Journal of proteome research

Article

pubs.acs.org/jpr

Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate

Gun Wook Park,[†,‡,#] Heeyoun Hwang,[†,#] Kwang Hoe Kim,[†,‡] Ju Yeon Lee,[†] Hyun Kyoung Lee,[†,‡] Ji Yeong Park,[†,‡] Eun Sun Ji,[†] Sung-Kyu Robin Park,[§] John R. Yates, III,[§] Kyung-Hoon Kwon,[†] Young Mok Park,[∥] Hyoung-Joo Lee,[⊥] Young-Ki Paik,[⊥] Jin Young Kim,[*,†] and Jong Shin Yoo[*,†,‡]

[†]Biomedical Omics Group, Korea Basic Science Institute, 162 YeonGuDanji-Ro, Ochang 363-883, Republic of Korea

[‡]Graduate School of Analytical Science and Technology, Chungnam National University, Daejeon 34134, Republic of Korea

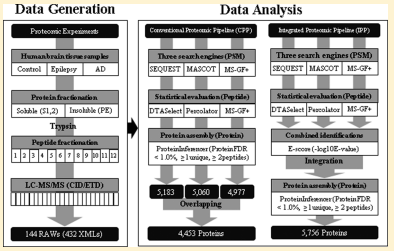
[§]Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California 92037, United States

[∥]Center for Cognition and Sociality, Institute for Basic Science, Daejeon 305-811, Republic of Korea

[⊥]Yonsei Proteome Research Center and Department of Integrated OMICS for Biomedical Science, and Department of Biochemistry, College of Life Science and Biotechnology, Yonsei University, Seoul 120-749, Republic of Korea

Supporting Information

ABSTRACT: In the Chromosome-Centric Human Proteome Project (C-HPP), false-positive identification by peptide spectrum matches (PSMs) after database searches is a major issue for proteogenomic studies using liquid-chromatography and mass-spectrometry-based large proteomic profiling. Here we developed a simple strategy for protein identification, with a controlled false discovery rate (FDR) at the protein level, using an integrated proteomic pipeline (IPP) that consists of four engrailed steps as follows. First, using three different search engines, SEQUEST, MASCOT, and MS-GF+, individual proteomic searches were performed against the neXtProt database. Second, the search results from the PSMs were combined using statistical evaluation tools including DTASelect and Percolator. Third, the peptide search scores were converted into E-scores normalized using an in-house program. Last, ProteinInferencer was used to filter the proteins containing two or more peptides with a controlled FDR of 1.0% at the protein level. Finally, we compared the performance of the IPP to a conventional proteomic pipeline (CPP) for protein identification using a controlled FDR of <1% at the protein level. Using the IPP, a total of 5756 proteins (vs 4453 using the CPP) including 477 alternative splicing variants (vs 182 using the CPP) were identified from human hippocampal tissue. In addition, a total of 10 missing proteins (vs 7 using the CPP) were identified with two or more unique peptides, and their tryptic peptides were validated using MS/MS spectral pattern from a repository database or their corresponding synthetic peptides. This study shows that the IPP effectively improved the identification of proteins, including alternative splicing variants and missing proteins, in human hippocampal tissues for the C-HPP. All RAW files used in this study were deposited in ProteomeXchange (PXD000395).



KEYWORDS: *false discovery rate, proteogenomics, integrated proteomic pipeline, E-value, E-score, ProteinInferencer, missing protein, alternative splicing variant*

INTRODUCTION

Proteogenomic analysis is a technique commonly used to identify protein-coding genes and transcripts from various organisms by mapping mass spectrometry (MS) data obtained from biologically derived proteins directly to genomic or transcript sequences.[1−4] This approach has been used to identify novel protein-coding genes, new alternative splicing and sequence variants, new translational initiation sites, short open reading frames, as well as missing proteins, and it has also been used to classify pseudogenes as protein-coding or noncoding genes.[1−4] A major challenge in proteogenomics has been the lack of sufficient proteomic analysis data.[5] Therefore, generating more proteomic data has been an ongoing, large-scale initiative in this field. As an example, the community-driven Chromosome-Centric Human Proteome Project (C-HPP) is a worldwide project initiated by proteomic researchers to create expression profiles for all human chromosomes.[6,7] Recently, proteogenomic approaches have attracted particular interest, with large-scale human proteomic analyses identifying large numbers of peptides and proteins.[5,8−10]

A number of different groups have reported the identification of human proteins from analyses of their own proteomes. For example, Kim et al.[8] conducted a large-scale analysis of MS data obtained from the human proteome. This study employed a nonredundant database containing three- and six-frame trans-

Special Issue: Chromosome-Centric Human Proteome Project 2016

Received: May 1, 2016
Published: August 18, 2016

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

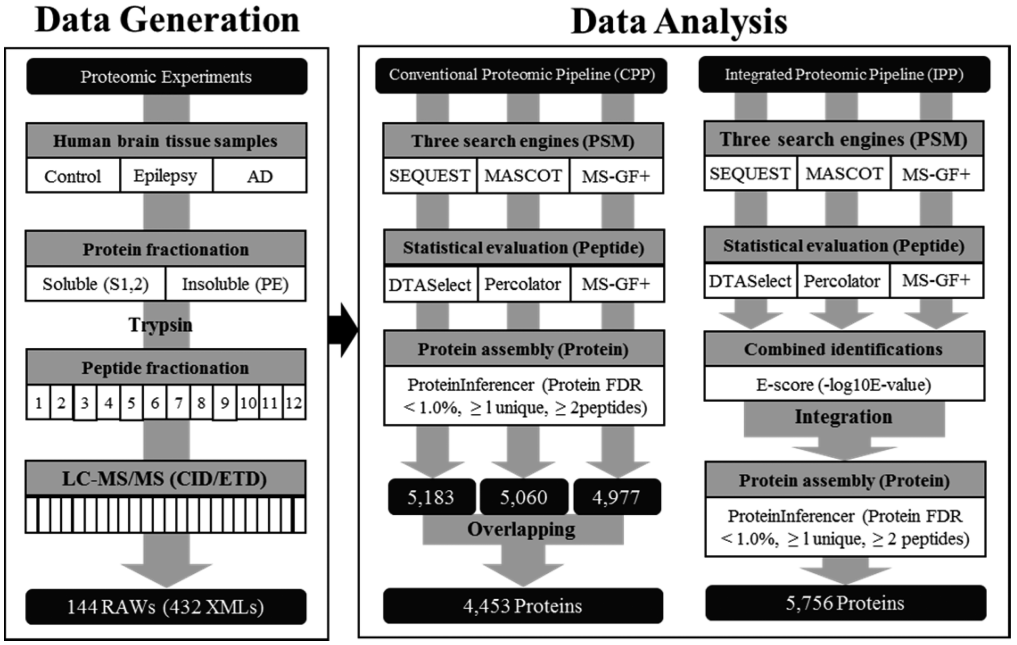


Figure 1. Proteomic data analysis comparison between a conventional proteomic pipeline (CPP) and an integrated proteomic pipeline (IPP) for protein identification. A total of 144 RAW files were processed by RawExtractor and MM File Conversion. The neXtProt database was used for a spectral library search by three different search engines. Identified peptides were assembled using ProteinInferencer, and protein false discovery rate (FDR) control was conducted. Using the IPP, 5756 proteins containing two or more peptides were identified (at least one unique peptide corresponding to one protein in the neXtProt database) with a controlled FDR of 1.0% at the protein level in human hippocampal tissues.

lations of Ensembl transcripts and gene models from the Encyclopedia of DNA Elements (ENCODE) database; 17 294 genes, 84% of the total annotated protein-coding genes in human, were mapped with no protein-level false discovery rate (FDR) and inclusion of single-peptide identifications as short as six amino acids. In other human proteogenomic work by Wilhelm et al. using a global target-decoy approach,[9] 18 097 (92%) of the 19 629 human genes were annotated in SwissProt with no protein-level FDR. In another example, Zhang et al.[10] used a 2% FDR for PSMs and a minimum of two unique peptide sequences to identify a given protein within the full data set (95 samples, 1425 raw data); they also reported that this approach led to an unacceptably high FDR of 32% at the protein level. Therefore, they ultimately used a 0.1% FDR for PSMs to achieve a protein-level FDR of 2.64%.

Although combining data from multiple experiments may increase protein coverage, the advantage of using multiple data set comes at the expense of higher false-positive protein identification rates.[11] Therefore, in a large-scale proteomic analysis, reliable protein identification is especially important, as demonstrated by a deep dive examination of the spectra for a large class of exceptional identifications (hundreds of olfactory receptors, none of which was confirmed).[12]

To improve the number of proteins identified in shotgun proteomics, Balgley et al.[13] and Jones et al.[14] demonstrated that different search engines do not result in the same peptide identifications for large-scale data sets, particularly for PSMs that score close to the threshold for acceptance or rejection. This suggests that it should be possible to select more proteins from the spectra set by employing multiple search engines if there is a framework suitable for combining the results. In other words, the results from several search engines can be combined to improve the rate of true positives.[14] Ma et al.[15] and Jones et al.[14] demonstrated that combining results from several search engines can increase the number of protein and peptide identifications with a high confidence level. Recently, the "picked" target-decoy strategy has been developed, which combines scores generated by multiple search engines for protein-level FDR estimation in large proteomic data sets;[16] it resulted in the removal >3000 claimed proteins from refs 8 and 9.

We developed an integrated proteomic pipeline (IPP) to improve the correct identification of proteins with a controlled FDR of 1.0% at the protein level for large-scale proteomic studies such as the C-HPP. We used three database search engines, SEQUEST,[17] MASCOT,[18] and MS-GF+,[19] as well as the validation tools DTASelect[20] and Percolator[21] (and Mascot Percolator[22]). Then, we assembled the validated peptides using ProteinInferencer,[23] resulting in combined peptide data with normalized peptide spectrum match (PSM) E-scores from three different search engines. We reported a chromosome 11-centric human proteome analysis from human brain hippocampus tissue, as a model study, with the gene clusters extracted from a specific biological process or molecular function in gene ontology.[24] We also showed that various protein variants, such as translated lncRNA variants, novel alternative splicing variants (ASVs), and single amino acid variants, were identified in a chromosome-based study using customized database.[25]

We represents a reanalysis of data set (PXD000395) published[25] in the previous paper to compare the performance between conventional and integrated proteomic pipelines (CPP and IPP, respectively) for protein identification with a controlled FDR of <1% at the protein level. Additionally, we used our pipeline to find more number of ASVs and missing proteins in human hippocampal tissues.

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

EXPERIMENTAL SECTION

Sample Preparation and LC–MS/MS Analysis

Using three sets of human hippocampal tissues (from controls and patients with epilepsy and Alzheimer disease), soluble fractions I and II and the insoluble fraction were used in proteomic analyses for the workflows using multiple search engines (Figure 1). This is the same method as used in our previous study.[24]

Liquid Chromatography Tandem Mass Spectrometry (LC–MS/MS) Conditions for the Analysis of Synthetic Peptides

Synthetic peptides were purchased from Anygen (Gwangju, South Korea). To confirm peptide fragmentation of missing proteins, we performed the analysis using LTQ-Orbitrap mass spectrometer (Elite version; Thermo Fisher Scientific, Waltham, MA) equipped with the EASY-nLC system (Thermo Fisher Scientific) using collision-induced dissociation (CID) MS/MS fragmentation. The details of the peptide preparation and MS conditions were similar to those reported by Hwang et al.[25] with minor modifications. A total of 11 synthetic peptides were diluted with 50 mM ammonium bicarbonate until the concentration of each peptide reached 1 pmol/μL. The diluted synthetic peptides were reduced using dithiolthreitol (final concentration at 5 mM) for 45 min at 60 °C with gentle shaking. Next, iodoacetamide was added from 100 mM stock to a final concentration at 10 mM. Reduced peptides were incubated for 45 min at room temperature in the dark. The sample was lyophilized in the SpeedVac system (Thermo Scientific, Wiesbaden, Germany). For MS analysis, a pooled sample of 11 peptides was diluted with 0.1% formic acid to 100 fmol/μL. Each synthetic peptide (500 fmole) was injected at a flow rate of 4.0 μL/min into the C18 trap column [180 μm internal diameter (ID) × 20 mm, 5 μm, 100 Å] and analyzed at a flow rate of 0.3 uL/min into a C18 analytical column (100 μm ID × 200 mm, 3 μm, 100 Å). The LC gradient using buffer A (99.9% water and 0.1% formic acid) and buffer B (99.9% acetonitrile and 0.1% formic acid) was as follows: 5% buffer B for 15 min, which was ramped up to 15% over 5 min, to 50% over 75 min, and to 95% over 1 min, at which it was maintained for 13 min and then decreased to 5% for another 11 min. The full scan resolution was 120 000 at m/z 400. The six most intense ions were sequentially isolated for tandem MS CID scans were acquired using the LTQ mass spectrometer with an activation time of 10 ms, charge state of 2+ or more, normalized collision energy in CID of 35%, and an isolation window CID of 2.0 Da. Previously fragmented ions were excluded for 180 s for all MS/MS scans, and the molecular ions of 11 peptides were included for MS/MS. The MS1 mass scan range was 400–2000 m/z. The electrospray voltage was maintained at 2.4 kV, and the capillary temperature was set at 250 °C.

Data Analysis

The 144 (3 samples × 2 protein fractions × 12 peptide fractions × 2 fragmentation methods) RAW MS data files (PXD000395) comprising a total of ~3 600 000 MS/MS spectra obtained from LTQ-Orbitrap Velos were converted into .ms2 (for SEQUEST) and .mgf (for MASCOT and MS-GF+) files using the freeware program RawExtractor version 1.9 (The Scripps Research Institute, La Jolla, CA) and MM File Conversion Tools version 3.9 (http://www.massmatrix.net/mm-cgi/downloads.py). The CID and electron-transfer dissociation (ETD) MS/MS spectra were searched separately against the neXtProt[26] database using three proteome search engines (release 09, 2014; http://www.nextprot.org), which contain 20 055 human protein sequences (target sequences) with reverse sequences as a decoy database. The Decoy database was generated by reverse reading the uploaded FASTA file of neXtProt database in the IP2 system (Integrated Proteomic Application, San Diego, CA) developed as a server side platform. As shown in Figure 1, the 144 MS/MS spectra files were searched using SEQUEST, MASCOT, and MS-GF+ with the same parameter set: precursor ion tolerance, 50 ppm (ppm); fragment ion tolerance, 0.8 Da; missed cleavages, 3; and modification: carbamidomethyl cysteine (fixed) and oxidized methionine (variable), and enzyme (full tryptic).

Then, for peptide validation, the score threshold for PSMs was set at 1% FDR. Estimated FDRs were calculated using DTASelect (version 2.0, http://fields.scripps.edu/DTASelect/) for SEQUEST search results and the MASCOTPercolator (version 2.02, http://www.sanger.ac.uk/Software/analysis/MascotPercolator/) and Percolator (version 1.14, http://noble.gs.washington.edu/proj/percolator/) for MASCOT search results. For all of the search results obtained from the three search engines combined, controlled FDR was calculated using ProteinInferencer. To determine whether a known single amino acid variation (SAAV) in neXtProt could convert the missing proteins into known proteins, we cross-checked all identified peptides with the result of SAAVs to avoid misidentification of proteins, despite the correct peptide identification of SAAVs using a customized database.[25] To find missing proteins, we checked all identified spectra in the latest version of neXtProt (release 02, 2016), and all identified extraordinary proteins including ASVs with two or more unique peptides were considered.

To verify the spectra from synthetic peptides, we converted the obtained MS RAW file into a .ms2 file using RawExtractor version 1.9 and searched against 11 synthetic peptide database using SEQUEST with the following parameter set: precursor ion tolerance, 50 ppm; fragment ion tolerance, 0.8 Da; missed cleavages, 0; and modification: carbamidomethyl cysteine (fixed) and oxidized methionine (variable).

Calculation of the E-Score

The E-scores were calculated for all PSMs in the three search engines via the following steps. Histograms of the XCorr value (SEQUEST), PEP score (MASCOT), and RAW score (MS-GF+) were generated for every log-transformed PSM subject to a neXtProt database search (Figure S1). Additionally, using the decoy-matched PSM, we generated a histogram of decoy PSMs, which is similar to incorrect target PSMs. Two histograms were log-transformed and calculated independently for each result. Then, least-squares lines (red line) were fitted using logarithmic number of decoy PSMs from each search engine. E-values of target PSMs were calculated using the slope (a) and intercept (b) of the least-squares line from decoy PSMs ((E-value = a*x (XCorr value, PEP score, and RAW score) + b)). The E-value is used to normalize the score and can be calculated for each PSM based on the score distribution of each spectrum obtained from multiple search engine results.[27,28] The E-value is a widely used statistical metric to determine significance levels, and its use in proteomics was previously validated.[28,29] Finally, the E-scores were calculated as the negative log of the E-values (E-score = $-\log_{10}$ E-value) from target PSMs. We added this discussion and followed the figures in the manuscript and the

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082–4090

supplementary figure (Figure S1). Regardless of same peptide or different peptide, E-scores were independently calculated by results from search engine such as SEQUEST, MASCOT, and MS-GF+, respectively. We used E-scores as input data for ProteinInferencer to control the global protein-level FDR. To use ProteinInferencer, the input file needs to be converted into DTASelect-filter.txt file format. Therefore, the files generated from Percolator and MS-GF+, such as .tab, .tsv, and .mzid, were converted into DTASelect-filter.txt files using an in-house program coded by Python script (version 2.7). ProteinInferencer can be downloaded for free at http://fields.scripps.edu/downloads.php. For details of the system requirements and input DTASelect-filter.txt file information, see the READ-ME file at the following Web site (http://www.proteomicswiki.com/wiki/index.php/ProteinInferencer).

RESULTS AND DISCUSSION

Protein-Level FDR Using Large Data Sets

In the initial protein assembly derived from the MASCOT results, we identified 62 023 unique peptides (9 253 proteins) with a q value of <0.01 at the PSM level. The q values can be assigned to each PSM and used to estimate the FDR. However, the application of an FDR of <1% at the PSM level resulted in an unacceptably high FDR of 59.7% at the protein level when combining the individual data set of 144 RAW files (Figure 2a).

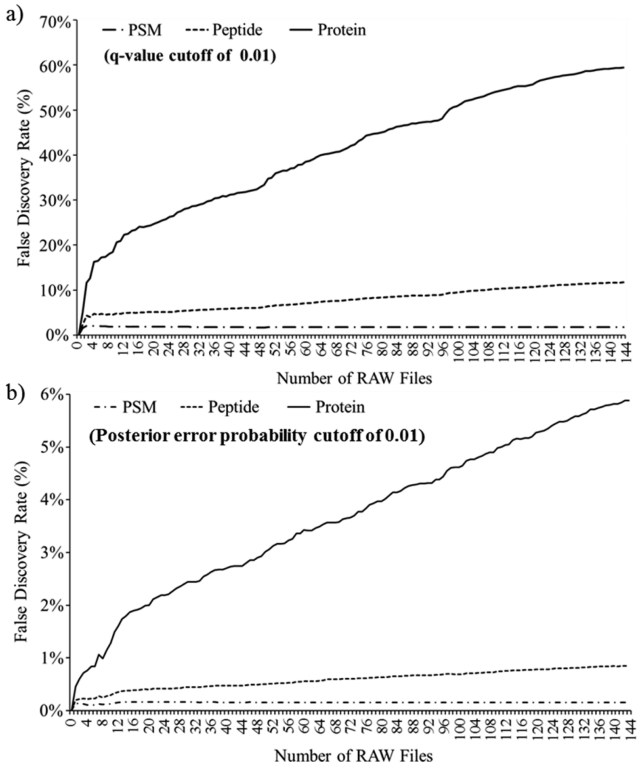


Figure 2. Comparison of accumulated false discovery rates (FDRs) among peptide spectrum matches (PSMs), peptides, and proteins from MASCOT search results using the human hippocampal data set. (a) Proportions of accumulated FDRs of proteins (black solid line), peptides (black dotted line), and PSMs (black dashed line) with accumulated estimated FDRs derived at the 1% q value from MASCOT search results. (b) Proportions of accumulated FDRs of proteins (black solid line), peptides (black dotted line), and PSMs (black dashed line) with accumulated estimated FDRs derived at 1% posterior error probabilities from the MASCOT search results.

On the contrary, to reduce the FDR at the protein level, we applied an advanced statistical method with a posterior error probability of <0.01 from the Percolator results at the PSM level. This filtering procedure resulted in the identification of a total of 52 262 unique peptide sequences from the 144 RAW files, representing 7225 proteins with a protein-level FDR of 5.9% (Figure 2b). These numbers were gradually increased by combining the PSMs from very large data sets from multiple experiments. Although these methods increase protein coverage, they also increase the protein-level FDR.[11] Therefore, protein inference must be carefully controlled in large-scale proteomic experiments.[12,16]

Improved of Protein Identification

Table 1 depicts the proteomic search results using SEQUEST, MASCOT, and MS-GF+ with the CPP, in which 5183, 5060, and 4977 proteins were identified, respectively, at a protein-level FDR of 1.0% with a threshold of two or more peptides for each search engine. Comparing protein matches among all three search engines and with any one of the search engines, a total of 661 proteins were identified, namely, 247 from Mascot, 176 from MS-GF+, and 238 proteins from SEQUEST (Figure 3a). These proteins are classified mainly as plasma membrane components ($5.96 \times 10^{-6} < p$ value $< 1.72 \times 10^{-4}$) with low abundance. When the protein data from the three search engines were combined, a total of 4453 (77.9%) proteins were commonly identified with a controlled FDR of 1.0% at the protein level. Because different search engines produce different peptide identifications, we employed more than one search engine, which could increase the number of proteins (and peptides) identified.[13] Jones et al.[14] demonstrated that combining search results using a well-controlled FDR not only enhanced the number of peptide identifications but also increased the confidence in these identifications. After applying the IPP using three different search engines to search, 5756 proteins with a controlled FDR of 1.0% at the protein level were identified against the neXtProt database, containing two or more peptides including at least one unique peptide in a protein from the neXtProt database (Table 1). Overall, a total of 63 895 peptides (345 736 PSMs) were identified, considering decoy hits from the decoy sequences with a protein-level FDR of 1.0% (Supplementary Excel Table 1). Among the 5756 proteins identified using the IPP, 4442 of them overlapped with the 4453 proteins identified from the combined protein data set using the three different search engines Therefore, 1314 proteins were uniquely identified using the IPP at the same protein FDR of <1% (Figure 3b). We believe that these 1314 proteins have low abundance (defined as proteins matching to fewer than four peptides). Assuming that low-abundance proteins with a small number of peptide assignments would be identified, higher percentages of assignments of peptide with

Table 1. Number of Proteins and Peptides Identified in the neXtProt Database Using the Conventional Proteomic Pipeline (CPP) versus Integrated Proteomic Pipeline (IPP)

	number of proteins	number of peptides	protein FDR (%)
SEQUEST	5183	60 925	1
MASCOT	5060	46 339	1
MS-GF+	4977	56 937	1
CPP	4453	66 571	1
IPP	5756	63 895	1

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

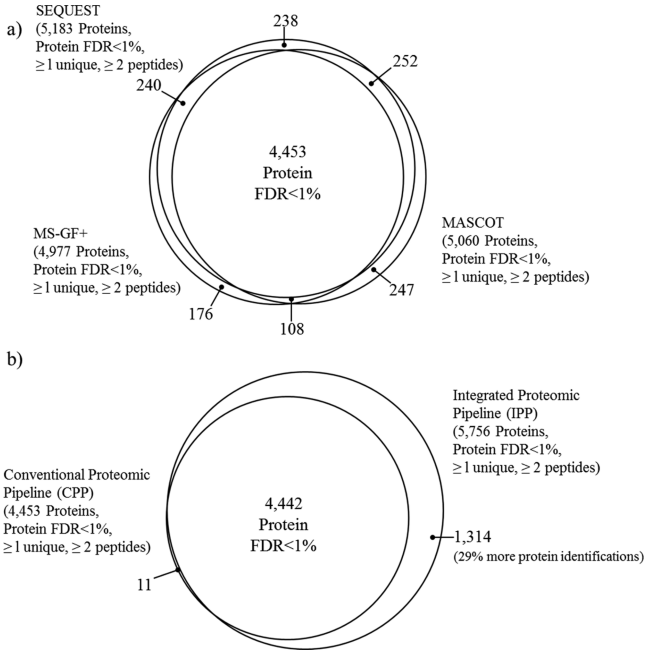


Figure 3. Comparison of the numbers of identified proteins from the neXtProt database using the integrated proteomic pipeline (IPP) versus conventional proteomic pipeline (CPP). (a) Number of proteins identified by three different proteomic search engines in the human hippocampal data sets. (b) Using the IPP, ~29% more high confidence proteins were identified than the overlapping 4453 proteins identified using the three different search engines at protein-level FDR of <1%.

a small number indicated that IPP can enhance the detection of low-abundance proteins

To evaluate the quality of the proteins identified using IPP, we initially examined the number of peptides per identified protein for all search results. Figure 4a shows the distribution of the number of identified peptides at a 1% protein-level FDR using SEQUEST, MASCOT, and MS-GF+ from ProteinInferencer. Using the CPP, proteins with assignments of two, three, or four peptides were lost, while such proteins were increased using the IPP (Figure 4b). Thus, use of the IPP may enhance the detection of low abundance proteins such as ASVs and missing proteins.

Alternative Splicing Variants

Paik and Hancock suggested that the identification and functional study of ASVs is one of the main objectives of the C-HPP.[30] Because ASVs have high homology from a single gene, it is necessary to apply a proteogenomic method to achieve accurate differentiation of ASV with one or more unique peptides rather than shared peptides identified from MS/MS spectra. In human hippocampal tissue, applying the CPP, we identified 359, 277, and 369 ASVs using SEQUEST, MASCOT, and MS-GF+, respectively (Supplementary Excel Table 2). When the ASVs from the three search engines were combined, a total of 182 (35.8%) ASVs were commonly identified with an FDR of 1% at the protein level (Figure 5a). On the contrary, after applying the IPP using the three different search engines, a total of 477 ASVs were identified, with a controlled FDR of 1.0% at the protein level including at least one unique peptide (Supplementary Excel Table 3). Comparing the CPP and IPP results, 15 and 310 ASVs, respectively, were identified in only one of these pipelines (Figure 5b).

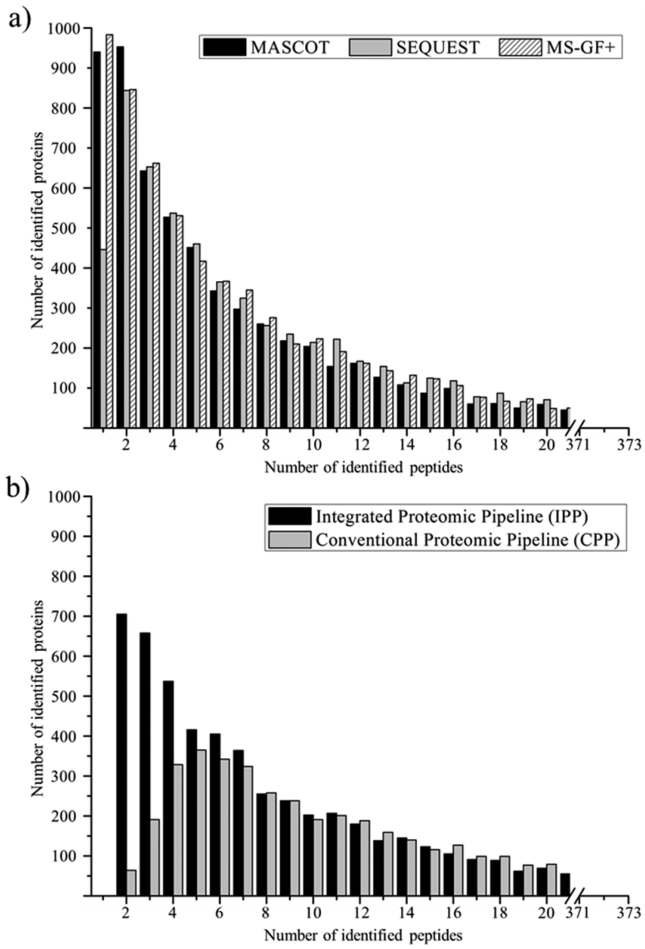


Figure 4. Comparison of the numbers of proteins identified using the integrated proteomic pipeline (IPP) versus conventional proteomic pipeline (CPP) using three different search engine results. (a) Distribution of unique peptide sequence of the identified proteins at a 1% false positive rate (FDR) using SEQUEST, MASCOT, and MS-GF+ with single hits. (b) Distribution of unique peptide sequence of the identified proteins at a 1% FDR using the IPP and CPP.

Within the controlled FDR at the protein level, the number of ASV unique peptides identified using the IPP results was greater than using the CPP (Figure 5c). The smaller the number of unique ASV peptides, the greater the number of identified ASVs. This indicates that the IPP could be used to maximize the number of identified peptides without a loss of accuracy, identifying 2.8 fold ASVs than the number identified using the CPP.

Missing Proteins

Missing proteins that lack sufficient experimental evidence from biological samples at the protein level are classified into five levels of protein evidence (PE), namely, evidence at the protein level (PE = 1), transcript level (PE = 2), inferred from homology (PE = 3), predicted (PE = 4), or uncertain (PE = 5), by the neXtProt database.[31] The published metrics for the Human Proteome Project 2015 contain a guideline for high-confidence identification of missing proteins and information for neXtProt (release 09, 2014), including 2948 missing and 616 uncertain proteins out of a total 20 055.[32] In February 2016, neXtProt reported a total number of human protein entries of 20 055, including 2949 missing and 588 uncertain proteins. Deutsch et al. suggested using a stringent FDR

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

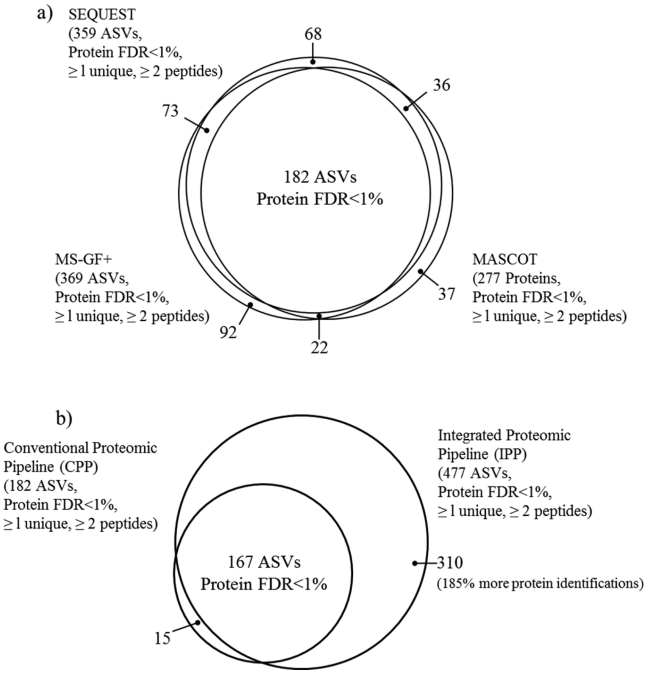


Figure 5. Comparison of the numbers of identified alternative splicing variants (ASVs) using the integrated proteomic pipeline (IPP) versus conventional proteomic pipeline (CPP). (a) A total of 182 ASVs were commonly identified among three different proteomic search engines in human hippocampus data sets at protein-level false discovery rate (FDR) of <1%. (b) Using the IPP, a total of 477 ASVs were identified at a protein-level FDR of <1%. Additionally, 310 ASVs were identified from human hippocampal tissues using the IPP. (c) Comparison of IPP and CPP in terms of the distribution of unique ASV peptides among the identified ASV proteins at a 1% FDR.

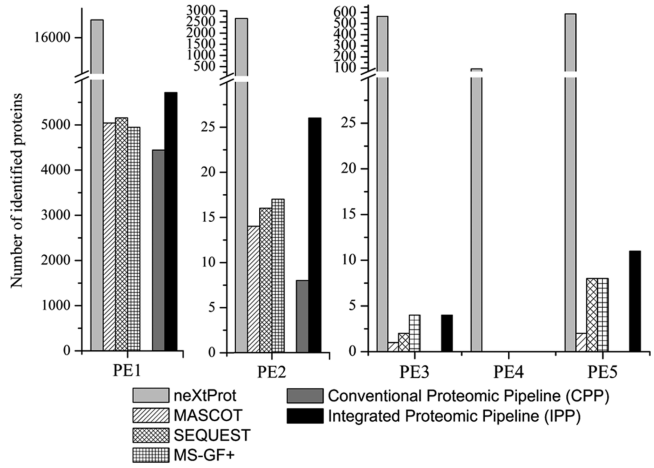


Figure 6. Distribution of protein evidence (PE) values obtained from the numbers of proteins identified in the neXtProt database using the conventional proteomic pipeline (CPP) versus integrated proteomic pipeline (IPP). The total number of proteins identified in neXtProt was compared with those identified using MASCOT, SEQUEST, and MS-GF+. The missing protein candidates were obtained from proteins identified at PE2, PE3, and PE4 levels.

threshold (<1% at the protein level) and discussed the misidentification of proteins in large data sets.[33]

In our study, using the IPP, a total of 25 missing protein candidates (PE2, PE3, and PE4) and 12 uncertain protein candidates (PE5) were found by two or more peptides at a protein-level FDR of 1% from the neXtProt database (release 09, 2014). Shown in Figure 6 are more missing proteins as well as uncertain proteins identified using the IPP than the CPP. In addition, the number of missing proteins identified using the IPP at the PE2 level was identified to be more than 3 times that using the CPP. Furthermore, uncertain proteins at the PE5 level were not identified using the CPP, but 12 were identified using with the IPP. However, we considered that the estimated FDR is an imperfect assumption, for identifying missing proteins; therefore, we performed a comparison of several reference databases and analyzed their corresponding synthetic peptide as follows. To obtain evidence of missing and uncertain protein, we examined the results from the repositories of GPMDB, ProteinAtlas (HPA), and PeptideAtlas (Table S1). The missing and uncertain protein candidates were present in at least one of the three repositories, while the 10 missing and 1 uncertain protein candidates were found in all three repositories, and the 10 missing and 3 uncertain protein candidates were uniquely identified using the IPP alone. We recognized that 10 missing protein candidates containing a small number of identified peptides were accepted at the PE1 level in the latest version of neXtProt (release 02, 2016) Additionally, nine of uncertain proteins were categorized into "canonical" or "marginally distinguished" in the latest version of Peptide Atlas (Table S1).

Using the IPP, we identified 10 missing proteins from neXtProt (release 09, 2014), as shown in Table 2, which corresponds to 1.5 times more proteins than the number identified using the CPP. This indicates that our IPP has the ability to identify low abundance proteins, as discussed in Figure 4. Comparing the IPP and CPP results, four missing proteins were identified by the IPP only (Table 2). In accordance with the new data interpretation guidelines for the C-HPP, we manually validated and compared all spectra of the missing protein candidates and their corresponding synthetic peptides at the PSM level. As a result, we finally claimed two missing proteins [glutamate receptor ionotropic kainate 5 (GRIK5, NX_Q16478) and ecotropic viral integration site 2A (EVI2A, NX_P22794)] against the latest version of neXtProt (release 02, 2016) (Table 2). For example, two different peptides, SFNYPSASLICAK and LYSAGAGGD-AGSAHGGPQR, from NX_Q16478 of GRIK5 were identified and matched to their corresponding synthetic peptides well (Figure 7).

■ CONCLUSIONS

In proteogenomic studies, false-positive protein identifications generated by large PSMs after database searching tend to result in higher protein-level FDRs. Here we have developed an IPP to improve the confidence in protein identification with a controlled protein-level FDR of <1% based on the integration of validated peptide search results using SEQUEST, MASCOT, and MS-GF+ with the neXtProt (release 09, 2014) database.

At a protein-level FDR controlled at <1%, 5756 proteins in human hippocampal tissues were identified using IPP in comparison with 4453 proteins using CPP. We identified 477 ASVs and 10 missing proteins using IPP, which corresponds to 2.8 and 1.5 times more proteins than when using CPP, respectively. Therefore, we were able to identify more missing proteins with a low abundance than using CPP at a given level of global protein-level FDR control. Further use of IPP should

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

Table 2. Peptides Identified from Missing Proteins in Human Hippocampal Tissues Using the neXtProt Database

	proteins accessions (description)	PE level (release 09.2014, neXtProt)	PE level (release 02.2016, neXtProt)	method (IPP or CPP)	peptides	unique/ shared	peptide E-score	ppeptide FDR (%)	number of spectra	PSM FDR (%)
1	NX_Q16478 (glutamate receptor ionotropic, kainate 5)	PE2	PE2	IPP	LYSAGAGGDAGSAHGGPQR (3+)[a]	unique	1.8	0.00	6	0.00
					SFNYPSASLICAK (2+)[a]	unique	−0.4	0.13	3	0.04
					EIGVWYSNRTLAMNATTLDINLSQTLANK (5+)	unique	−0.5	0.22	1	0.06
					LSNGKLYSAGAGGDAGSAHGGPQRLLDDPGPPSGAR (4+)	unique	−0.7	0.36	1	0.09
2	NX_P22794 (protein EVI2A)	PE2	PE2	IPP/CPP	QLTGPNLVMQSTGVLTATR[a]	unique	3.8	0.00	10	0.00
					SNGDFLASGLWPAESDTWK[a]	unique	1.5	0.00	3	0.00
					SNGDFLASGLWPAESDTWKR[a]	unique	0.7	0.01	3	0.00
3	NX_Q5JV73 (FERM and PDZ domain-containing protein 3)	PE2	PE1	IPP/CPP	TYSLAVHPALSPQLSEQK (2+)	unique	3.0	0.00	1	0.00
					IQSCAAQAVLTAPYSLGRPDPNPSLQPIATGQSPGPPGAR (2+)	unique	1.2	0.00	3	0.00
4	NX_Q7Z407 (CUB and sushi domain-containing protein 3)	PE2	PE1	IPP/CPP	TCQLNGHWSGSQPHCSGDATGTCGDPGTPGHGSR (2+)	unique	0.6	0.01	6	0.00
					ICQQDHNWSGQLPSCVPVSCGHPGSPIYGR (4+)	unique	0.1	0.02	6	0.01
5	NX_O43300 (leucine-rich repeat transmembrane neuronal protein 2)	PE2	PE1	IPP	DQFASFSQLTWLHLDHNQISTVK (2+)	unique	0.0	0.03	3	0.01
					LRELHLEHNQLTK (4+)	unique	−0.5	0.21	2	0.06
6	NX_Q9ULB4 (cadherin-9)	PE2	PE1	IPP/CPP	RTVPLWENIDVQDFIHR (2+)	unique	1.0	0.00	3	0.00
					VVYSILQGQPYFSVDPESGIIK (2+)	unique	−0.2	0.06	3	0.02
					EQYQVVIQAK (2+)	shared	−0.3	0.09	4	0.02
7	NX_P0CW23 (uncharacterized protein C18orf42)	PE4	PE1	IPP/CPP	DHIQLGVGELTK (2+)	unique	2.8	0.00	10	0.00
					QIVQNAILQAVQQVSQESQR (2+)	unique	2.4	0.00	8	0.00
					VFAPGEKPGNEPEEVKLQNASK (4+)	unique	0.6	0.01	2	0.00
8	NX_Q8IYE0 (coiled-coil domain-containing protein 146)	PE2	PE1	IPP	DLTEKEM(15.9949)IQKLDKLELQLAKK (3+)	unique	−0.1	0.04	2	0.01
					LCSKTQGCKQDTLLLAKKMNGYQRRIK (4+)	unique	−0.4	0.12	3	0.03
					ERHKM(15.9949)SLNELEILR (4+)	unique	−0.6	0.29	1	0.08
					HANNVTIRESMQNDVRKIVSK (2+)	unique	−1.1	0.38	1	0.09
9	NX_C4AMC7 (putative WAS protein family homologue 3)	PE2	PE1	IPP/CPP	MQHSLAGQTYAVPLIQPDLR (2+)	unique	3.7	0.00	8	0.00
					ATLLESIR (2+)	unique	1.6	0.00	5	0.00
					THVMLGAETEEK (2+)	unique	1.6	0.00	8	0.00
					EEAVQQMADALQYLQK (2+)	unique	1.1	0.00	6	0.00
					YVFLDPLAGAVTK (2+)	unique	0.6	0.01	4	0.00
					KYVFLDPLAGAVTK (2+)	unique	−0.4	0.12	3	0.03
10	NX_Q9Y5F6 (protocadherin gamma-C5)	PE2	PE1	IPP/CPP	EATSHYIIELLASDAGSPSLHK (2+)	unique	1.4	0.00	5	0.00
					VTAVDADAGHNAWLSYSLLPQSTAPGLFLVSTHTGEVR (2+)	unique	0.2	0.02	5	0.00
					VGIPENAPIGTLLLR (2+)	unique	−0.4	0.13	3	0.04
					QNVYIPGSNATLTNAAGK (2+)	shared	3.7	0.00	6	0.00
					YGPQFTLQHVPDYR (2+)	shared	1.5	0.00	10	0.00
					YGPQFTLQHVPDYRQNVYIPGSNATLTNAAGK (2+)	shared	−0.5	0.21	2	0.06

[a]Compared with the spectrum of their corresponding synthetic peptide.

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

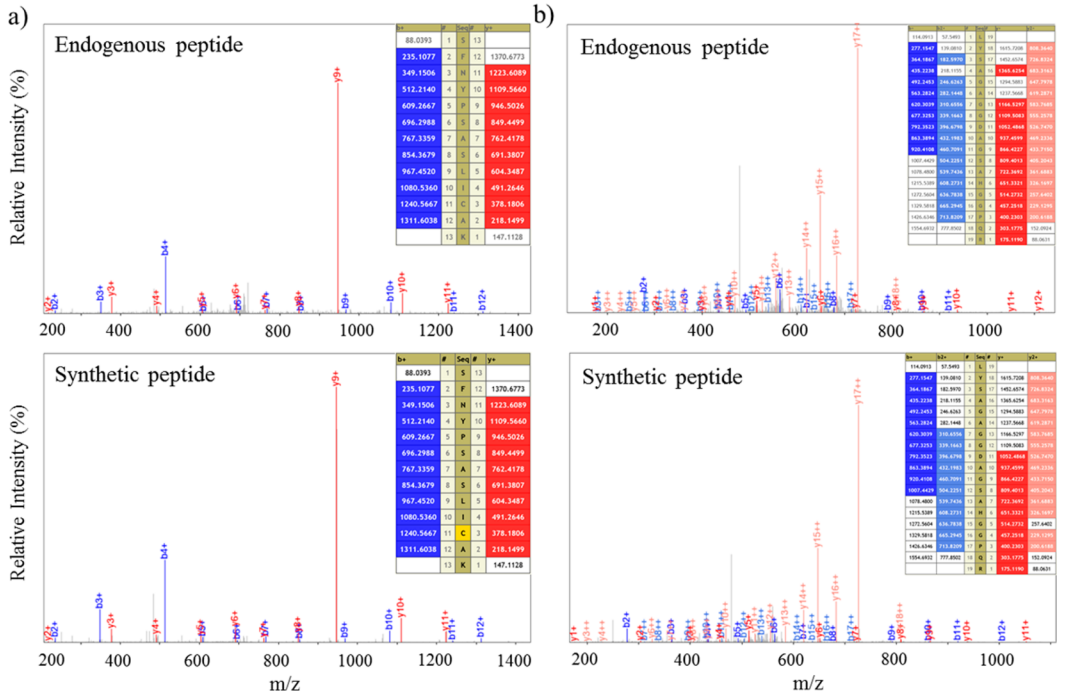


Figure 7. Example of tandem mass spectrometry (MS/MS) spectra of glutamate receptor ionotropic kainate 5 protein, which is one of the missing proteins identified by the integrated proteomic pipeline (IPP). Two different peptides, (a) SFNYPSASLICAK (2+) and (b) LYSAGAGGDAGSAHGGPQR (3+) from NX_Q16478 of GRIK5, were identified. MS/MS spectra of an identified endogenous peptide from a tissue sample (top) and the corresponding synthetic peptide (bottom) showing the same fragmentation pattern for validation.

be of great benefit to proteogenomic analyses in C-HPP studies to identify more ASVs as well as missing proteins.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00376.

Figure S1. Calculation of E-values. Figure S2. Tandem mass spectra of endogenous peptides of ecotropic viral integration site 2A (NX_P22794) protein and their corresponding synthetic peptides. Table S1. Candidates of 27 missing and 12 uncertain proteins identified from the neXtProt database using the integrated proteomic pipeline (IPP). (PDF)

Supplementary Excel Table 1. List of proteins and peptides identified, including decoy hits, using the integrated proteomic pipeline (IPP) from the neXtProt database. (XLSX)

Supplementary Excel Table 2. List of identified alternative splicing variants using CPP. (XLSX)

Supplementary Excel Table 3. List of identified alternative splicing variants using IPP. (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*J.S.Y.: Tel: +82-43-240-5145. Fax: +82-240-5159. E-mail: jongshin@kbsi.re.kr.

*J.Y.K.: E-mail: jinyoung@kbsi.re.kr.

Author Contributions

#G.W.P. and H.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

All RAW files used in this study were deposited in ProteomeXchange (PXD000395).

■ ACKNOWLEDGMENTS

This research was supported by the National Research Council of Science and Technology (CAP-15-03-KRIBB); by the Korea Health Technology R&D Project, through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (HI13C2098); and by the research program through the Korea Basic Science Institute (G35110).

■ ABBREVIATIONS

FDR, false discovery rate; CID, collision-induced dissociation; ETD, electron-transfer dissociation; PSM, peptide spectrum match; CPP, conventional proteomic pipeline; IPP, integrated proteomic pipeline; C-HPP, Chromosome-Centric Human Proteome Project; ASV, alternative splicing variant

■ REFERENCES

(1) Arthur, J. W.; Wilkins, M. R. Using proteomics to mine genome sequences. *J. Proteome Res.* **2004**, *3* (3), 393−402.

(2) Jaffe, J. D.; Berg, H. C.; Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004**, *4* (1), 59−77.

(3) Brosch, M.; Saunders, G. I.; Frankish, A.; Collins, M. O.; Yu, L.; Wright, J.; Verstraten, R.; Adams, D. J.; Harrow, J.; Choudhary, J. S.; Hubbard, T. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.* **2011**, *21* (5), 756−67.

(4) Khatun, J.; Yu, Y.; Wrobel, J. A.; Risk, B. A.; Gunawardena, H. P.; Secrest, A.; Spitzer, W. J.; Xie, L.; Wang, L.; Chen, X.; Giddings, M. C.

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090

Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **2013**, *14*, 141.
(5) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114−25.
(6) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J. Proteome Res.* **2011**, *10* (1), 210.
(7) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.
(8) 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.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−81.
(9) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−7.
(10) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513* (7518), 382−7.
(11) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (11), 2405−17.
(12) Ezkurdia, I.; Vazquez, J.; Valencia, A.; Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **2014**, *13* (8), 3854−5.
(13) Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol. Cell. Proteomics* **2007**, *6* (9), 1599−608.
(14) Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **2009**, *9* (5), 1220−9.
(15) Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J. Proteome Res.* **2009**, *8* (8), 3872−3881.
(16) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. Molecular & cellular proteomics. *Mol. Cell. Proteomics* **2015**, *14*, 2394.
(17) 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* (11), 976−89.
(18) 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* (18), 3551−67.
(19) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7* (8), 3354−63.
(20) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1* (1), 21−6.
(21) Kall, 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), 923−5.
(22) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176−81.
(23) Zhang, Y.; Xu, T.; Shan, B.; Hart, J.; Aslanian, A.; Han, X.; Zong, N.; Li, H.; Choi, H.; Wang, D.; Acharya, L.; Du, L.; Vogt, P. K.; Ping, P.; Yates, J. R., 3rd ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. *J. Proteomics* **2015**, *129*, 25−32.
(24) Kwon, K. H.; Kim, J. Y.; Kim, S. Y.; Min, H. K.; Lee, H. J.; Ji, I. J.; Kang, T.; Park, G. W.; An, H. J.; Lee, B.; Ravid, R.; Ferrer, I.; Chung, C. K.; Paik, Y. K.; Hancock, W. S.; Park, Y. M.; Yoo, J. S. Chromosome 11-centric human proteome analysis of human brain hippocampus tissue. *J. Proteome Res.* **2013**, *12* (1), 97−105.
(25) Hwang, H.; Park, G. W.; Kim, K. H.; Lee, J. Y.; Lee, H. K.; Ji, E. S.; Park, S. K.; Xu, T.; Yates, J. R., 3rd; Kwon, K. H.; Park, Y. M.; Lee, H. J.; Paik, Y. K.; Kim, J. Y.; Yoo, J. S. Chromosome-Based Proteomic Study for Identifying Novel Protein Variants from Human Hippocampal Tissue Using Customized neXtProt and GENCODE Databases. *J. Proteome Res.* **2015**, *14* (12), 5028−37.
(26) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; Bairoch, A.; Lane, L. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12* (1), 293−8.
(27) 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* (4), 768−74.
(28) Alves, G.; Ogurtsov, A. Y.; Wu, W. W.; Wang, G.; Shen, R. F.; Yu, Y. K. Calibrating E-values for MS2 database search methods. *Biol. Direct* **2007**, *2*, 26.
(29) Eng, J. K.; Fischer, B.; Grossmann, J.; Maccoss, M. J. A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **2008**, *7* (10), 4598−602.
(30) Paik, Y. K.; Hancock, W. S. Uniting ENCODE with genome-wide proteomics. *Nat. Biotechnol.* **2012**, *30* (11), 1065−7.
(31) Lane, L.; Argoud-Puy, G.; Britan, A.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gaudet, P.; Gleizes, A.; Masselot, A.; Zwahlen, C.; Bairoch, A. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* **2012**, *40* (Database issue), D76−83.
(32) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**, *14* (9), 3452−60.
(33) Deutsch, E. W.; Sun, Z.; Campbell, D.; Kusebauch, U.; Chu, C. S.; Mendoza, L.; Shteynberg, D.; Omenn, G. S.; Moritz, R. L. State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet. *J. Proteome Res.* **2015**, *14* (9), 3461−73.

DOI: 10.1021/acs.jproteome.6b00376
J. Proteome Res. 2016, 15, 4082−4090