Journal of proteome research

Article

pubs.acs.org/jpr

Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-Centric Human Proteome Project

Jin-Young Cho,[†] Hyoung-Joo Lee,[†] Seul-Ki Jeong,[†] Kwang-Youl Kim,[†] Kyung-Hoon Kwon,[‡] Jong Shin Yoo,[‡] Gilbert S. Omenn,[§] Mark S. Baker,[∥] William S. Hancock,[⊥] and Young-Ki Paik*[,†]

[†]Yonsei Proteome Research Center, Department of Integrated OMICS for Biomedical Science and Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, 50 Yonsei-Ro, Seodaemoon-gu, Seoul 120-749, Korea

[‡]Korea Basic Science Institute, Ochang, Korea

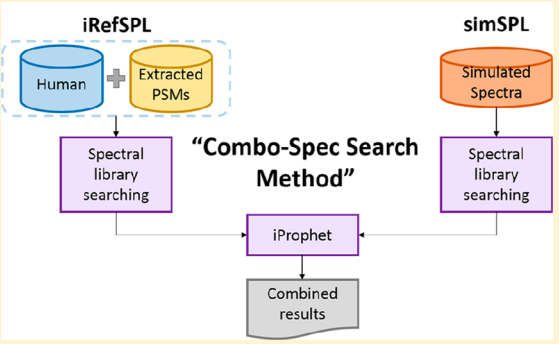
[§]Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor 48109, Michigan United States

[∥]Department of Biomedical Science, Faculty of Medicine and Health Science, Macquarie University, New South Wales 2109, Australia

[⊥]Northeastern University, Boston, Massachusetts 02115, United States

Supporting Information

ABSTRACT: Approximately 2.9 billion long base-pair human reference genome sequences are known to encode some 20 000 representative proteins. However, 3000 proteins, that is, ∼15% of all proteins, have no or very weak proteomic evidence and are still missing. Missing proteins may be present in rare samples in very low abundance or be only temporarily expressed, causing problems in their detection and protein profiling. In particular, some technical limitations cause missing proteins to remain unassigned. For example, current mass spectrometry techniques have high limits and error rates for the detection of complex biological samples. An insufficient proteome coverage in a reference sequence database and spectral library also raises major issues. Thus, the development of a better strategy that results in greater sensitivity and accuracy in the search for missing proteins is necessary. To this end, we used a new strategy, which combines a reference spectral library search and a simulated spectral library search, to identify missing proteins. We built the human iRefSPL, which contains the original human reference spectral library and additional peptide sequence-spectrum match entries from other species. We also constructed the human simSPL, which contains the simulated spectra of 173 907 human tryptic peptides determined by MassAnalyzer (version 2.3.1). To prove the enhanced analytical performance of the combination of the human iRefSPL and simSPL methods for the identification of missing proteins, we attempted to reanalyze the placental tissue data set (PXD000754). The data from each experiment were analyzed using PeptideProphet, and the results were combined using iProphet. For the quality control, we applied the class-specific false-discovery rate filtering method. All of the results were filtered at a false-discovery rate of <1% at the peptide and protein levels. The quality-controlled results were then cross-checked with the neXtProt DB (2014-09-19 release). The two spectral libraries, iRefSPL and simSPL, were designed to ensure no overlap of the proteome coverage. They were shown to be complementary to spectral library searching and significantly increased the number of matches. From this trial, 12 new missing proteins were identified that passed the following criterion: at least 2 peptides of 7 or more amino acids in length or one of 9 or more amino acids in length with one or more unique sequences. Thus, the iRefSPL and simSPL combination can be used to help identify peptides that have not been detected by conventional sequence database searches with improved sensitivity and a low error rate.



KEYWORDS: *Chromosome-Centric Human Proteome Project, proteomics, missing protein, spectral library search*

INTRODUCTION

Approximately 2.9 billion long base-pair human reference genome sequences are now known to encode some 20 000 representative proteins.[1] By inference, many proteins are not only directly encoded by a genome sequence but are also diversified by additional processing such as post-transcriptional and post-translational modifications. The direct analysis of cell and tissue protein expression is therefore necessary to collect and create a list of parts.[2,3] The Chromosome-centric Human

Received: June 20, 2015
Published: September 2, 2015

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966

Proteome Project (C-HPP) consortium was founded to map and annotate all of the proteins that are encoded by genes on each of the chromosomes found in humans.[4,5] A total of 25 C-HPP working groups from 20 nations integrate proteomics data into a genomic framework and annotate human proteins using a range of unique and often rare clinical samples. All of the currently available techniques are used to improve our understanding of complex human biological systems and disease states; however, despite the efforts of the teams, about 3000 proteins still have no clear proteomic evidence (supported by mass spectrometry (MS) or antibody detection). These proteins have been colloquially termed "missing proteins."[4−6]

A bottom-up proteomic approach is commonly used to identify proteins by MS analysis coupled to high-pressure liquid chromatography.[7,8] The proteins are extracted from the samples and digested by a protease(s) (e.g., trypsin) to produce a peptide mixture, which is subsequently injected into a reverse-phase high-pressure liquid chromatograph. When the peptide passes through the column, it is separated by its physicochemical properties (i.e., hydrophobicity, charge, and molecular size). The molecular ions of each peptide are then introduced into the mass spectrometer. The ions are fragmented, frequently by collision-induced dissociation (CID), and their mass-to-charge ratio (m/z) and intensity are recorded in subsequent MS/MS spectra, which are used to identify the peptides and eventually the proteins in the sample. Sequence database searching[9,10] is the most widely used method for MS-based proteomics,[11−16] in which sequence-spectrum matching is performed by automated sequence database search tools such as SEQUEST,[12] MASCOT,[13] X! TANDEM,[17] MyriMatch,[16] and MS-GF+;[18] however, in this approach, only m/z values are used to match the sequence-spectrum, and any other spectral information, such as residue-specific effects in cleavage and variable fragment mass peak intensities, are ignored, which may result in low sensitivity and potential errors in the handling of low-quality experimental spectra, especially those contaminated by any polymer or other noise peaks.[19]

Spectral libraries have been used for the MS-based identification of small molecules since the 1980s.[20,21] Spectral library searching takes into account all of the spectral features, such as peak intensities, the natural loss of fragments, and various unknown fragments that are specific to certain peptides and thus shows greater sensitivity and leads to a better matching of results than sequence database searching.[22,23] Yates et al.[24] suggested that this approach could be used for the identification of peptides and proteins. Spectral library searching was recently reported to outperform sequence database searching,[25−27] and spectral library search algorithms and software, such as SpectraST (2007),[22] X!Hunter (2006),[23] and BiblioSpec (2006),[28] were released at around the same time and are now widely used in this approach. The National Institute of Standards and Technology (NIST) now provides reference spectral libraries for humans and eight other species, and the PeptideAtlas, developed by the Institute for Systems Biology (ISB), provides almost 61 million human peptide spectra and various spectral libraries of individual human organs (e.g., the brain, heart, kidney, liver, and plasma).[29]

The accumulation of data depends on high-quality tandem MS spectra with high-scored peptide sequence assignment following stringent quality control criteria to build a spectral library. This ensures the reliability of the spectral library but explains its low proteome coverage and slower increase in data accumulation rate than those of the sequence database.[30] Several strategies have been proposed to expand the proteome coverage of the reference spectral library by including the predicted spectra of unobserved peptides.[19,31] For example, it has been suggested that the fragmentation patterns of a peptide in MS can be predicted by its sequence and physicochemical properties.[32,33] The CID spectra of similar peptides show extremely similar intensity patterns, which implies that the MS spectra of a peptide can be predicted by the neighbor-based approach based on its sequence.[34] Information-driven semi-empirical spectra of the reference spectral library were also demonstrated to be useful for the detection of novel phosphorylated peptides.[30,35]

In this study, we describe a new strategy, which uses a combination of multiple spectral libraries (e.g., a reference spectral library and a simulated spectral library) for spectrum−spectrum matching to identify the proteins of interest in cells or tissues. We demonstrate that, compared with conventional sequence database searching, this method can provide improved sensitivity and a lower error rate in the identification of missing proteins by extended proteome coverage.

MATERIALS AND METHODS

Data Sets

The data sets used in this study were obtained from the ProteomeXchange database (PXD). First, we obtained data set files that were generated by 47 purified human recombinant protein mixtures (Sigma UPS, Sigma-Aldrich) spiked into the biological sample (published by Ahrné et al., PXD000331)[36] and exploited this Sigma UPS data set to evaluate the performance and effectiveness of our approach. Second, we used the MS data set obtained from human placental tissue that was previously analyzed by Lee et al. (PXD000754).[37] This data set was generated using various protein enrichment techniques and MS for the comprehensive proteomic analysis of human placental tissue and was used to reanalyze and evaluate our new method for the search for novel peptides that are possibly derived from missing proteins. The more detailed metadata of the data sets are given in Supplementary Table S-3.

Integration of the Human Reference Spectral Library

The reference spectral libraries were obtained from PeptideAtlas (ISB) and the NIST public library repository. We selected the libraries that contained only the CID-fragmented ion spectra, as listed in Supplementary Table S-1. All of the human reference spectral libraries obtained were combined as a consensus spectral library (human refSPL). Proteome coverage of the original human refSPL was expanded by extracting peptide-spectrum match (PSM) entries from spectral libraries of other species. Because each PSM entry in the spectral libraries from PeptideAtlas and NIST had already been validated, we did not put a limit on the maximum sequence length. Thus, the PSM entries from the nonhuman spectral library were selected using the human tryptic peptide list, which contains peptides with a minimum of seven amino acids and a maximum of two missed cleavage sites, generated from the SwissProt human protein sequence database (2015-04). All impure spectra were removed or marked by SpectraST software (version 5.0, Build 201408281759-6544:6594 M by Henry Lam). All of the selected PSM entries were added to the human refSPL to build a human iRefSPL.

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966

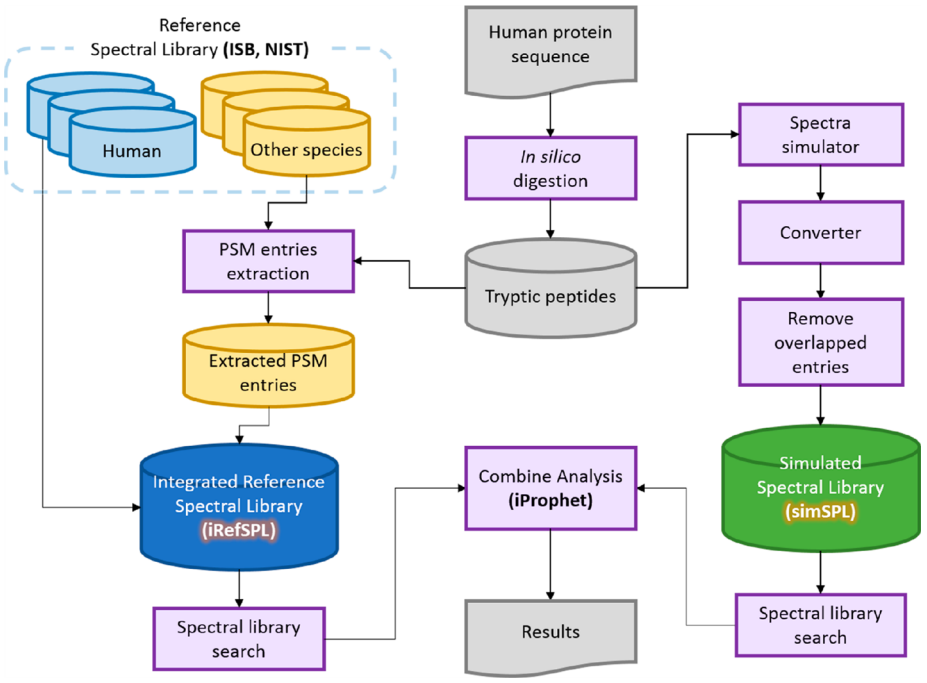


Figure 1. Workflow for building the integrated spectral library and multiple search results approach. Using the human tryptic peptide list, additional PSM entries were obtained from the other spectral libraries to expand the proteome coverage of the human reference spectral library called iRefSPL. We also constructed simSPL to identify novel peptides that were not covered by the iRefSPL search. In practice, the two spectral libraries were used independently in spectrum-spectrum matching and all of the results were combined later using iProphet.

Generation of Simulated Spectral Library

We obtained 41 061 protein sequences from neXtProt (2014-09-19) and compiled a tryptic peptide list of the proteins with a length of 7 to 35 amino acids and a maximum of 2 missed cleavage sites, as previously mentioned. In total, 2 227 896 sequences were selected for the simulation of their MS/MS spectra. MassAnalyzer (version 2.3.1) was applied to simulate the MS/MS spectra of the selected peptides using the simulation parameters: Orbitrap instrument profile; CID fragmentation mode; isolation width, 2.5; resolution, 800; collision energy (V), 35; and activation time, 30 ms. We considered two charge states, + 2 and +3 precursors, and added two types of modification into the simulated spectra: carbamidomethylation at cysteine residues for fixed modifications and oxidation at methionine residues for variable modifications. The predicted spectra were converted to the *.splib format by SpectraST,[25] and all PSM entries already included in iRefSPL were removed. The simulated spectral library was called the "human simSPL".

Protein Identification and Data Analysis

All MS data files were converted into "mgf" and "mzXML" formats by msconvert (Build date: June 17, 2013). Three protein sequence database search engines were used for sequence database searching: Mascot Server (version 2.2.07, Matrixscience), X!Tandem (2013.06.15.1 – LabKey, Insilicos, ISB), and Comet (version 2014.02 rev. 2, University of Washington). The sequence database search parameters were: trypsin for protein digestion, carbamidomethylation at cysteine residues (+57 Da) for fixed modifications, oxidation at methionine (+16 Da) for variable modifications, a maximum of two missed cleavages, 5 ppm MS tolerance, and 0.6 Da MS/MS tolerance. Two charge states, 2+ and 3+, were considered. To filter the false discovery rate (FDR), reversed protein sequences were included in the target sequence database using the TOPPAS DecoyDatabase builder (version 1.11.1).[38] SpectraST was used to build and search the spectral library. All results that had a lower F value than 0.45 were excluded. To estimate the FDR, we generated an equal-size artificial decoy library and appended it to the target spectral library following the method described by Lam et al.[39] The result of each experiment was analyzed using PeptideProphet,[40] and all of the results were combined using iProphet[41] (built in TransProteome Pipeline version 4.8.0 PHILAE, Build 201411201551-6764) with default parameters. We used decoy hits and a nonparametric model to ascertain the negative frequency and determined two peptide probability thresholds by class-specific FDR filtering.[42] Each threshold was established in separate FDR estimations in two classes (peptide hits from iRefSPL as class I and from simSPL as class II). The FDR of each class was limited to <1%.

RESULTS AND DISCUSSION

Construction of the Integrated Reference Spectral Library (iRefSPL) Which Contains Peptide Spectrum Matches from Humans and Eight Nonhuman Species

We designed a method that uses two spectral libraries to expand proteome coverage for spectral library searching to detect additional peptides (Figure 1). To expand the proteome coverage of the human reference spectral library, we prepared an integrated reference spectral library called the iRefSPL. The library was built by combining the original human reference spectral library and PSM entries obtained from the spectral libraries of other species. The rationale for this approach was provided by a previous report that indicated a close correlation between the peptide fragmentation pattern and the sequence, the charge state, and modifications.[32,33] We expected that the proteome coverage of the spectral library of interest could be expanded by the addition of PSM entries and that this would

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959–4966

Table 1. Similarity of Common PSM Pairs in Humans and Eight Other Nonhuman Species (*Caenorhabditis elegans*, Chicken, *Drosophila melanogaster*, *Escherichia coli*, Mouse, Rat, Yeast, and Zebrafish) Provided by NIST. The Dot Scores Were Calculated by Matching The Human and Each Species PSM.

		C. elegans	chicken	*D. melanogaster*	*E. coli*	mouse	rat	yeast	zebrafish
similarity (dot score)	1–0.9	775	6	1377	19	26180	11949	257	909
	0.9–0.8	333	6	734	22	15203	6669	130	2287
	0.8–0.7	92	5	296	4	4946	2284	49	839
	0.7–0.6	19	2	58	2	812	495	11	151
	0.6–0.5	1	3	0	0	67	44	1	9
	0.5–0.4	0	2	0	1	4	0	0	2
	0.4–0.3	0	0	0	0	1	0	0	0
	0.3–0.2	0	0	0	0	0	0	0	0
	0.2–0.1	0	0	0	0	0	0	0	0
	0.1–0	0	0	0	0	0	0	0	0

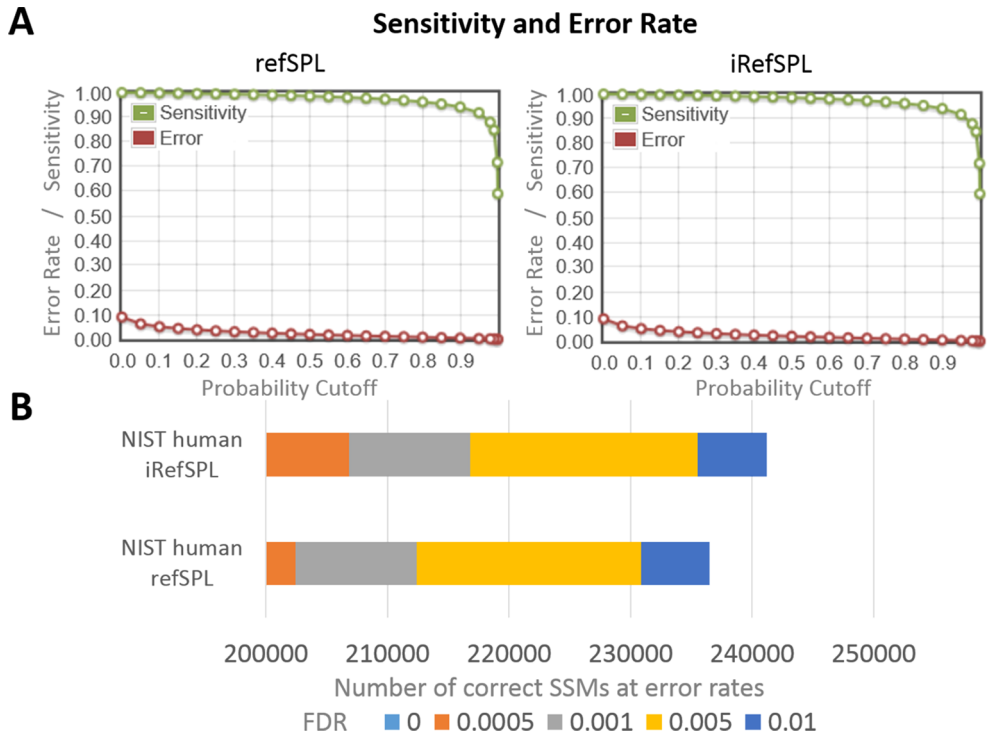


Figure 2. Comparison of the spectral library search results from iRefSPL and refSPL. (A) Comparison of each sensitivity and error rate model of iRefSPL and refSPL. (B) Comparison of the number of spectrum-spectrum matches through different error rates.

not introduce false-positive problems. To estimate the dependence of the fragmentation pattern on the physicochemical properties of the peptide (e.g., sequence, charge state, and modification) through various spectral libraries, we selected common PSM entries from the NIST human reference spectral library and spectral libraries of eight other species. In total, 77 056 PSM pairs were collected to compare their similarity through various spectral libraries. The similarity of the PSM pairs was estimated by the dot scoring method.[22] Table 1 outlines the distributions of PSM pairs, as expressed by their dot scores. Many PSM pairs tended to show a dot score of close to 1, suggesting that the fragmentation and peak intensity patterns of the peptides were highly correlated with their sequence, charge, and modification state. On the basis of the result, we extracted a total of 51 374 PSM entries from 13 nonhuman spectral libraries to expand the proteome coverage of human refSPL (Supplementary Table S-2). We added the PSM entries obtained from the 13 nonhuman species spectral libraries to the human refSPL to produce the human iRefSPL.

To test the effectiveness of adding PSM entries, we analyzed the placental tissue data set using both the human iRefSPL and human refSPL (called the Combo-Spec Search method). Figure 2A presents a prediction model that shows the estimated sensitivity and error rate of both the human iRefSPL and the human refSPL. The two results did not differ significantly. The human iRefSPL identified more peptides, with an especially low error rate (≤0.0005), than the human refSPL (Figure 2B), suggesting that PSM entries extracted from other spectral libraries can be used to successfully expand the proteome coverage of the human refSPL without introducing false-positive problems.

Comparison of the Sensitivity and Error Rate of Various Search Methods

We examined the performance of the Combo-Spec Search method in comparison with other conventional approaches in identifying additional peptides at a low error rate using the Sigma UPS data set. Three protein sequence database search

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959–4966

engines (Mascot, X!Tandem, and Comet) and the original reference UPS spectral library were used as conventional approaches. The FASTA sequence database and the reference spectral library of the Sigma UPS standard protein mix (UPS refSPL) were obtained from the NIST (released 2011-05-24). We did not prepare the iRefSPL for analysis of the Sigma UPS data set in this test because the original refSPL from NIST for Sigma UPS data set analysis already has sufficient proteome coverage (~85% of the sequences of all of the 47 standard proteins). Thus, we used the refSPL of Sigma UPS data set rather than build an additional iRefSPL.

We compared the number of correct matches through different error rates obtained by refSPL only and each of the three sequence database search engines. As expected, the number of matches detected by refSPL only (second bar in Figure 3A) was greater than that obtained by each single sequence search engine (bottom three bars in Figure 3A).

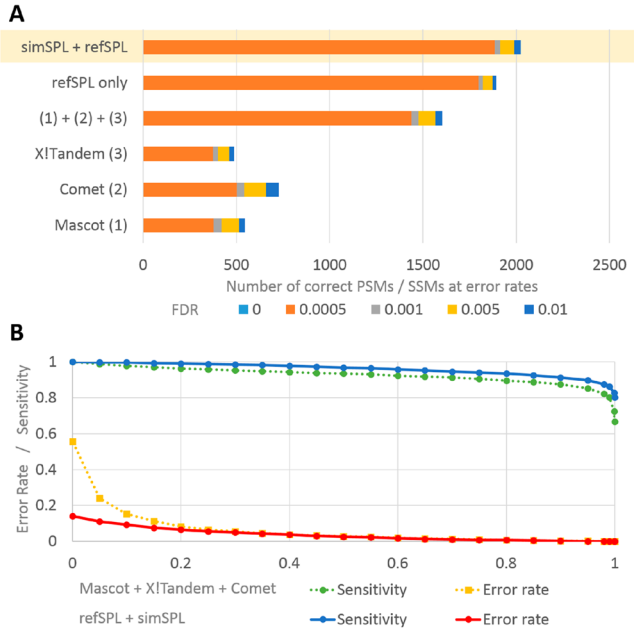


Figure 3. Comparison of spectral library searches using refSPL and simSPL and conventional methods for the analysis of the Sigma UPS data set. (A) Comparison of matches between a combination of the refSPL and simSPL, refSPL only, and three sequence search engines. (B) Comparison of the sensitivity and error rates of the refSPL-simSPL combination and multiple sequence database searching.

The first bar in Figure 3A shows the effectiveness of the simSPL. The refSPL had an 85% proteome coverage of the UPS data, so we built the simSPL using 15% of the gaps to complete the coverage with no overlap with refSPL because the simSPL shows a better positive/negative number of sibling peptide distributions in the refSPL-simSPL combination than the complete proteome coverage version of simSPL. (See Supplementary Figure S-1.)

With an FDR of <1%, we detected 427 different peptides using the refSPL only; however, using the combination of simSPL and refSPL, we detected 33 more different novel peptides, showing that the combination of both refSPL and simSPL (the refSPL-simSPL combination method) can detect more peptides with a low error rate than other conventional methods (refSPL only, single or multiple sequence database searching). The use of a combination of multiple search engines is known to produce highly improved identification rates,[43] and the combination of three sequence database search engines (Multiple DB Search) showed a significantly increased number of matches with a low error rate (≤0.0005). To evaluate the sensitivities of both multiple search strategies (the Multiple DB Search and the Combo-Spec Search method), we determined the relationship between sensitivity and error rate. Figure 3B shows that the Combo-Spec Search method had slightly greater sensitivity than the Multiple DB Search, but the difference was not significant. Both methods showed good sensitivity for various probability thresholds; however, the Combo-Spec Search method showed lower error rates than the Multiple DB Search with extremely low probability thresholds (≤0.2), indicating that the Combo-Spec Search method has greater effective restriction power for errors than the Multiple DB Search.

Application of the Combo-Spec Search Method to Identify Missing Proteins

To test the performance of the human Combo-Spec Search method in identifying missing proteins, we attempted to reanalyze the human placental tissue data set (PXD000754)[37] independently using the Combo-Spec Search method and the SpectraST and combining the results using iProphet (Figure 4).

The combined results were filtered at an FDR of <1% at the protein level. All combined matched results were classified into two groups (matched by human iRefSPL and human simSPL), to which a probabilistic threshold was separately applied (0.8299 for iRefSPL group and 0.9303 for simSPL group) to provide an FDR of <1% at the peptide level in each group. In total, 4104 proteins were identified, slightly fewer (135) than the previous result of 4239 proteins,[37] which may have been due to the use of CID spectra only in this study, whereas various types of spectrum (CID, higher-energy collisional dissociation, and electron-transfer dissociation) were used in the previous study. The human iRefSPL and simSPL used in this study can only support the CID type spectra for spectral library searching. Using multiple sequence database search engines (Mascot, X!Tandem, and Comet), 3607 proteins were identified at an FDR of <1% at the protein level. When the two results generated by the Multiple DB Search Method and the Combo-Spec Search method were compared, the Combo-Spec Search method showed the higher rate of protein identification than the former. When the previous search results (4239 proteins) were applied to the old version of neXtProt DB (2012-10-07 release), 42 proteins were found to be newly identified missing proteins;[37] however, when neXtProt DB (2014-09-19 release) was applied to the Combo-Spec Search method, 12 proteins were newly found missing proteins. (See each spectrum and matched peak information in Supplementary Figure S-3.) The 12 missing proteins passed our consensus criterion of at least 2 peptides of 7 or more amino acids in length or 1 of 9 or more amino acids in length and 1 or more unique sequence. Using the Multiple DB Search Method, no newly identified missing proteins were found.

Three of the proteins were identified by simSPL, and the unique peptides of three proteins were not included in any reference spectral library, implying that simSPL is complementary to iRefSPL in terms of novel peptide searches. Thus, the use of both iRefSPL and simSPL showed a synergistic effect for identifying known and novel peptides from large data sets with high sensitivity and a low error rate and identified peptides

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966

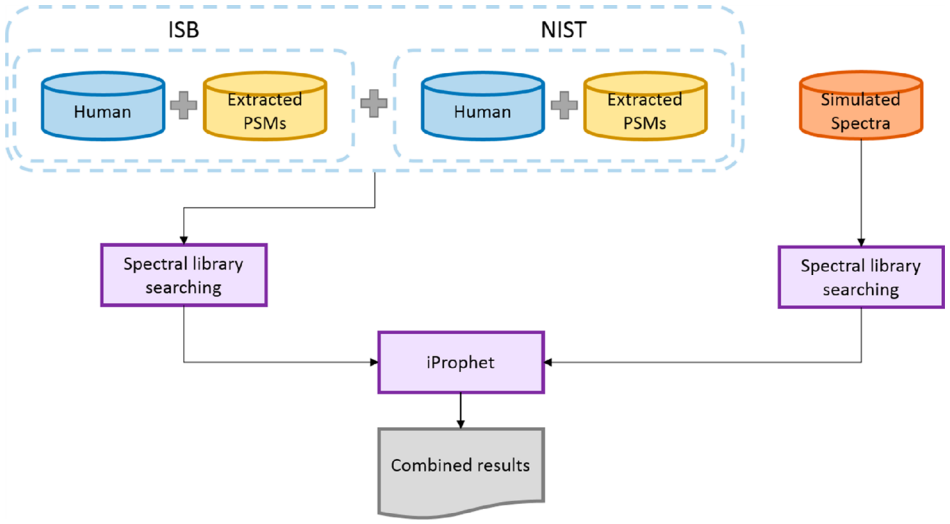


Figure 4. Workflow of the human placental tissue data set (PXD000754) analysis obtained by searching three spectral libraries and integrating the results using iProphet.

Table 2. List of Identified Missing Proteins in This Study

Chr	protein accession (gene name) / peptide sequence/charge	coverage (%) / length	total PSMs / PSMs	protein prob. / peptide prob.	PE / dot	*F* value	matched library
1	Q5VVM6 (CCDC30)	2.9	3	0.8953	2		
	DHFLIAC$_{160}$DLLQRENSELETKVLK/2	23	3	0.8953	0.758	0.622	iRefSPL
3	Q8NGV6 (OR5H6)	6.8	2	0.987	2		
	AVSTCGAHLLSVSLYYGPLTFK/3	22	2	0.987	0.898	0.783	iRefSPL
6	Q8IZF3 (GPR115)	2	88	0.9773	2		
	QVNGLVLSVVLPER/3	14	88	0.9919	0.891	0.722	iRefSPL
7	Q8WXK1 (ASB15)	2	5	0.9955	2		
	KGSYDMVSTLIK/3	12	5	0.9955	0.939	0.571	iRefSPL
9	Q8NE28 (STKLD1)	3.7	3	0.8783	2		
	QM$_{147}$VPASITDM$_{147}$LLEGNVASILEVMQK/3	25	3	0.8783	0.713	0.607	iRefSPL
11	Q6IEU7 (OR5M10)	3.5	11	0.9987	2		
	DVILAIQQM$_{147}$I/2	10	11	0.9987	0.757	0.613	simSPL
13	O75343 (GUCY1B2)	2.1	2	0.9949	2		
	DQEALQAAFLKMK/3	13	2	0.9949	0.908	0.698	iRefSPL
18	Q9H2F9 (CCDC68)	5.1	5	0.9721	2		
	DLQLLEM$_{147}$NKENEVLKIK/3	17	5	0.9721	0.749	0.608	iRefSPL
19	C9J6K1 (C19orf81)	7.1	8	0.9683	4		
	RM$_{147}$LEALGAEPNEEA/3	14	8	0.9683	0.852	0.545	iRefSPL
19	Q96RP8 (KCNA7)	3.1	3	0.9957	2		
	GLQILGQTLRASM$_{147}$R/3	14	3	0.9957	0.816	0.623	simSPL
20	Q8N687 (DEFB125)	10.3	4	0.9243	2		
	NKLSCCISIISHEYTR/2	16	4	0.9243	0.837	0.697	iRefSPL
21	P57055 (RIPPLY3)	2.9	18	0.979	2		
	MEPEAAAGAR/2	10	18	0.979	0.653	0.552	simSPL

that had not been detected by some conventional sequence database search engines in the previous study. Using the Combo-Spec Search method, we were able to detect 12 missing proteins from the previously published data set, suggesting that the method could be useful for reanalyzing other previously published data sets and detecting additional missing proteins.

CONCLUSIONS

Although the rigorous protein search analyses were carried out on MS data produced under optimal performance conditions, it is inevitable that some proteins will have remained undetected; therefore, a better search strategy that provides greater sensitivity and more accurate analysis in the search for missing proteins needs to be developed. This study demonstrated that the application of the Combo-Spec Search method to a previously analyzed data set[37] can provide additional opportunities to identify missing proteins that have never been detected by sequence database searches. Original reference spectral libraries usually have insufficient proteome coverage (30−40%) compared with the sequence database. We suggest that the combination of multiple spectral libraries with different proteome coverage could be one solution to avoid this limitation. The improved performance of the Combo-Spec Search method in the identification of missing proteins is due to its expanded proteome coverage. These promising results

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966

indicate that it would also be worth reanalyzing previously reported data sets deposited in the ProteomeXchange repository in the hope of detecting additional missing proteins. Using this method, we were able to detect 12 new missing proteins, two of which were olfactory receptors, which is an exceptional result when considering the sample type used in this study. We made a thorough search again through the currently updated PeptideAtlas, but we were unable to find any evidence that the two olfactory receptors were false-positive matches; however, we cannot exclude the possibility of a single nucleotide polymorphism (SNP) or any modifications because our newly built spectral libraries (iRefSPL and simSPL) do not contain such rare modification types or SNP. This issue can be re-examined together with the 12 newly identified missing proteins when the upgraded versions of iRefSPL and simSPL, into which artificial modifications and SNP will be introduced, become available in the future. Some useful public spectral library and mass spectral data repositories (PeptideAtlas, NIST Peptide Library, and GPMdb) are available, which are updated at regular intervals (e.g., quarterly or yearly). Using the latest data, we can obtain a more expanded and sophisticated spectral library for use in the Combo-Spec Search method. Finally, we propose that the Combo-Spec Search method could serve as a common practice in the search for missing proteins and could thus replace the conventional sequence database search approach.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00578.

Supplementary Table S-1: List of the reference spectral libraries used in this study. Supplementary Table S-2: Number of extracted PSM entries from nonhuman species spectral libraries using the human whole tryptic peptide list. Supplementary Table S-3: Metadata of data sets. Supplementary Figure S-1: Comparison of the effects of two simSPLs with different proteome coverage in the Combo-Spec Search method. Supplementary Figure S-2: Statistics from the human placental tissue data set. Supplementary Figure S-3: The spectra and matched peaks of the newly identified missing proteins. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +82-2-2123-4242. Fax: +82-2-393-6589. E-mail: paikyk@yonsei.ac.kr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Eric Deutsch for his comments and guidance in designing the concept of the use of spectral libraries as described in this manuscript. This work was supported by a grant from the Korean Ministry of Health and Welfare (to Y.K.P., HI13C2098-International Consortium Project) and the National Research Foundation of Korea (2011-0028112 to Y.K.P.).

■ ADDITIONAL NOTE

Intended as part of the The Chromosome-Centric Human Proteome Project 2015 special issue.

■ REFERENCES

(1) Maher, B. ENCODE: The human encyclopaedia. *Nature* **2012**, *489*, 46−8.

(2) Dhingra, V.; Gupta, M.; Andacht, T.; Fu, Z. F. New frontiers in proteomics research: a perspective. *Int. J. Pharm.* **2005**, *299*, 1−18.

(3) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **1999**, *19*, 1720−30.

(4) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30*, 221−3.

(5) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; et al. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11*, 2005−13.

(6) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; et al. Metrics for the Human Proteome Project 2013−2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13*, 15−20.

(7) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198−207.

(8) Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down? *Science* **2006**, *314*, 65−6.

(9) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 699−711.

(10) Zhang, Z.; Wu, S.; Stenoien, D. L.; Pasa-Tolic, L. High-throughput proteomics. *Annu. Rev. Anal. Chem.* **2014**, *7*, 427−54.

(11) Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1*, 195−202.

(12) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.

(13) 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*, 3551−67.

(14) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−7.

(15) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; et al. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958−64.

(16) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6*, 654−61.

(17) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−74.

(18) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(19) Yen, C. Y.; Houel, S.; Ahn, N. G.; Old, W. M. Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol. Cell. Proteomics* **2011**, *10*, M111.007666.

(20) Lam, H.; Aebersold, R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. *Methods* **2011**, *54*, 424−31.

(21) 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*, 859−66.

(22) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655−67.

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966

(23) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **2006**, *5*, 1843−9.
(24) Yates, J. R., 3rd; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* **1998**, *70*, 3557−65.
(25) Hu, Y.; Lam, H. Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications. *J. Proteome Res.* **2013**, *12*, 5971−7.
(26) 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*, 1075−85.
(27) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **2008**, *5*, 873−5.
(28) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **2006**, *78*, 5678−84.
(29) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005**, *6*, R9.
(30) Hu, Y.; Li, Y.; Lam, H. A semi-empirical approach for predicting unobserved peptide MS/MS spectra from spectral libraries. *Proteomics* **2011**, *11*, 4702−11.
(31) Yen, C. Y.; Meyer-Arendt, K.; Eichelberger, B.; Sun, S.; Houel, S.; et al. A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. *Mol. Cell. Proteomics* **2009**, *8*, 857−69.
(32) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2005**, *77*, 6364−73.
(33) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 3908−22.
(34) Ji, C.; Arnold, R. J.; Sokoloski, K. J.; Hardy, R. W.; Tang, H.; et al. Extending the coverage of spectral libraries: a neighbor-based approach to predicting intensities of peptide fragmentation spectra. *Proteomics* **2013**, *13*, 756−65.
(35) Suni, V.; Imanishi, S. Y.; Maiolica, A.; Aebersold, R.; Corthals, G. L. Confident site localization using a simulated phosphopeptide spectral library. *J. Proteome Res.* **2015**, *14*, 2348.
(36) Ahrne, E.; Molzahn, L.; Glatter, T.; Schmidt, A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* **2013**, *13*, 2567−78.
(37) Lee, H. J.; Jeong, S. K.; Na, K.; Lee, M. J.; Lee, S. H.; et al. Comprehensive genome-wide proteomic analysis of human placental tissue for the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2013**, *12*, 2458−66.
(38) Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; et al. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *J. Proteome Res.* **2012**, *11*, 3914−20.
(39) 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*, 605−10.
(40) 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*, 5383−92.
(41) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, M111 .07690.
(42) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114−25.
(43) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics* **2013**, *12*, 2383−93.

DOI: 10.1021/acs.jproteome.5b00578
J. Proteome Res. 2015, 14, 4959−4966