Sensitive and specific spectral library searching with COSS and Percolator

Genet Abay Shiferaw [1,2], Ralf Gabriels [1,2], Elien Vandermarliere [1,2], Lennart Martens [1,2], Pieter-Jan Volders [1,2,3]

[1] VIB-UGent Center for Medical Biotechnology, VIB, 9000 Ghent, Belgium

[2] Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium

[3] Cancer Research Institute Ghent, Ghent University, 9000 Ghent, Belgium

Corresponding Author

*Prof. Dr. Lennart Martens, Technologiepark-Zwijnaarde 75, 9052 Ghent, Belgium.

E-mail: lennart.martens@vib.ugent.be , Tel: +32 92249840

ABSTRACT: Maintaining high sensitivity while limiting false positives is a key challenge in peptide identification from mass spectrometry data. Here, we therefore investigate the effects of integrating the machine learning-based post-processor Percolator into our spectral library searching tool COSS. To evaluate the effects of this post-processing, we have used twenty data sets from two different projects and have matched these against the NIST spectral library. The matching is carried out using two performant spectral library search engines (COSS and MsPepSearch), both with and without Percolator post-processing, and using sequence database

search engine MS-GF+ as a baseline comparator. The addition of the Percolator rescoring step was particularly effective for COSS, resulting in a substantial improvement in sensitivity and specificity of the identifications. Importantly, the false discovery rate was especially strongly affected, resulting in much more reliable results. COSS is freely available as open source under the permissive Apache2 license, and binaries and source code are found at https://github.com/compomics/COSS

Keywords: spectral library, peptide identification, rescoring, COSS, percolator, mass spectrometry

Introduction

MS-based peptide identification typically relies on matching measured spectra against theoretical spectra in a database searching approach[1]. However, identification can also be obtained by matching measured spectra against a library of previously measured and identified spectra[2]. Several spectral library searching tools have been developed for this purpose, with notable examples including SpectraST[3], the National Institute of Standards and Technology(NIST) MS-Search[4], MSPepSearch[a], ANN-SoLo[5], X!Hunter[6], and COSS[7]. The direct comparison of a newly measured spectrum against the spectra in such a spectral library can both increase sensitivity[8] and reduce computational complexity compared to database searching[9].

However, similar to database searching, spectral library searching is not perfect, and several causes can lead to incorrect peptide identification[10]: quality of the spectral data, unexpected post translational modifications, charge state issues, or a poor scoring function. To control this erroneous identification, the validation of search results is a crucial step in the identification process. Typically, this is handled through a target-decoy approach in both database and spectral library searches[1]. This approach allows estimation of the False Discovery Rate (FDR)[11,12], which is the expected proportion of incorrect peptide to spectrum matches (PSMs) among the selected set of accepted identifications.

Besides the validation of PSMs using target-decoy based FDR control, it has also become common to employ post-processing methods to database search results to increase its sensitivity. The most popular of these is Percolator[13], which significantly improves the sensitivity of multiple database search engines, including SEQUEST, Mascot[14], and MS-GF+[15]. Percolator itself is a semi-supervised machine learning algorithm based on a linear support vector machine, which is designed to discriminate between correct and incorrect peptide matches by rescoring peptide identifications. For this, Percolator considers a set of features that describe each PSM and uses these features as well as the annotation of target and decoy PSMs in an iterative process to re-rank PSMs using a new score and associated q-value. Yet, although Percolator is commonly used in database searches, its use has thus far not been reported for spectral library searching.

---

[a] https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch

Here, we have therefore examined the utility of Percolator in spectral library searching by extending our COSS spectral library search tool. Importantly, we have found that Percolator improves the output of COSS in two ways. First, Percolator rescoring increases the total number of identifications obtained. Second, Percolator use also results in better controlled FDR. We have moreover compared the effect of Percolator on COSS output with that of Percolator on a different spectral library search tool, MSPepSearch. This comparison shows that Percolator controls FDR better for COSS than for MsPepSearch, while also yielding more identifications for COSS.

We therefore integrated Percolator in the latest version of our freely available, open-source COSS tool, which is moreover capable of handling multiple file formats, and can also be used to analyze large data sets.

MATERIAL AND METHODS

Experimental data sets and spectral library

We obtained raw data files from eleven runs from the Human Proteome Map[16] (ProteomeXchange[17] ID PXD000561) and nine runs from the deep proteome and transcriptome abundance atlas[18] data set (ProteomeXchange ID PXD010154) as benchmarking data sets (Supplementary Table S-1). All these twenty raw files were converted to Mascot Generic Format (mgf) format using the msconvert tool (ProteoWizard[19] version 3.0.19014), with the peak picking algorithm activated. The well-known NIST spectral library[b] was used to perform our searches against.

Estimating FDR and FDP

COSS uses a reverse sequence decoy generation technique to generate decoy sequences as described by Zhang *et al.*[20]. This method is described and benchmarked in detail in previous work[7]. After running the data sets against concatenated decoy libraries, the corrected FDR is calculated as described previously by Sticker *et al.*[12].

To evaluate how re-rescoring with Percolator affects the number of false positive identifications, we used a modified entrapment method[21]. For this, we obtained a tandem mass spectrometry data set of *Pyrococcus furiosus*[21] (ProteomeXchange[17] ID PXD001077) which contains a total of 15,615 spectra. Unlike the original entrapment method, where *Pyrococcus furiosus* spectra are appended to the library spectra, we have appended these to our query spectra and then ran the search against the library concatenated with the generated decoy spectra. The false discovery proportion (FDP), which serves as a measure of the actual FDR, is then calculated as follows:

$$\frac{\frac{\#target_{Pyrococcus}}{f}}{\#target_{Sample}}$$

[b] https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503

With *f* the fraction of *Pyrococcus furiosus* spectra over non-*Pyrococcus* spectra in the search. The effect of rescoring with Percolator on the actual FDR can then be estimated by comparing the FDP before and after rescoring.

Spectral library searches

All twenty data sets are searched with COSS against the NIST spectral library, which is appended with the corresponding decoy spectra. The following search settings were used: precursor mass tolerance set to 10 ppm, fragment mass tolerance set to 0.05 Dalton (Da), MSROBIN scoring function[7], and a fragment mass window to select peaks set to10 Da.

To run MSPepSearch (version 0.96), we first created an MSPepSearch readable spectral library using the COSS decoy generator. Next, we converted the MSP file of the generated library to MSPepSearch's binary file format using Lib2NIST (version 1.0.6.5) and performed the searches using a precursor mass tolerance of 10 ppm and fragment mass tolerance 0.05 Da.

To compare the FDR and FDP, we also searched each of the twenty data sets concatenated with *Pyrococcus furiosus* spectra using the same search settings as described above for COSS and MSPepSearch.

After obtaining all search results from COSS, MSPepSearch, and MSGF+ (see section below on sequence database searches), all results were rescored with Percolator using the features described in supplementary Table S-2.

Sequence database search

In parallel with the spectral library searches, we performed a sequence database search with the well-established MS-GF+[22] (version v2018.04.09) tool through SearchGUI[23] (version 3.3.16) and PeptideShaker[24] (version 1.16.42). The search database is constructed from the human proteome (UP000005640) as obtained from UniProt[25] (consulted on 9/10/2018). Carbamidomethylation of cysteine and oxidation of methionine are used as fixed, and as variable modification respectively. Trypsin is used as protease and a maximum of two missed cleavages was allowed. Precursor mass tolerance was set to 10 ppm, and fragment mass tolerance to 0.05 Da. Precursor charges from 2 to

4 are considered. The SearchGUI configuration file used to run MS-GF+ is provided in Supplementary methods.

Percolator rescoring

We integrated Percolator[13] version v3-04 in COSS[7] such that Percolator's input features are automatically generated from COSS results and Percolator can be executed automatically on the results. The calculated features are: PSM title, score, Library (whether the PSMs is a decoy or not), scan number, retention time, precursor mass, precursor charge, modification, number of matched peaks between query and matched library spectrum, sum of matched peak intensities in the query spectrum, sum of matched peak intensities in the library spectrum, peptide sequence, and protein. All features are generated by COSS and are output as a tab-separated text file which will be saved in the working directory.

For MSPepSearch, Percolator was provided with the full list of available features: Title, label, scan number, Score, probability, peptide, protein. The full list of features used for all tools, including MS-GF+, are found in supplementary Table S-2.

RESULTS AND DISCUSSION

Effects of different feature sets on the rescored result

Percolator uses a machine learning algorithm to rescore the output of peptide identification tools to achieve higher sensitivity and specificity. As with all machine learning algorithms, the result of this process depends strongly on the quality and utility of the features provided to Percolator. To examine the effect of available features on rescoring in spectral library searching, we analyzed different combinations of output features with Percolator. The full set of features consists of: PSM title, score, Library (whether the PSMs is a decoy or not), scan number, retention time, precursor mass, precursor charge, modification, number of matched peaks between query and matched library spectrum, sum of matched peak intensities in the query spectrum, sum of matched peak intensities in the library spectrum, peptide sequence, and protein. When all these features are used on COSS results, the identification rate at 1% FDR is 28% in PXD000561, and 13.6% in PXD010154 (Figure 1). This marks a slight improvement over the non-rescored results of 1% in PXD000561, and of 0.4% in PXD010154. When features are omitted, the worst performance is observed when the MSROBIN score is omitted from the feature set, resulting in only 23.4% identifications in PXD000561, and 10.9% in PXD010154 (Figure 1). This seems to indicate that the remaining features are incapable of compensating for the score omission, which is expected to be a key parameter[26], with total identification performance. By contrast, omitting intermediate score elements of the COSS scoring function (total matched intensity in query spectrum, total matched intensity in library spectrum, and number of matched peaks) still increases the total identification rate by 0.6% in PXD000561, and by 0.3% in PXD010154 (Figure 1). The precursor charge, protein and modification features appear the least relevant, as their exclusion has the smallest negative effect on the rescoring when compared to the inclusion of all features (Figure 1). These results emphasize the importance of the compound score feature for COSS in Percolator rescoring, an importance that cannot be compensated for by providing only the intermediate score elements.

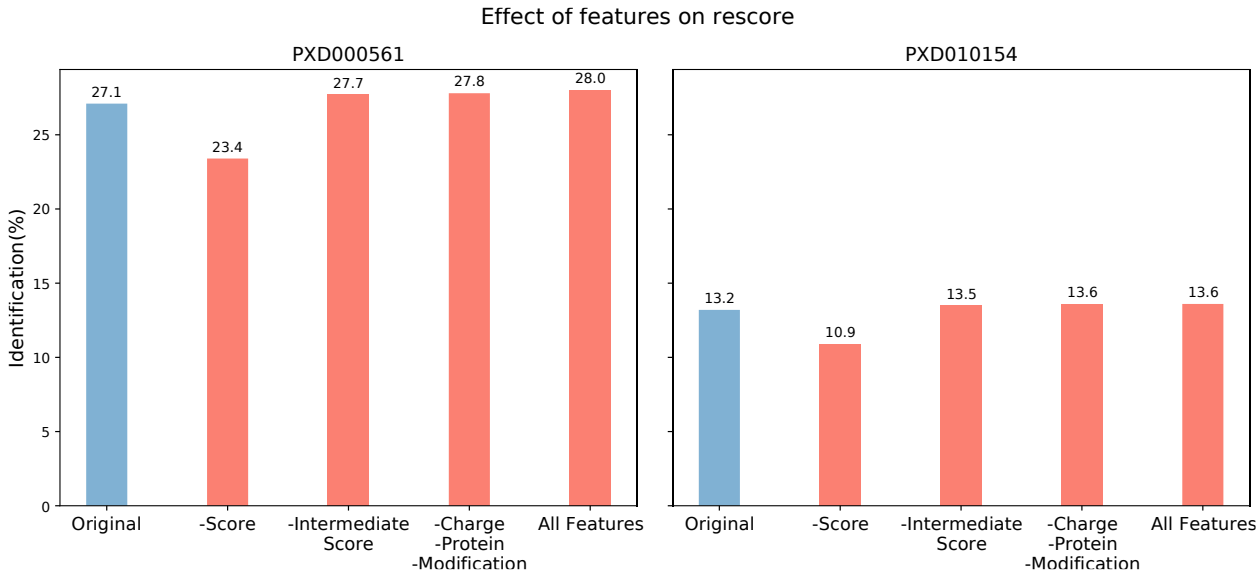

Figure 1. Effect of different feature sets provided to Percolator as measured by identification rate at 1% FDR, and this for two distinct data sets.

Comparison of COSS, MSPepSearch and MS-GF+ with and without Percolator

To put the improvement observed in rescoring COSS results into context, we compared these with rescoring results from MSPepSearch, a different but performant spectral library search tool, and MS-GF+, a popular database search engine. Figure 2 shows the identification rates obtained either at 1% FDR (without Percolator), or for a q-value at or below 0.01 (with Percolator) for each of the two test data sets as analyzed by COSS, MSPepSearch, and MS-GF+. For PXD000561, COSS consistently outperforms both MSPepSearch and MS-GF+, while for PXD010154, MS-GF+ obtains a slightly higher identification rate. The latter can likely be attributed to incomplete spectral coverage of the NIST library in this case, given the even lower performance of the MSPepSearch tool on this data set. Interestingly, the effect of applying Percolator on the results yields a roughly equal increase in identification rate for COSS and MS-GF (around 1% for PXD000561, and 0.5% for PXD010154), but roughly double that for MSPepSearch (1.8% for PXD000561, 1.1% for PXD010154).

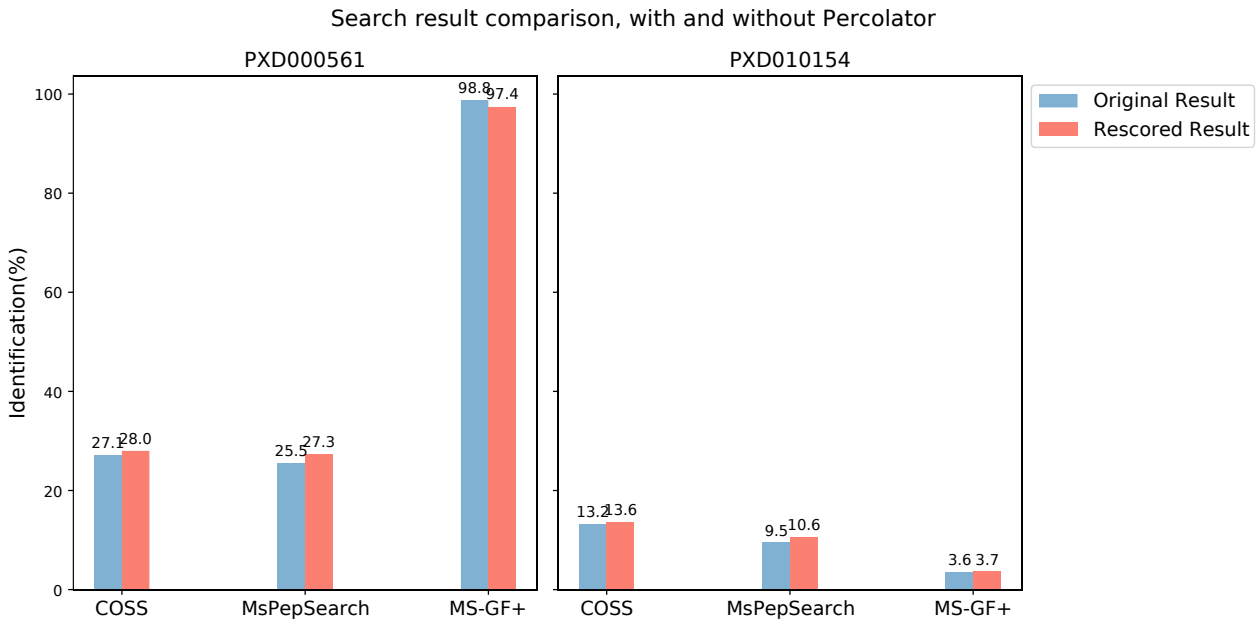

Figure 2. Comparison of identification rates achieved by COSS, MSPepSearch and MS-GF+, both without (blue) and with (red) Percolator. All results are either taken at 1% FDR (without Percolator; blue) or at a q-value cut-off at 0.01 (with Percolator; red).

Comparison of obtained identifications before and after Percolator rescoring

To gain further insight into the results obtained with and without Percolator rescoring, identification overlap at 1% FDR before and after rescoring was analyzed at the PSM level. Figure 3 shows this identification agreement for COSS with and without Percolator. For both data sets, Percolator removes some of the original COSS results, while adding a larger set of new results. The vast majority of the identifications is maintained, however. Supplementary Figure S-1 shows these plots for MSPepSearch and MSGF+, obtaining very similar results overall, but with a relatively larger number of added identifications for MSPepSearch for each data set.

Figure 3. Overlap in search results at the 1% FDR level from COSS before and after Percolator rescoring.

Evaluation of FDR control after Percolator rescoring through an inverted *Pyrococcus furiosus* search

The previous sections showed that the application of Percolator on the results of all tested identification engines improves the total number of PSMs, thus boosting sensitivity. However, it is also important to evaluate the effect of rescoring on the ability to control FDR. We therefore performed an inverted *Pyrococcus furiosus* search as described previously[7]. Briefly, by appending MS/MS spectra obtained from *Pyrococcus furiosus* samples to the query set, a false discovery proportion (FDP) can be calculated from the relative amount of identified *Pyrococcus* spectra, which serves as an estimation of the true FDR of the result set. As shown in Figure 4, the overall underestimation of true FDR by COSS is clearly strongly improved after rescoring. The effect for MSPepSearch is much more limited, however, with essentially no impact in the typically used range (q-value up to 0.05). Due to this difference, COSS manages to outperform MSPepSearch in terms of specificity by a substantial margin after Percolator rescoring. This discrepancy in the utility of Percolator to enhance specificity may be attributed to the higher number of relevant Percolator input features for COSS than for MSPepSearch.

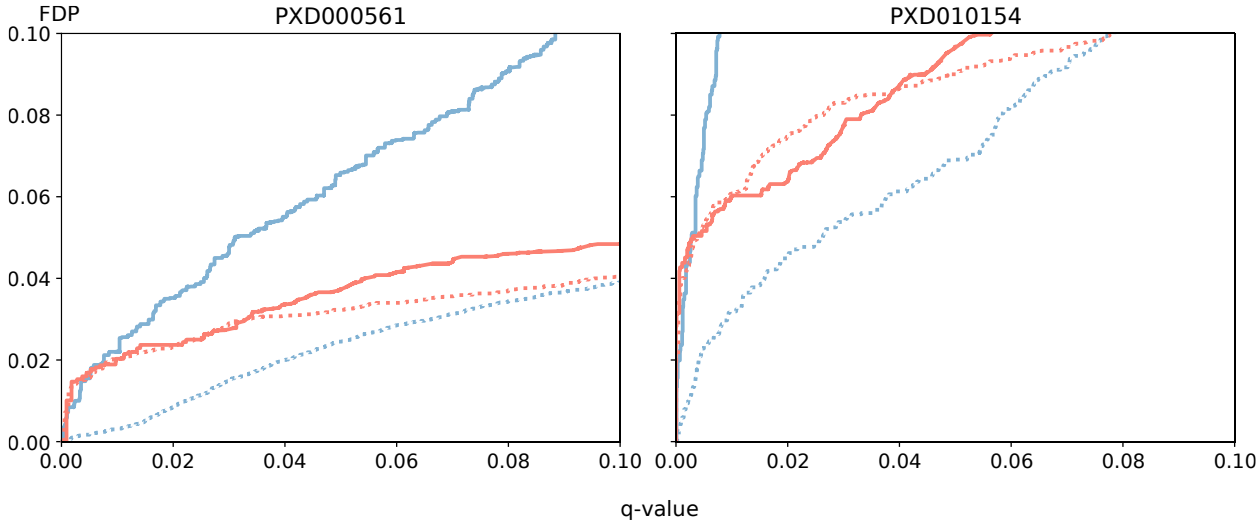

Figure 4. FDP plotted against q-value for COSS and MSPepSearch, without Percolator (full line) and with Percolator (dotted line) for both data sets. The results shown are based on eleven aggregated data sets for PXD000561, and on nine aggregated data sets for PXD010154.

CONCLUSION

Increasing sensitivity while maintaining specificity is key in peptide and protein identification from mass spectrometry data. Rescoring peptide identifications using tools such as Percolator to increase sensitivity and specificity is thus common practice in sequence database searching. However, until now it was unclear if such post-processing had any benefits for spectral library searching. Here, we have shown that combining Percolator with such tools enhances sensitivity, and that it can also enhance specificity. Specifically, our COSS spectral library search tool shows increased sensitivity, and dramatically enhanced specificity when combined with Percolator. We have thus shown that, for COSS at least, the combination of these benefits justifies the added complexity of the percolator post-processing step. Based on these findings, we can thus recommend use of rescoring in spectral library searching. To this end, the latest version of COSS has been fitted with Percolator integration, allowing it to be used by default through the graphical user interface.

AVAILABILITY

The COSS software and its source code can be freely downloaded from https://github.com/compomics/COSS and is licensed under the permissive, open-source Apache License, version 2.0.

ACKNOWLEDGMENTS

This project is supported by the National Institute of Health (NIH) [NCI-ITCR grant number 1U24CA199347 to G.A.S.], Research Foundation - Flanders (FWO) [grant number 1S50918N to R.G., grant number G042518N to L.M., and grant number 1253321N to P.V.], and by the European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to LM]. The authors would like to thank all CompOmics group members for their ideas, discussions and support.

REFERENCES

(1) Verheggen, K.; Ræder, H.; Berven, F. S.; Martens, L.; Barsnes, H.; Vaudel, M. Anatomy and Evolution of Database Search Engines — a Central Component of Mass Spectrometry Based Proteomic Workflows. **2020**, No. July 2017, 292–306.

(2) Lam, H.; Aebersold, R. Using Spectral Libraries for Peptide Identification from Tandem Mass Spectrometry (MS/MS) Data. *Curr. Protoc. Protein Sci.* **2010**, *2010*.

(3) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *J. Proteome Res.* **2010**, *9* (1), 605–610.

(4) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994**.

(5) Bittremieux, W.; Meysman, P.; Sta, W.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **2018**, *17*, 3463–3474.

(6) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification Research Articles. **2006**, 1843–1849.

(7) Shiferaw, G. A.; Vandermarliere, E.; Hulstaert, N.; Gabriels, R.; Martens, L.; Volders, P. COSS: A Fast and User-Friendly Tool for Spectral Library Searching. *J. Proteome Res.* **2020**, *19*.

(8) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the Improved Sensitivity of Spectral Library Searching over Sequence Database Searching in Proteomics Data Analysis. *Proteomics* **2011**, *11* (6), 1075–1085.

(9) Ahrné, E.; Masselot, A.; Binz, P. A.; Müller, M.; Lisacek, F. A Simple Workflow to Increase MS2 Identification Rate by Subsequent Spectral Library Search. *Proteomics* **2009**, *9* (6), 1731–1736.

(10) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide

Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667.

(11) Benjamini, Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. **2014**, No. November 1995.

(12) Sticker, A.; Martens, L.; Clement, L. Mass Spectrometrists Should Search for All Peptides, but Assess Only the Ones They Care About. *Nat. Methods* **2017**, *14*, 643.

(13) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(14) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8*, 3176–3181.

(15) Manuscript, A. Fast and Accurate Database Searches with MS-GF+Percolator. *J. Proteome Res.* **2015**, *13* (2), 890–897.

(16) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A Draft Map of the Human Proteome A. *Nature* **2014**, *509* (7502), 575–581.

(17) Rosenberger, G.; Navarro, P.; Gillet, L.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Aebersold, R.; Diseases, M. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223–226.

(18) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; et al. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, 1–16.

(19) Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Frewen, B.; Baker, T. A.; Brusniak, M.; Paulse, C.; Lefebvre, B.; Kuhlmann, F.; et al. HHS Public Access. *Nat. Biotechnol.* **2013**, *30* (10), 918–920.

(20) Zhang, Z.; Burke, M.; Mirokhin, Y. A.; Tchekhovskoi, D. V; Markey, S. P.; Yu, W.; Chaerkady, R.; Hess, S.; Stein, S. E. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches.

2018.

(21) Vaudel, M.; Burkhart, J. M.; Breiter, D.; Zahedi, R. P.; Sickmann, A.; Martens, L. A Complex Standard for Protein Identification, Designed by Evolution. *J. Proteome Res.* **2012**, *11* (10), 5065–5071.

(22) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(23) Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17* (7), 2552–2555.

(24) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets: To the Editor. *Nat. Biotechnol.* **2015**, *33* (1), 22–24.

(25) The Uniprot Consortium. UniProt : A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (November 2018), 506–515.

(26) Silva, A. S. C.; Bouwmeester, R.; Martens, L.; Degroeve, S. Data and Text Mining Accurate Peptide Fragmentation Predictions Allow Data Driven Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions. **2019**, *35* (May), 5243–5248.