TECHNICAL BRIEF

D-score: a search engine independent MD-score

Marc Vaudel[1], Daniela Breiter[1,2], Florian Beck[1], Jörg Rahnenführer[2], Lennart Martens[3,4,*], René P. Zahedi[1]

[1] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany

[2] Department of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany

[3] Department of Medical Protein Research, VIB, Ghent, Belgium

[4] Department of Biochemistry, Ghent University, Ghent, Belgium

* Corresponding author

Keywords: peptide identification, PTM, PTM localization scores

Abstract

While peptides carrying post translational modifications (PTMs) are routinely identified in gel-free mass spectrometry, the localization of the PTMs onto the peptide sequences remains challenging. Search engine scores of secondary peptide matches have been used in different approaches in order to infer the quality of site inference, by penalizing the localization whenever the search engine scored similarly two candidate peptides with different site assignments. In the present work, we show how the estimation of posterior error probabilites (PEP) for peptide candidates allows the estimation of a PTM score called the D-score, for multiple search engine studies.

We demonstrate the applicability of this score to three popular search engines: Mascot, OMSSA and X!Tandem and evaluate its performance using an already published high resolution data set of synthetic phosphopeptides. For those peptides with phosphorylation site inference uncertainty, the number of spectrum matches with correctly localized phosphorylation increased by up to 25.7% when compared to using Mascot alone, although the actual increase depended on the fragmentation method used. Since this method relies only on search engine scores, it can be readily applied to the scoring of the localization of virtually any modification at no additional experimental or *in silico* cost.

Main text

The advent of search engines like Sequest [1], Mascot [2], OMSSA [3] or X!Tandem [4] for the identification of peptides measured by mass spectrometry has enabled the routine identification of millions of spectra [5]. These peptide to spectrum matches (PSMs) are usually filtered at a controlled false discovery rate (FDR) using target/decoy searches [6]. However, while this error rate allows the discrimination of random matches, it does not account for more refined mistakes such as post translational modification (PTM) localization errors [7]. The corresponding false localization rate (FLR) is therefore typically higher than the FDR. In proteomics, the resulting need for modification localization quality control is addressed by two distinct approaches [8]: (A) the use of secondary hits [9-11] and (B) probabilistic scores [9, 12, 13]. Since the latter inspect the spectra themselves, they can find experimental proof of modification localization *via* the detection of so-called site determining ions [9]. The secondary hit based approaches on the other hand have the benefit of being modification independent, and can be easily and quickly computed from search engine results.

Secondary hit scores were first applied for PTM localization with the Sequest delta score, and with Mascot using the normalized delta-ions score [9] further extended to the MD-score [11]. The first modification site score adapted to multiple search engines came with the ‘‘maximum confidence approach’’ [10] which compares site assumptions from Sequest, Mascot, OMSSA and X!Tandem. Here, we present a generic approach for secondary hit scoring by multiple algorithms: first, we translate search engine scores of first and secondary hits in a dataset derived from a multi-search engine analysis into posterior error probabilities (PEP) [14]; secondly, we estimate the PEP difference between the two most likely modification sites in a search engine independent way – thus obtaining a single confidence score named D-score. The performance of this D-score was tested on an already published high resolution dataset of synthetic phosphopeptides [11].

Synthetic peptides dataset

Raw files were downloaded from the Tranche database and transformed into peak lists in the mgf format using Proteome Discoverer (Version 1.3.0.339, Thermo Scientific). MS1 peaks were chosen for peak picking using peptides with charge states ranging from 2+ to 5+, and with masses from 300 to 5000 Da. The resulting peak lists were searched with Mascot version 2.3.2 as well as with OMSSA [3] version 2.1.9 and X!Tandem [4] 'Cyclone' (2010.12.01.1) *via* SearchGUI [15] version 1.8.9.

All searches were conducted against a concatenated target/decoy [6] database derived from the human complement of the UniProtKB/Swiss-Prot database [16] (downloaded on 4th of November 2010, containing 20,260 target sequences), the decoy sequences were obtained within SearchGUI by reversing the target sequences. Search settings were: a maximum of two allowed missed cleavages, peptide charges from 2+ to 4+, a peptide mass tolerance of 10 ppm, a fragment ion mass tolerance of 0.5 Da for CID, ETD and MSA data, and of 0.02 Da for HCD data, carbamidomethylation of Cys (+57.0214 Da) as fixed, and phosphorylation of Ser, Thr and Tyr (+79.9663 Da) as well as oxidation of Met (+15.9949 Da) as variable modifications. For ETD data, *c* and *z* ions were searched for, while *b* and *y* ions were searched for in all other spectra – for Mascot searches the respective instrument setting was used. All other OMSSA and X!Tandem settings were kept at the default values provided by SearchGUI.

For every spectrum analysed, each search engine returns a set of peptide candidates, so-called peptide to spectrum matches (PSMs). Mascot, OMSSA and X!Tandem data were parsed using the MascotDatfile [17], OMSSA [18] and XTandem [19] parsers, respectively, and the PSM with the lowest e-value per spectrum was retained. The search engine e-value of the best peptide candidate was rescored into a posterior error probability (PEP) as canonically done in proteomics by comparing the score distributions of target and decoy matches [14, 20].

It is here important to note that the correctness of the PEP estimation – and subsequently of the D-score – is heavily dependent on the search engines used which should be complying with independence requirements inherent to the use of target decoy databases [21]. The use of multi-stage search strategies can thus impair the accuracy of the scoring as known from the literature [22]. The accuracy of the PEP estimation can be verified by the use of samples of known contents[23-25], this is however not the scope of the present study.

All identification results are freely available in PRIDE [26] (accessions 27179-27198) as a ProteomeXchange data set (http://www.proteomexchange.org, accession PXD000021).

PSMs were sorted by increasing PEP and were filtered to an FDR of 1%. Subsequently, only those synthetic peptides with multiple phosphorylatable amino acid residues (here Ser, Thr, Tyr) were retained. Moreover, when two different phosphorylation isoforms were known to be present in the same mix, the peptide was ignored.

Secondary hits

As displayed in Figure 1A with spectrum matches of Mascot, OMSSA and X!Tandem filtered at 1% FDR each, different search engines identify different peptides; they thus clearly present the benefit of complementarity. For every search engine, we estimated a posterior error probability (PEP) for the best hit of every spectrum by comparing the target and decoy distributions: when considering 100 PSMs at a PEP of 5%, one expects only five false positives. Subsequently, we attached to every secondary hit the PEP corresponding to its score; i.e., the PEP it would have had if it had been the best candidate for the spectrum.

Secondary hit based PTM localization scores rely on the rationale that whenever the search engine is unsure about a modification site, it will return the second best position as the secondary hit. For every spectrum leading to the identification of a modified peptide, the PEP of the best secondary hit with the same sequence but another modification site minus the PEP of the first hit is thus a search-engine independent metric for the certainty of the site assignment.

Assuming that the PEP of the highest e-value reached by the search engine corresponds to a truly random hit, *i.e.* a PEP of 1, a secondary PEP of 1 is taken whenever no secondary hit is found in the search engine results. It is worth noting that this hypothesis might lead to slightly optimistic results since the probability of a match is never truly zero.

We applied this approach to the synthetic phosphopeptide dataset. Here, phosphorylation sites are known *a priori*, making it possible to verify the number of true and false localizations introduced at decreasing scores, as displayed in Figure 1B. It is clear that the different search engines will return different number of PSMs with confidently localized modifications: OMSSA outperforms Mascot whereas X!Tandem provides quantity rather than quality. Interestingly, excluding matches presenting a D-score equal to 0 – i.e. where the best secondary hit and the retained peptide present the same PEP – allows decreasing the FLR significantly for Mascot and OMSSA whereas no such hit is present for X!Tandem alone.

Multiple search engine approach

In the previous example, all search engines were considered separately. In order to combine the results from the different search engines, a single PSM score was given to every peptide candidate by computing the product of the PEPs given by the different search engines. The peptide with the lowest PSM score was retained as best candidate and whenever two peptides scored equally, the number of search engines supporting a candidate was used to discriminate between them. As displayed in Figure 1C, combining results from different search engines provides an extended list of peptide candidates *per* spectrum. The identification process thus benefits from much more information to decide on the modification location: the hits from one search engine can be used to confirm, question or even correct the results of another.

The number of PSMs validated at 1% FDR is also substantially enhanced by combining the search engines (see Figure 2A).

For every spectrum, we can thus estimate the D-score based on the PEPs of secondary hits. As before, the peptide candidate PEP is subtracted from the PEP of the best secondary hit presenting the same sequence but a different modification site. This score has the important advantage to be, by design, directly comparable between search engines. When working with a single search engine, it is moreover similar to the Mascot-MD score [11]. When working with multiple search engines, as illustrated in Figure 2B, however, a substantial gain in the number of PSMs with correctly localized modifications can be obtained.

Among the CID PSMs at 1% FDR, the number of spectra matching a peptide with different potential phosphosites increased from 443 to 557 when using the combined three search engines instead of only Mascot. The number of these matches was thus increased by 25.7% at a stable FLR (actually going down from 15.4% to 12.8%). Here again, excluding the matches presenting a D-score equal to 0 substancially reduced the FLR. A similar increase was observed with other fragmentation methods used on the synthetic phosphopeptide dataset (HCD: +4.3%, MSA: +20.4%, ETD: +8.0%) while the FLR always stayed at the same order of magnitude (decreasing from 23.7% to 20.8% for HCD spectra, and increasing slightly from 15.4% to 16.7% for MSA spectra and from 4.6% to 7.6% for ETD spectra).

Discussion

This study demonstrates how modification-oriented studies can benefit from the systematic use of multiple search engines for both identification as well as modification localization. The estimation of the posterior error probability (PEP) for every match makes it possible then to estimate a search engine independent score for modification site scoring: the D-score. This score has the advantage of being search engine independent and directly computable from the search engine results without requiring additional spectrum processing.

Computing the score on data obtained with different fragmentation methods showed consistent improvement but strong variations of the FLR at a given score as further illustrated in Figure 2C. Like with all other PTM scores, setting a threshold for a given experiment thus remains problematic. Since the accuracy of the search engine scores can be monitored at the FDR threshold, we propose to set a D-score threshold at the PEP reached at the FDR threshold in order to enhance the modification localization stringency. As displayed in Figure 2D and detailed in Table 1, this adaptive threshold allows the setting of a default limit for the D-score while dramatically reducing the FLR of the result set.

Acknowledgments

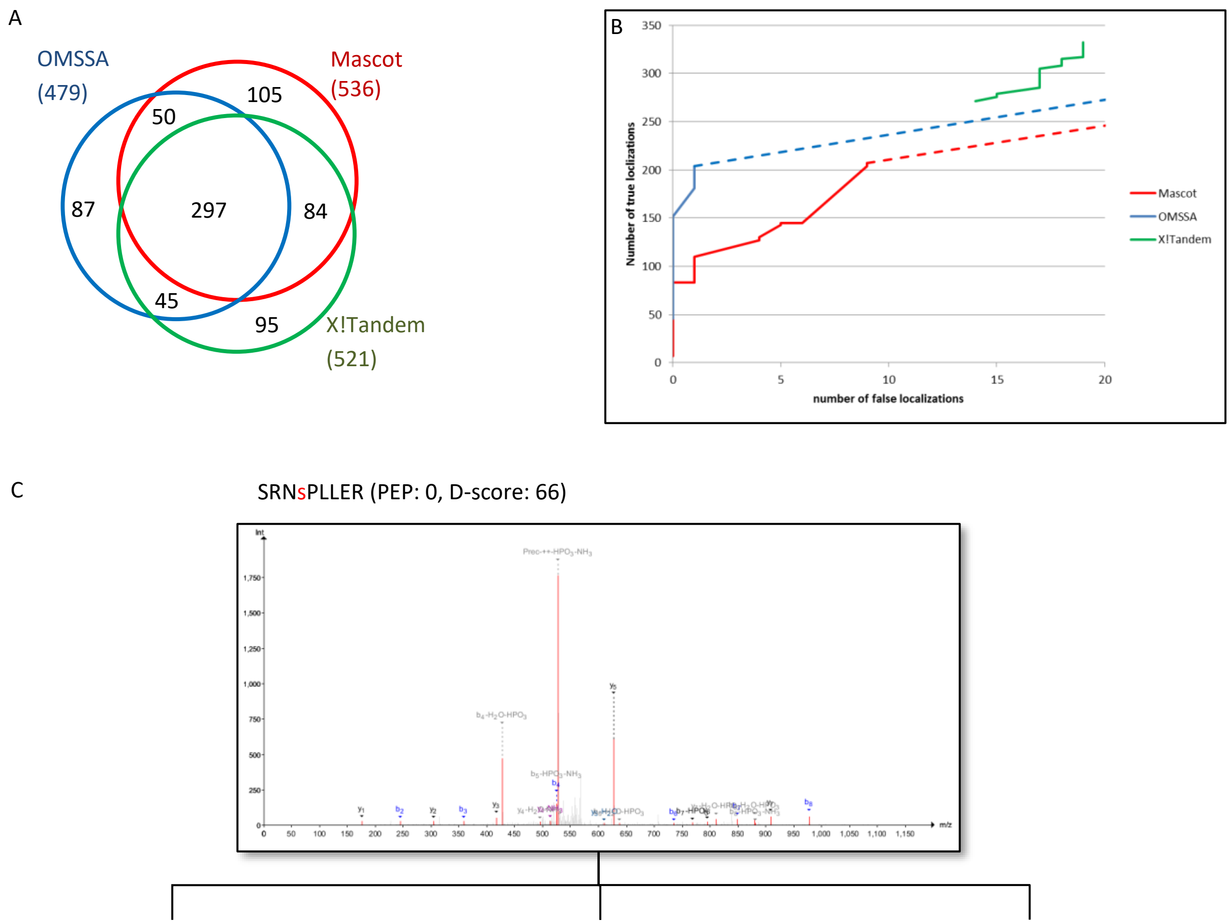
The financial support by the Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen and by the Bundesministerium für Bildung und Forschung (SARA, DYNAMO) is gratefully acknowledged. L.M. acknowledges the financial support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and the PRIME-XS project funded by the European Union 7th Framework Program under grant agreement number 262067.

Competing financial interests

The authors declare no competing financial interests

Figures

Figure 1:



Mascot			
#	PEPTIDE	e-value	PEP
1	ALELsWQPK	0.12	6%
2	SRNsPLLER	0.367	19%
6	sRNSPLLER	3.999	66%

OMSSA			
#	PEPTIDE	e-value	PEP
1	ALELsWQPK	11.435	74%

X!Tandem			
#	PEPTIDE	e-value	PEP
1	SRNsPLLER	0.005	0%

Figure 2:

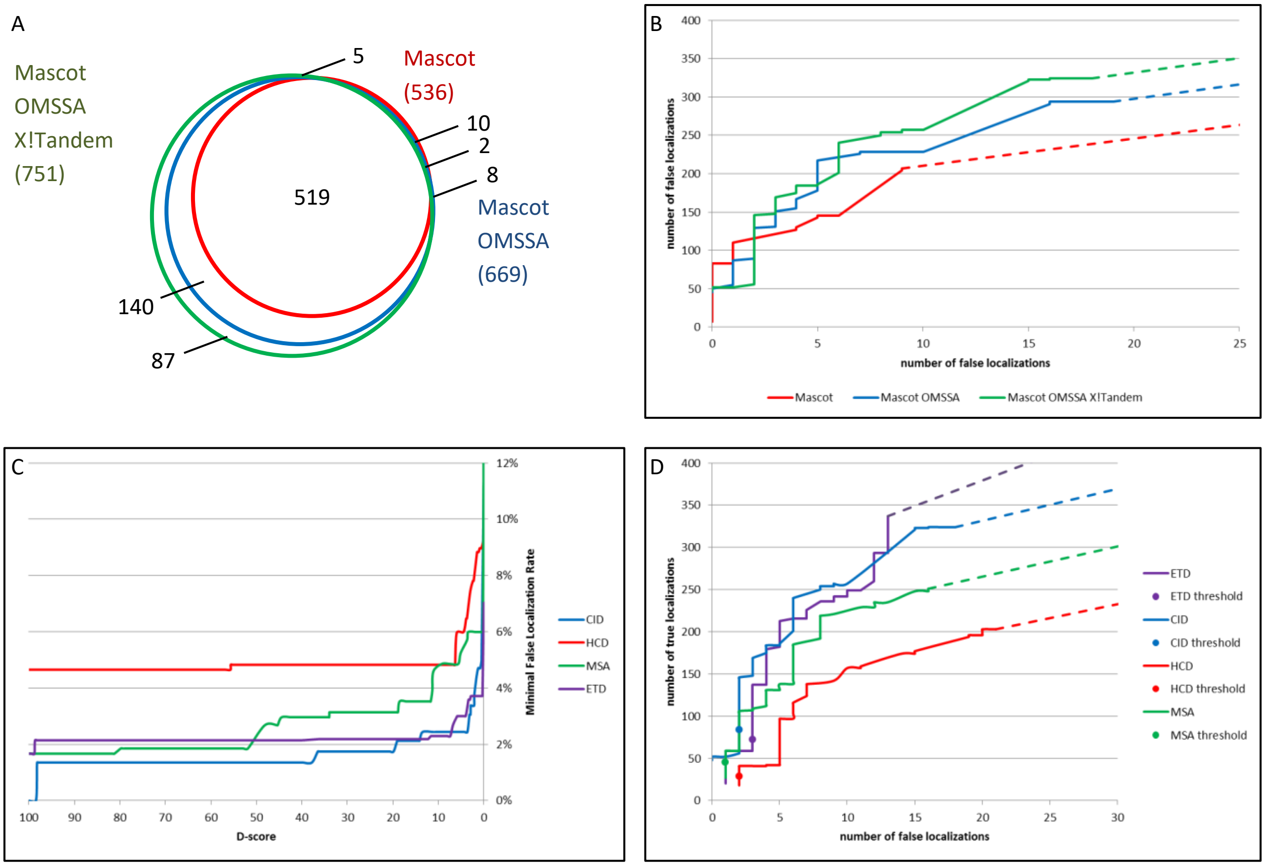

A
Mascot
OMSSA
X!Tandem
(751)
5
Mascot
(536)
10
2
8
519
Mascot
OMSSA
(669)
140
87
B
number of false localizations
number of false localizations
0
50
100
150
200
250
300
350
400
0
5
10
15
20
25
Mascot
Mascot OMSSA
Mascot OMSSA X!Tandem
C
D-score
100
90
80
70
60
50
40
30
20
10
0
Minimal False Localization Rate
0%
2%
4%
6%
8%
10%
12%
CID
HCD
MSA
ETD
D
number of true localizations
number of false localizations
0
50
100
150
200
250
300
350
400
0
5
10
15
20
25
30
ETD
ETD threshold
CID
CID threshold
HCD
HCD threshold
MSA
MSA threshold

Figure Legends

Figure 1: **Secondary hits and PTM scoring for multiple search engine approaches.** (A) Venn diagram of the identifications obtained from different search engines (Mascot, OMSSA and X!Tandem) separately. Here obtained on the synthetic, phosphorylated peptides spectra obtained after CID fragmentation [11] at 1% false discovery rate (FDR). While all three search engines perform similarly, they present complementary spectrum coverage (B) On the same identification matches, after ordering the hits with potential phosphosite uncertainty at decreasing difference between the score of the best secondary hit with the same sequence but another modification site and the score of the best peptide, the number of true localizations is plotted against the number of false localizations. Here, OMSSA performs best, followed by Mascot. X!Tandem returns many possible peptides but with a considerable share of false localizations. Note that hits with a D-score equal to zero are represented with a dashed line. Excluding these reduces the False Localization Rate to 4.2% for Mascot and 0.5% for OMSSA. (C) At the end of the identification process, for each spectrum, different possible PSMs from three search engines are available. Every PSM comes with an e-value from which we estimate the probability that the search engine made a mistake: the posterior error probability (PEP). The best candidate for this spectrum, *i.e.* presenting the lowest PEP, is the peptide SRNsPLLER, identified by X!Tandem and Mascot, with a PSM-level PEP of 0%. The second best possibility for this peptide with another modification site comes at rank 6 in the Mascot results with a PEP of 66%. The D-score thus reaches: 66% - 0% = 66%.

Figure 2: **Performance of PTM scoring and estimating a threshold.** (A) Venn diagram showing the combined results from the CID dataset for three different search engines as obtained when processing the spectra with Mascot alone (red), Mascot combined with OMSSA (blue) and Mascot combined with OMSSA and X!Tandem (green). Note that the number of PSMs at 1% FDR increases substantially compared to the analysis where each search engine is considered separately (see Figure 1A): a 25% gain is achieved when adding OMSSA results to Mascot, 40% when adding OMSSA and X!Tandem. (B) Chart with the performance

effect of incrementally combining search engines, starting with Mascot alone, then adding OMSSA results, and finally X!Tandem results. Here on the same dataset, PSMs for which the PTM site must be inferred are sorted by decreasing D-score and the number of correct localizations is plotted against the number of incorrect localizations. Notably, the performance increases with each added search engine. Here, excluding hits with a D-score equal to zero (dashed line) reduces the False Localization Rate (FLR) to 4.2% for Mascot alone, 6.1% when adding OMSSA and 5.3% when using all three search engines. (C) For those PSMs validated at 1% FDR but requiring PTM site inference, when comparing the minimal false localization rate (FLR) achievable at a given D-score, the yield varies considerably between fragmentation methods, as displayed here for CID (blue), HCD (red), MSA (green) and ETD (purple). This effect was already observed for the MD-score [11]. (D) For these PSMs ordered by decreasing D-score, the number of hits presenting a correct localization is plotted against the number of incorrect localizations for the different fragmentation methods CID (blue), HCD (red), MSA (green) and ETD (purple). Here, excluding hits with a D-score equal to zero, illustrated by the dashed line reduces the FLR to 5.3% for CID spectra, 9.4% for HCD spectra, 6.0% for MSA spectra and 3.7% for ETD spectra. The performance is again clearly fragmentation dependent, however, when thresholding the lists at a D-score equal to the PEP corresponding to the FDR threshold, illustrated here by the points, the stringency is increased and the FLR reduced as detailed in Table 1.

Table

Table 1:

Fragmentation	# PSMs total	FLR total	D-score threshold	# PSMs	FLR
CID	557	12.8%	85.45%	86	2.3%
HCD	555	20.8%	80.51%	31	6.5%
MSA	560	16.7%	88.37%	46	2.2%
ETD	512	7.6%	88.62%	75	4.0%

Table Legend

Table 1: Validating Peptide to Spectrum Matches (PSMs) from different search engines at a 1% False Discovery Rate (FDR) retains all PSMs with a high identification confidence – where the confidence is the complement of the PEP: 1-PEP. Here, 557 PSMs with multiple possible modification sites were retained for CID spectra, accounting a False Localization Rate (FLR) of 12.8%. The confidence corresponding to this threshold is 85.45% (PEP of 14.55) for CID spectra, 80.51%, 88.37% and 88.62% for HCD, MSA and ETD spectra, respectively. Matches with a lower confidence will not be validated: we propose to use this value for the D-score threshold. The PSM selection is more stringent, for example, only 86 out of 557 PSMs were retained for CID spectra. The FLR is subsequently substantially reduced, from 12.8% to 2.3% for CID spectra; from 20.8% to 6.5% for HCD spectra, from 16.7% to 2.2% for MSA spectra and from 7.6% to 4% for ETD spectra.

References

[1] Yates, J. R., 3rd, Eng, J. K., McCormack, A. L., Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 1995, *67*, 1426-1436.
[2] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551-3567.
[3] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L.*, et al.*, Open mass spectrometry search algorithm. *J Proteome Res* 2004, *3*, 958-964.
[4] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466-1467.
[5] Burkhart, J. M., Vaudel, M., Gambaryan, S., Radau, S.*, et al.*, The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* 2012.
[6] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007, *4*, 207-214.
[7] Colaert, N., Degroeve, S., Helsens, K., Martens, L., Analysis of the resolution limitations of peptide identification algorithms. *J Proteome Res* 2011, *10*, 5555-5561.
[8] Chalkley, R. J., Clauser, K. R., Modification site localization scoring: strategies and performance. *Mol Cell Proteomics* 2012, *11*, 3-14.
[9] Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 2006, *24*, 1285-1292.
[10] Beck, F., Lewandrowski, U., Wiltfang, M., Feldmann, I.*, et al.*, The good, the bad, the ugly: validating the mass spectrometric analysis of modified peptides. *Proteomics* 2011, *11*, 1099-1109.
[11] Savitski, M. M., Lemeer, S., Boesche, M., Lang, M.*, et al.*, Confident phosphorylation site localization using the Mascot Delta Score. *Mol Cell Proteomics* 2011, *10*, M110 003830.
[12] Olsen, J. V., Blagoev, B., Gnad, F., Macek, B.*, et al.*, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006, *127*, 635-648.
[13] Taus, T., Kocher, T., Pichler, P., Paschke, C.*, et al.*, Universal and confident phosphorylation site localization using phosphoRS. *J Proteome Res* 2011, *10*, 5354-5362.
[14] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 2010, *73*, 2092-2123.
[15] Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., Martens, L., SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 2011, *11*, 996-999.
[16] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C.*, et al.*, UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004, *32*, D115-119.
[17] Helsens, K., Martens, L., Vandekerckhove, J., Gevaert, K., MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* 2007, *7*, 364-366.
[18] Barsnes, H., Huber, S., Sickmann, A., Eidhammer, I., Martens, L., OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics* 2009, *9*, 3772-3774.
[19] Muth, T., Vaudel, M., Barsnes, H., Martens, L., Sickmann, A., XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* 2010, *10*, 1522-1524.
[20] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002, *74*, 5383-5392.
[21] Barboza, R., Cociorva, D., Xu, T., Barbosa, V. C.*, et al.*, Can the false-discovery rate be misleading? *Proteomics* 2011, *11*, 4105-4108.
[22] Everett, L. J., Bierl, C., Master, S. R., Unbiased statistical analysis for multi-stage proteomic search strategies. *J Proteome Res* 2010, *9*, 700-707.

[23] Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J.*, et al.*, The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J Proteome Res* 2008, *7*, 96-103.
[24] Granholm, V., Noble, W. S., Kall, L., On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *J Proteome Res* 2011, *10*, 2671-2678.
[25] Vaudel, M., Burkhart, J. M., Breiter, D., Zahedi, R. P.*, et al.*, A complex standard for protein identification, designed by evolution. *J Proteome Res* 2012.
[26] Martens, L., Hermjakob, H., Jones, P., Adamski, M.*, et al.*, PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537-3545.