human and reversed decoy database used in the main text, spectra were searched against a combined human and shuffled decoy database. To enable comparison of similarly shaded squares, the log likelihoods (relative to the maximum likelihood) are written in blue. (**b**) Venn diagram (drawn to scale) showing the overlap in the target proteins identified with $(k, n) = (1, 1)$ (the parameters chosen using the npCI), compared to the overlap produced by using the two-peptide rule with a peptide FDR of 1%.

Supplementary Figure 2: npCI and two-peptide estimates and the relationship between sensitivity and FDR

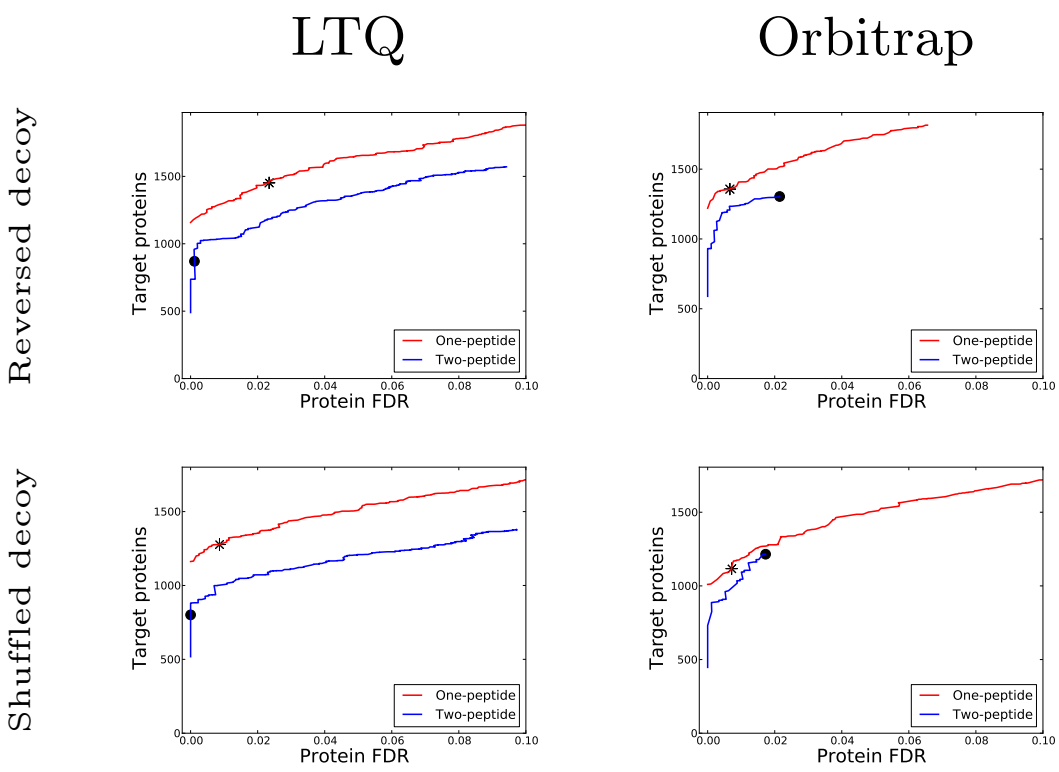

Figure 2: Comparison of npCI and two-peptide rule with fixed peptide (FDR). For both LTQ and Orbitrap instruments, the relationship between number of target identifications and protein-level FDR is shown (ungrouped to match Venn diagrams in **Figure 2**). Asterisks indicate the protein set chosen by the npCI, and a circle indicates the protein set chosen by the two-peptide rule with a 1% peptide FDR cutoff. On the LTQ, using the two-peptide rule with a 1% peptide FDR is unnecessarily conservative; 870 proteins are identified at $\approx 0\%$ FDR; however, the two-peptide rule can identify 959 proteins at the same protein FDR. Conversely, it may be inconsistent and unnecessarily permissive (if only slightly) on the Orbitrap. Using one peptide ($k = 1$) appears to outperform two peptides ($k = 2$), regardless of the decoy database used. This concurs with **Figure 2–3**. Series are truncated when no more proteins can be identified with the given peptide rule. For the high mass accuracy Orbitrap data, this truncation results in a low maximum peptide FDR.

1 Supplementary Methods

A nonparametric approach to evaluating identified proteins Here we present an unbiased probabilistic measure for evaluating protein identifications. Importantly, our approach does not use assumptions to parametrically model and directly evaluate the identified proteins and peptides. The npCI does not attempt to characterize present proteins (by stating, for example, that present proteins have at least two peptides matching spectra with Mascot scores[1] exceeding 80). Instead, our measure is motivated by a simple and complementary idea: the best set of identified proteins is the set where the distribution of remaining target peptide scores (*i.e.* the scores from the peptides not assigned to any of the proteins listed as being identified) most closely resembles the distribution of scores from decoy peptides, which can be used to model empirically absent peptides[2,3] (**Figure 1**). We define the npCI of a set of proteins as the likelihood that the peptides unaccounted for by those proteins are from the same score distribution as empirically absent peptides. In contrast to the standard approach, decoy proteins are not used in any way by the npCI method (they are used in this manuscript as an *ad hoc* measure to estimate the reliability of the threshold chosen by the npCI approach).

Let D denote the observed target peptide scores, a collection of sets comprised of the peptide-spectrum match (PSM) scores for every peptide. X denotes the set of present proteins in an experiment; x would denote a particular set of identified proteins. Partition D into $D^{(present)}$ and $D^{(absent)}$, where $D^{(absent)}$ denotes the peptide scores from peptides not found in any present proteins and $D^{(present)}$ denotes the set of peptide scores from peptides found in at least one present

protein (these peptides *may* be present, but are not *necessarily* present). Assuming that the peptide scores from the two classes do not influence each other directly (given $X = x$), then the likelihood of a set of present proteins $X = x$ can be factored into two terms (via conditional independence):

$$\Pr(D|X = x) = \Pr(D^{(absent)}|X = x) \Pr(D^{(present)}|X = x)$$

An empirical distribution of present peptide scores is unknown, and is experiment-specific; therefore, we cannot model it in an unbiased manner. For this reason, we effectively ignore the present peptide scores by assuming that all present peptide score distributions are equally likely:

$$\Pr(D|X = x) \propto \Pr(D^{(absent)}|X = x)$$

The decoy peptides (with scores denoted $D^{(decoy)}$) are drawn from roughly the same distribution as the absent target peptides[2,3]; therefore, we estimate $\Pr(D^{(absent)}|X = x)$ as the probability that $D^{(absent)}$ and $D^{(decoy)}$ are drawn from the same distribution. Hence, the likelihood of the identified protein set $\Pr(D|X = x)$ is proportional to the probability that $D^{(absent)}$ and $D^{(decoy)}$ are drawn from the same distribution.

Justification for a novel measure of divergence: Existing commonly used measures of divergence do not adequately quantify the divergence between the absent and decoy peptide scores. For example, entropy-based measures like the symmetrized Kullback-Liebler (KL) divergence require either binning or smoothing; in both cases, the divergence will reach a minimum when either the fewest

bins or the most aggressive smoothing are used (for the sake of simplicity, we refer to this level of discretization or binning as the degree of smoothing). Furthermore, the optimum (*i.e.* the least divergent peptide set) varies substantially based on the degree of smoothing. Lastly, entropy-based measures require assumptions to make a probabilistic interpretation; the more assumptions we use, the less agnostic and robust our score will become.

Likewise, existing frequentist measures like the Kolmogorov-Smirnoff (K-S) test statistic are not sufficient for our application. The K-S statistic approaches zero with enough samples when the two distributions tested are identical; however, it does not, in our case, reasonably quantify the divergence between the two distributions when they are not the same. For example, the K-S statistic values are determined by the largest difference in the cumulative distributions. This value is dominated by the score regions which contain most peptides; the absence or over-enrichment of the much rarer high-scoring peptides is often the most informative in our application, but does not substantially alter the K-S statistic. As a result, the K-S statistic decreases in a concave up manner, but contains a large range where the test statistic is nearly indistinguishable. To resolve the best-scoring protein set using this metric, it would be necessary to specify a prior distribution on the number of included proteins (to weigh a minor improvement through a decreased K-S statistic against the many additional proteins necessary).

Nonparametric Bayesian probabilistic measure of divergence: Given two finite samples of continuous values ($D^{(absent)}$ and $D^{(decoy)}$), we propose a simple Bayesian method to estimate the probability that the two samples are drawn from the same distribution.

We first present the method by using discrete $k-$bin $(b_1 \dots b_k)$ Dirichlet distributions and then generalize to a Dirichlet process.

We then smooth the observed data and nonparametrically estimate each probability density function using kernel density estimation.

$$K(s_1, s_2) = e^{\frac{\|s_1 - s_2\|^2}{2\sigma^2}}$$

$$pdf_{D^{(decoy)}}(b_i) \propto \sum_{s \in D^{(decoy)}} K(s, b_i)$$

We then use the smoothed peptide counts (the product of the estimated density and the number of peptides) as the set of concentration parameters in two Dirichlet distributions. We symmetrically compute, for all bins, the probability that the decoy PSMs are drawn from the same distribution as the absent target PSMs and then compute the geometric mean probability that all binned densities would be drawn from the other Dirichlet distribution. A distorted geometric mean is taken to ensure the relative likelihoods do not depend on the number of bins k, while requiring that each bin contributes equally:

$$D^{(absent)} \sim Dirichlet(k, \{1 + n_{decoy} pdf'_{D^{(decoy)}}(b_i) \forall i \in \{1 \dots k\}\})$$

$$D^{(decoy)} \sim Dirichlet(k, \{1 + n_{absent} pdf'_{D^{(absent)}}(b_i) \forall i \in \{1 \dots k\}\})$$

where, for s_{low} and s_{high} equal to the lowest and highest scores of interest, respectively, pdf' denotes $(s_{high} - s_{low})pdf$, because the pmf function should be proportional to the pdf after the bucket width has been accounted for. Thus

$$\Pr(D^{(absent)}|pdf'_{D^{(decoy)}}) = \left(\frac{1}{\beta(n_{decoy}pdf'_{D^{(decoy)}})}\prod_i pdf'_{D^{(absent)}}(b_i)^{n_{decoy}pdf'_{D^{(decoy)}}(b_i)}\right)^{1/k}$$

$$\Pr(D^{(decoy)}|pdf'_{D^{(absent)}}) = \left(\frac{1}{\beta(n_{decoy}pdf'_{D^{(absent)}})}\prod_i pdf'_{D^{(decoy)}}(b_i)^{n_{absent}pdf'_{D^{(absent)}}(b_i)}\right)^{1/k}$$

where

$$\beta(\alpha) = \frac{\prod_i^k \Gamma(1+\alpha_i)}{\Gamma(k+\sum_j \alpha_j)}$$

and n_{decoy} and n_{absent} indicate the number of decoy and absent peptides, respectively.

To achieve a non-discretized derivation, we notice that as $k \to \infty$, the Dirichlet process yields a value $log(\Pr(D^{(absent)}|pdf'_{D^{(decoy)}}))$ that converges to the average value (over i) of

$$\frac{1}{\beta(pdf'_{D^{(decoy)}})}pdf'_{D^{(absent)}}(b_i)^{pdf'_{D^{(decoy)}}(b_i)}$$

Thus (without loss of generality),

$$log(\Pr(D^{(absent)}|pdf'_{D^{(decoy)}})) = \frac{1}{s_{high}-s_{low}}\int_{s_{low}}^{s_{high}} \frac{1}{\beta(n_{decoy}pdf'_{D^{(decoy)}})}pdf'_{D^{(absent)}}(s)^{pdf'_{D^{(decoy)}}(s)}\partial s$$

The resulting integral can be computed efficiently in practice by using a standard numeric Riemann sum. The provided npCI software uses a resolution of 100 bins for this summation.

Finally, we compute the symmetric probability that both the absent target PSMs and decoy PSMs are drawn from the same distribution:

$$\Pr(D^{(absent)}|pdf'_{D^{(decoy)}})\Pr(D^{(decoy)}|pdf'_{D^{(absent)}})$$

Note that the number of remaining (*i.e.* absent) peptides will vary for different sets of identified proteins; this will scale all of the exponents in the integral. For this reason, we set the number of remaining peptides (n_{absent}) equal to the number of decoy peptides (n_{decoy}) in the integral. This ensures that both factors receive equal weight.

The free parameter σ (which is used by the smoothing kernel to determine how aggressively the data are smoothed) is estimated automatically by taking the joint maximum likelihood (ML) over σ and the present protein set x. When the two distributions are very similar for some σ_0, the Dirichlet probability is higher than smoothing with $\sigma_1 >> \sigma_0$. Thus the maximum likelihood estimate (MLE) of σ using this scheme allows us to estimate σ without a prior, and guarantees that an appropriate value of σ is used (chosen as the value producing the best likelihood for the best protein set).

The most computationally expensive task performed is smoothing the densities with different values of σ; however, this step can be trivially parallelized in a manner that resembles parallelization of matrix-vector multiplication (which does not require communication between threads). On a machine with several processors (or a processor with several cores), a substantial speedup can be

achieved. Likewise, using a compiled language will provide a large constant speedup compared to the current python implementation. Furthermore, linearity can be exploited to smooth in only the changes to reduce the runtime by an order of magnitude.

Generalization to replicate experiments: The empirical absent score distribution (from decoy PSMs) will only match the score distribution of absent targets from the same experiment; therefore, pooling the scores over multiple distributions would not be appropriate. However, these distributions will be conditionally independent (given the set of present proteins $X = x$). This conditional independence can be exploited, while still respecting the different distributions of decoy scores from different replicates. As a result, we can trivially compute the product over n replicate experiments:

$$\Pr(D|X = x) = \prod_{i}^{n} \Pr(D^{(i)}|X = x)$$

where $\Pr(D^{(i)}|X = x)$ computes (as shown above) the probabilistic divergence between the remaining targets and decoys from replicate i.

Each replicate experiment is assigned it's own estimate of σ with the same methodology used for a data set consisting of only one experiment. This is important because it allows the different data sets to have different amounts of stochasticity and numbers of samples.

This method could also be used to trivially allow simultaneous analysis of multiple peptide

scores.

Generalization to simultaneously utilize multiple scores: The npCI can be easily adapted to include multiple scores for the same peptide (*e.g.* Mascot, PeptideProphet[4,5], XCorr, MS1-based features, etc.). Scores and features that are conditionally independent given the present protein set X can be included in the same manner as replicate experiments. Scores that are correlated even after conditioning on the present protein set can also be included, but should use a joint distribution rather than two separate distributions. Essentially, the joint distribution would treat those multiple scores as a single multidimensional score, which is a tuple composed of the individual scores. In this case, the distribution of remaining peptide scores would no longer be univariate, but would become a heatmap (*i.e.* a multivariate density) constructed from these tuples.

Protein grouping The npCI is, by definition, not altered by protein grouping. Including one protein from an identically connected set will result in the same set of eliminated peptides and the same npCI; therefore, for simplicity, these proteins are grouped using the standard method[6]. In **Figure 1**, protein counts are given using one protein per group (the minumum number that would result in an identical npCI). In the Venn diagrams (**Figure 2**), protein counts must be given ungrouped, because grouping may occasionally be different between the results from the different instruments.

When counting target and decoy identifications or estimating false discovery rate (FDR)[7] on grouped results, proteins groups containing at least one target are counted as a single target, while

protein groups containing only decoys are counted as a single decoy.

When using Mascot scores, edges are added to the protein-to-peptide graph so that the set of all proteins that contain an identified peptide are adjacent. This ensures that results are not influenced by Mascot's "peptide assignment," the process by which shared peptides are assigned to a single adjacent protein.

Peptide scores Regardless of the type of peptide score used (*e.g.* PeptideProphet or Mascot), each peptide is associated with every PSM score for which that peptide is the best-ranking match for the spectrum. This prevents information loss that can occur if only the best score is used for that peptide. For instance, a peptide may match a single spectrum with a fairly high score, but match several other spectra with poor scores. It is useful to distinguish this peptide from another peptide with high-scoring matches to several spectra.

cRAP proteins sp|ALBU_BOVIN|, sp|AMYS_HUMAN|, sp|CAS1_BOVIN|, sp|CAS2_BOVIN|, sp|CASB_BOVIN|, sp|CASK_BOVIN|, sp|CTRA_BOVIN|, sp|CTRB_BOVIN|, sp|K1C10_HUMAN|, sp|K1C15_SHEEP|, sp|K1C9_HUMAN|, sp|K1H1_HUMAN|, sp|K1H2_HUMAN|, sp|K1H4_HUMAN|, sp|K1H5_HUMAN|, sp|K1H6_HUMAN|, sp|K1H7_HUMAN|, sp|K1H8_HUMAN|, sp|K1HA_HUMAN|, sp|K1HB_HUMAN|, sp|K1M1_SHEEP|, sp|K1M2_SHEEP|, sp|K22E_HUMAN|, sp|K2C1_HUMAN|, sp|K2M1_SHEEP|, sp|K2M2_SHEEP|, sp|K2M3_SHEEP|, sp|KRA33_SHEEP|, sp|KRA34_SHEEP|, sp|KRA3A_SHEEP|, sp|KRA3_SHEEP|, sp|KRA61_SHEEP|, sp|KRB2A_SHEEP|, sp|KRB2B_SHEEP|, sp|KRB2C_SHEEP|, sp|KRB2D_SHEEP|, sp|KRHB1_HUMAN|, sp|KRHB2_HUMAN|, sp|KRHB3_HUMAN|, sp|KRHB4_HUMAN|, sp|KRHB5_HUMAN|, sp|KRHB6_HUMAN|, sp|KRUC_SHEEP|, sp|LALBA_BOVIN|,

sp|LYSC_LYSEN|, sp|PEPA_BOVIN|, sp|PEPA_PIG|, sp|PEPB_PIG|, sp|PEPC_PIG|, sp|SSPA_STAAU|, sp|TRY1_BOVIN|, sp|TRY2_BOVIN|, sp|TRYP_PIG|, sp|ALBU_HUMAN|, sp|ANT3_HUMAN|, sp|ANXA5_HUMAN|, sp|B2MG_HUMAN|, sp|BID_HUMAN|, sp|CAH1_HUMAN|, sp|CAH2_HUMAN|, sp|CATA_HUMAN|, sp|CATD_HUMAN|, sp|CATG_HUMAN|, sp|CO5_HUMAN|, sp|CRP_HUMAN|, sp|CYB5_HUMAN|, sp|CYC_HUMAN|, sp|EGF_HUMAN|, sp|FABPH_HUMAN|, sp|GELS_HUMAN|, sp|GSTA1_HUMAN|, sp|GSTP1_HUMAN|, sp|HBA_HUMAN|, sp|HBB_HUMAN|, sp|IGF2_HUMAN|, sp|IL8_HUMAN|, sp|KCRM_HUMAN|, sp|LALBA_HUMAN|, sp|LEP_HUMAN|, sp|LYSC_HUMAN|, sp|MYG_HUMAN|, sp|NEDD8_HUMAN|, sp|NQO1_HUMAN|, sp|NQO2_HUMAN|, sp|PDGFB_HUMAN|, sp|PPIA_HUMAN|, sp|PRDX1_HUMAN|, sp|RASH_HUMAN|, sp|RETBP_HUMAN|, sp|SODC_HUMAN|, sp|SUMO1_HUMAN|, sp|SYH_HUMAN|, sp|TAU_HUMAN|, sp|THIO_HUMAN|, sp|TNFA_HUMAN|, sp|TRFE_HUMAN|, sp|TRFL_HUMAN|, sp|UB2E1_HUMAN|, sp|UBE2C_HUMAN|, sp|UBE2I_HUMAN|, sp|UBIQ_HUMAN|, sp|GAG_SCVLA|, sp|CYC_HORSE|, sp|CAH2_BOVIN|, sp|ADH1_YEAST|, sp|ALDOA_RABIT|, sp|LYSC_CHICK|, sp|MYG_HORSE|, sp|OVAL_CHICK|, sp|BGAL_ECOLI|, sp|DHE3_BOVIN|, sp|GFP_AEQVI|

1. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

2. Käll, L., Canterbury, J., Weston, J., Noble, W. S. & MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923–25 (2007).

3. Granholm, V., Noble, W. S. & Käll, L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of Proteome Research* **10**, 2671–2678 (2011).

4. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Analytical Chemistry* **74**, 5383–5392 (2002).

5. Choi, H. & Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research* **7**, 254–265 (2008).

6. Serang, O. & Noble, W. S. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and Its Interface* **5**, 3–20 (2012).

7. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300 (1995).