Journal of proteome research

Article

pubs.acs.org/jpr

Enhanced Missing Proteins Detection in NCI60 Cell Lines Using an Integrative Search Engine Approach

Elizabeth Guruceaga,[†,‡] Alba Garin-Muga,[†] Gorka Prieto,[§] Bartolomé Bejarano,[⊥] Miguel Marcilla,[¶] Consuelo Marín-Vicente,[#,∥] Yasset Perez-Riverol,[▲] J. Ignacio Casal,[#] Juan Antonio Vizcaíno,[▲] Fernando J. Corrales,[¶] and Victor Segura*[,†,‡]

[†]Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona 31008, Spain

[‡]IdiSNA, Navarra Institute for Health Research, Pamplona 31008, Spain

[§]Department of Communications Engineering, University of the Basque Country (UPV/EHU), Bilbao 48013, Spain

[⊥]Fundación Jímenez Díaz, Madrid 28040, Spain

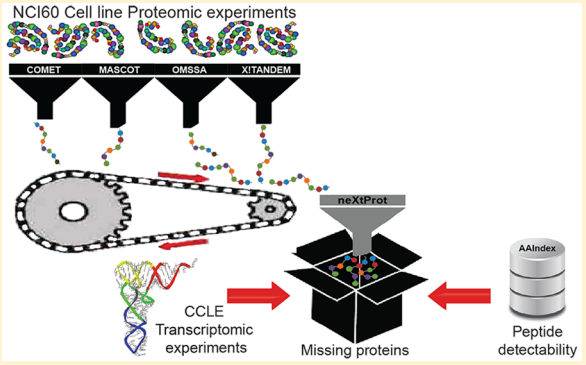
[¶]Proteomics Unit, Spanish National Biotechnology Centre, CSIC, Madrid 28049, Spain

[#]Functional Proteomics, Department of Cellular and Molecular Medicine and [∥]Proteomic Facility, Centro de Investigaciones Biológicas (CIB-CSIC), Ramiro de Maeztu 9, Madrid 28040, Spain

[▲]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

S *Supporting Information*

ABSTRACT: The Human Proteome Project (HPP) aims deciphering the complete map of the human proteome. In the past few years, significant efforts of the HPP teams have been dedicated to the experimental detection of the missing proteins, which lack reliable mass spectrometry evidence of their existence. In this endeavor, an in depth analysis of shotgun experiments might represent a valuable resource to select a biological matrix in design validation experiments. In this work, we used all the proteomic experiments from the NCI60 cell lines and applied an integrative approach based on the results obtained from Comet, Mascot, OMSSA, and X! Tandem. This workflow benefits from the complementarity of these search engines to increase the proteome coverage. Five missing proteins C-HPP guidelines compliant were identified, although further validation is needed. Moreover, 165 missing proteins were detected with only one unique peptide, and their functional analysis supported their participation in cellular pathways as was also proposed in other studies. Finally, we performed a combined analysis of the gene expression levels and the proteomic identifications from the common cell lines between the NCI60 and the CCLE project to suggest alternatives for further validation of missing protein observations.



KEYWORDS: *C-HPP, missing proteins, NCI60, CCLE, integration of search engines, peptide detectability*

INTRODUCTION

Since 2010, the Human Proteome Project (HPP)[1,2] has brought together the efforts of the international research community in the field of proteomics, bioinformatics, and molecular biology to (1) define the complete catalog of human proteins (C-HPP initiative[3]) and (2) study the functions of proteins in biology and disease (B/D-HPP initiative[4−6]). Although there have been successful scientific and technological advances over these years,[7−12] significant challenges still remain uncovered. In terms of the human proteome characterization, the main objective is the detection of the proteins without sufficient experimental evidence using mass-spectrometry, also known as the "missing proteins" or "missing proteome".[13]

The neXtProt human protein knowledgebase[14] (https://www.nextprot.org) has been consolidated as the key resource for the evaluation of the C-HPP initiative advances in the description of the human proteome. In this database, different experimental evidence categories are assigned to each protein. The codes PE2 (experimental evidence at transcript level), PE3 (protein inferred from homology), and PE4 (predicted protein) correspond to missing proteins, while PE1 is the annotation for proteins with strong evidence from mass spectrometry or other experimental methods, and PE5 is the code for uncertain proteins. neXtProt not only includes the most up-to-date annotation of the human proteome, but also other information

Special Issue: Chromosome-Centric Human Proteome Project 2017

Received: June 6, 2017
Published: September 29, 2017

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

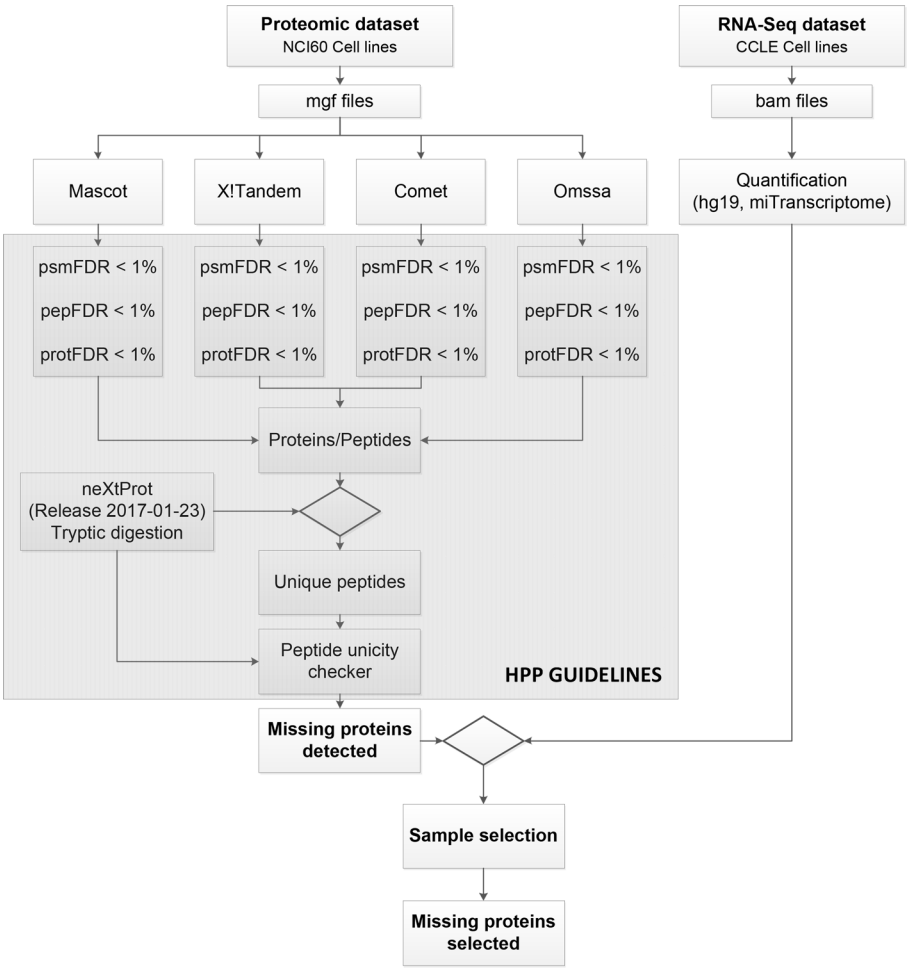


Figure 1. (A) Overall scheme of the analysis pipeline developed to identify missing proteins. An integrative strategy based on the results of four search engines was used with the shotgun experiments of the NCI60 data set and the RNA-Seq experiments of the CCLE project.

resources of great relevance for the proteomic research community. For example, in recent versions it is possible to verify if a certain peptide is unique to a protein (proteotypic definition by HUPO) and if it has been detected in a biological matrix. For this purpose, neXtProt included the "Peptide uniqueness checker" utility to determine the unambiguous peptides of a protein taking into account the known variants stored in the database (more than 5 million SNPs and disease mutations).[15] In this way, this resource allows compliance with the C-HPP guidelines for the proper detection of a protein in an MS experiment.[16] These guidelines consider an MS evidence as accurate if the following thresholds are applied: 1% FDR at PSM, peptide, and protein level, and more than one unique peptide detected of nine or more amino acids and without any ambiguity in their spectrum assignment (SNPs, mutations, or incorrect amino acid assignments).

Different bioinformatic workflows were implemented to detect missing proteins using alternative methods in a shotgun experiment.[12,17,18] Some of the obtained proteins that were identified with only one peptide, also known as one-hit wonders, did not fulfill the specified C-HPP criteria. Therefore, they can only be considered when further validation of the presence of the protein in the sample is provided. The validation of the detected peptides can include targeted proteomic experiments using synthetic peptide standards and manual evaluation and annotation of the obtained high-resolution spectra for each peptide. However, because of their low expression or their tissue specificity, the selection of the biological matrix in which these proteins are expressed is one of the main difficulties for the design of the proteomic experiments.[8,10,12,19−21] Several algorithms were developed to predict where the probability of detection of missing proteins is higher using, for example, an integrative approach based on publicly available genomic, transcriptomic, and proteomic experiments.

In one of these studies, the authors developed a bayesian classifier to guide the search of missing proteins based on the analysis of thousands of microarray experiments obtained from the Gene Expression Omnibus (GEO) database.[10] Another study analyzed RNA-Seq experiments from the ENCODE project and the Illumina Human Body Map 2.0 (HBM)[22] obtaining a database of expressed proteins along different normal tissues and cell lines. Finally, a different approach was implemented using a set of shotgun experiments obtained from the PRIDE database[12,23] as source of information. All previous results suggested testis, brain, skeletal muscle, and embryonic tissues as promising biological sample sources due to their enriched number of expressed missing proteins. The in-depth analysis of the spermatozoa proteome[8,20] and the HEK293 cell line[12] confirmed the validity of the mentioned methodologies. However, even in these cases, improved bioinformatic methods

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

and manual curation of the mass spectra corresponding to the detected peptides were required.

In this manuscript, we propose the integration of all the proteins identified using four different search engines (Comet, Mascot, OMSSA, and X!Tandem) from the analysis of all the cell lines available in the NCI60 data set.[24] Previous studies using different statistical approaches demonstrated that the analysis of the results using different search engines effectively increases the coverage of a proteome using a single shotgun experiment.[25−27] Such a study was even applied for the detection of missing proteins.[11] In this case, an unique FDR for the integration of the results obtained with all the search engines was calculated but only one biological sample was analyzed. Here we present the analysis of a huge proteomic data set consisting of 59 cell lines (NCI60) with more than 900 fractions in total using four search engines. The union of the results of the four search engines increases the number of identifications and has proven to be crucial in the detection of the missing proteins.[28] We were able to integrate the results of more than 3600 proteomic searches using a sample-based strategy for the calculation of the FDR. The NCI60 cell lines are available in molecular laboratories all over the world, being good candidates for the validation of our results. In addition, we performed a peptide detectability study for the unique peptides, detected or not in different analysis configurations, using a classifier approach based on the peptide information stored in the GPMDB database.[29] The obtained predictions were used to evaluate the results from the different search engines, and we were able to draw several conclusions, which could be relevant to increase our knowledge about the missing proteins from a computational and a biological point of view.

We detected five missing proteins with two unique peptides and less than 1% FDR at PSM, peptide, and protein levels: FREM3 (Chr 4), LAMB4 (Chr 7), MYEOV (Chr 11), RAD21L1 (Chr 20), and TLDC2 (Chr 20). Validation experiments with synthetic peptides were performed for these proteins with questionable results. In addition, we detected 165 one-hit wonder missing proteins that should be further validated. These validation process starts with the selection of those cell lines in which the probability of finding these missing proteins is high. To do so, the results obtained for the mentioned NCI60 experiments were combined with the analysis of the gene expression profiles in the set of RNA-Seq samples available in the CCLE project.[30] This information is provided for the laboratories involved in the C-HPP initiative to facilitate the design of targeted proteomic experiments for the one-hit wonders of their corresponding chromosomes in the most appropriate cell lines.

MATERIAL AND METHODS

Bioinformatic Workflow

In this work, we developed a bioinformatic workflow (Figure 1) for the detection of missing proteins based on three pillars: (1) the strict application of the C-HPP guidelines for the detection of proteins using MS/MS experiments; (2) the analysis of shotgun experiments of 59 different cell lines using an integrative approach based on four search engines; and (3) the quantification of the expression level of the protein coding genes in these cell lines as a guidance for predicting the suitable sample sources for the targeted proteomic validation experiments.

The main goal of the method we propose is to increase the proteome coverage obtained from the analysis of a given proteomic experiment, increasing our capacity to find missing proteins through the reanalysis of public experiments. Briefly, the pipeline takes into account all the peptide identifications from the search engines used. In this way, peptide and protein identifications following the C-HPP guidelines using each of the searching algorithms contribute to the total number of proteins detected, including missing proteins.

To combine the strengths of different approaches, we selected four search engines with a different strategy for peptide detection: one commercial (Mascot) and three open source ones (X!Tandem, Comet, and OMSSA). Mascot uses a probabilistic scoring algorithm adapted from the MOWSE algorithm, which is a methodological approach to detect peptides based on the calculation of the probability of whether an observed PSM has occurred by chance. The peptide detection with the lowest probability of occurring by chance is returned as the most significant one.[31] Instead, X!Tandem represents the experimental spectrum using only peaks that match peaks in the theoretical spectrum and then calculates the dot product. The scoring algorithm is called hyperscore, which is based on the number of assigned b and y ions using the hypergeometric distribution.[32] X!Tandem uses this score distribution to extrapolate empirical *E*-values and assess the significance of a PSM. On the other hand, Comet is a search engine originated from the University of Washington's academic version of SEQUEST. It implements a fast cross-correlation algorithm[33] to score the PSMs in a shotgun experiment. For every candidate peptide in the protein database, the cross-correlation is calculated by a simple sum of peak intensities at each calculated fragment ion mass. This eliminates the need to create theoretical spectra. The score histogram is then used to generate an expectation value or *E*-value.[34] Finally, OMSSA ranks the detected peptide matches using a probability score developed using classical hypothesis testing, the same statistical method used in BLAST.

We tested our bioinformatic pipeline with proteomic and transcriptomic public data available: shotgun experiments from the NCI60 project and RNA-Seq experiments from the CCLE data set.

Proteomic and Transcriptomic Public Data Sets of Cell Lines

The NCI60 anticancer drug screen was developed in the late 1980s by the US National Cancer Institute (NCI) to identify compounds with growth-inhibitory or toxic effects on particular tumor types. As a result, panels of cell lines were assembled that represented nine distinct tumor types: breast, brain, colon, leukemia, lung, melanoma, ovarian, prostate, and renal tumors. On the other hand, the Cancer Cell Line Encyclopedia (CCLE) project is a collaboration between the Broad Institute, and the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation to conduct a detailed genetic and pharmacologic characterization of a large panel of human cancer models. To do so, they developed integrated computational analyses that link distinct pharmacologic vulnerabilities to genomic patterns and to translate these genomic patterns into cancer patient stratification. The CCLE provides public access to genomic data, analysis, and visualization for about 1000 cell lines.

We found 43 cell lines with both proteomic and transcriptomic experiments available (Table 1). Using these

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

Table 1. Molecular Characteristics of the Cell Lines for Which Shotgun and RNA-Seq Experiments Were Available[a]

cell line	disease	tissue of origin	epithelial	source	p53	transcriptomics	proteomics
BT-549	BRCA	breast	yes	metastasis		RNA-Seq	shotgun
Hs 578T	BRCA	breast	yes	primary	MT	RNA-Seq	shotgun
MCF-7	BRCA	breast	yes	pleural effusion	WT	RNA-Seq	shotgun; deep proteome
MDA-MB-231	BRCA	breast	yes	pleural effusion	MT	RNA-Seq	shotgun
T-47D	BRCA	breast	yes		MT	RNA-Seq	shotgun
HCT 116	COAD	colon	yes			RNA-Seq	shotgun
HCT 15	COAD	colon	yes			RNA-Seq	shotgun
HT-29	COAD	colon	yes	primary	MT	RNA-Seq	shotgun
KM12	COAD	colon	yes		MT	RNA-Seq	shotgun
SW620	COAD	colon	yes		MT	RNA-Seq	shotgun
786-O	KIRC	renal	yes		MT	RNA-Seq	shotgun
A-498	KIRC	renal	yes		WT	RNA-Seq	shotgun
ACHN	KIRC	renal	yes		WT	RNA-Seq	shotgun
Caki-1	KIRC	renal	yes	metastasis	WT	RNA-Seq	shotgun
UO-31	KIRC	renal	yes		WT	RNA-Seq	shotgun
HL-60	LCLL	leukemia	no	PBL	MT	RNA-Seq	shotgun
K-562	LCLL	leukemia	no	pleural effusion	MT	RNA-Seq	shotgun
SF295	LGG	cns	no		MT	RNA-Seq	shotgun
SF268	LGG	cns	no		MT	RNA-Seq	shotgun
SF539	LGG	cns	no		WT	RNA-Seq	shotgun
SNB-75	LGG	cns	no		MT	RNA-Seq	shotgun
A-549	LUSC	non-small cell lung	yes		WT	RNA-Seq	shotgun
EKVX	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun
HOP-62	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun
HOP-92	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun (n = 2)
NCI-H226	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun
NCI-H23	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun
NCI-H460	LUSC	non-small cell lung	yes	pleural effusion	WT	RNA-Seq	shotgun; deep proteome
NCI-H522	LUSC	non-small cell lung	yes		MT	RNA-Seq	shotgun
RPMI-8226	MM	leukemia	no	PB	WT	RNA-Seq	shotgun
IGROV-1	OV	ovarian	yes		MT	RNA-Seq	shotgun
OVCAR-3	OV	ovarian	yes	ascites	MT	RNA-Seq	shotgun
OVCAR-4	OV	ovarian	yes		WT	RNA-Seq	shotgun
OVCAR-8	OV	ovarian	yes		MT	RNA-Seq	shotgun
SK-OV-3	OV	ovarian	yes	ascites		RNA-Seq	shotgun; deep proteome
DU145	PRAD	prostate	yes	metastasis		RNA-Seq	shotgun
PC-3	PRAD	prostate	yes		MT	RNA-Seq	shotgun; deep proteome
LOX-IMVI	SKCM	melanoma	no		WT	RNA-Seq	shotgun
Malme-3M	SKCM	melanoma	no	metastasis	WT	RNA-Seq	shotgun
SK-MEL-28	SKCM	melanoma	no		MT	RNA-Seq	shotgun
SK-MEL-5	SKCM	melanoma	no	metastasis	WT	RNA-Seq	shotgun
UACC-257	SKCM	melanoma	no		WT	RNA-Seq	shotgun
UACC-62	SKCM	melanoma	no		WT	RNA-Seq	shotgun

[a]WT, cell line with wild-type P53; MT, cell line with mutant P53.

experiments, we compared the expression levels of protein coding genes and the number of detected proteins in the shotgun experiments.

Analysis of Shotgun Proteomic Data

All the proteomic experiments available from the NCI60 cell lines (61 shotgun experiments and 9 deep proteomes) were downloaded from the NCI60 proteome resource (http://129.187.44.58:7070/NCI60/main/index). This database stores the proteome profile of the cell lines performed using a conventional one-dimensional PAGE followed by in-gel digestion and liquid chromatography−tandem mass spectrometry (GeLC−MS/MS) approach with an LTQ Orbitrap XL ETD mass spectrometer.[35] To increase the tissue-specific proteome coverage, one cell line from each of the nine tissues represented was analyzed in more depth with an Orbitrap Elite mass spectrometer (deep proteomes).[36]

We converted raw data files to MGF files using the MSConvertGUI software. For each cell line, 12 fractions in the case of shotgun experiments and 24 fractions in the case of deep proteomes were generated, given a total number of more than 900 MGF files to analyze. The protein identification analyses were performed following the C-HPP guidelines for the identification of proteins using MS/MS experiments.[16] We searched all the MGF files against the UniprotKB human database (release 2017.01.v2) using the target-decoy strategy. Decoy database was created using the peptide pseudoreversed method, and separate searches were performed for target and decoy databases.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

Searches were performed using the following four search engines: Comet v. 2016.01 rev. 2,[37] an in-house Mascot Server v. 2.3 (Matrix Science, London, U.K.), OMSSA v. 2.1.9,[38] and X!Tandem v. 2015.12.15.2.[39] In all the cases, search parameters were set as follows: carbamidomethylation of cysteine as a fixed modification and oxidation of methionine as variable modification. Precursor and fragment mass tolerance were set to 10 ppm and 0.05 Da, respectively, for deep proteome data sets and 20 ppm and 0.5 Da, respectively, for proteome profile data sets, and one missed cleavage was allowed. FDR at the PSM, peptide and protein level were calculated using in-house scripts written in R/Bioconductor[40] (see the Supporting Information for more details and R code). Protein identifications were obtained applying the criteria of PSM FDR < 1%, peptide FDR < 1%, and protein FDR < 1% following the C-HPP guidelines and then converted to neXtProt protein accessions (neXtProt release 2017−02). Each proteomic experiment was analyzed independently, and the resulting sets of proteins and peptides detected for each sample were compared afterward. The use of a global FDR is mandatory by the C-HPP guidelines when the results are obtained from the combination of the samples analyzed. This strategy is not the one used in our analysis because our main goal is to find MS evidence for missing proteins in a given sample. Nevertheless, the global FDR was calculated and included in Supporting Table 5.

Protein inference process was greatly simplified using exclusively unique peptides to ensure reliable identifications as required by C-HPP initiative. Proteogest software[41] was used to perform the *in silico* digestion of the proteins contained in the neXtProt database and only those proteins with unique peptides between 9 and 30 amino acid in lengths were selected for further analyses. We applied the standard rules of cleavage for trypsin enzyme and allowed oxidation of methionine and one missed cleavage.

Detection of Missing Proteins

In this manuscript, we propose the integration of the results obtained from the analysis of the same shotgun experiment using different search engines (Comet, Mascot, OMSSA, and X!Tandem) as an alternative method to increase the number of missing protein identifications in a biological sample. Once peptides and proteins for sample were identified as stated before, the neXtProt protein evidence codes were used to distinguish the class of the missing proteins (PE2, PE3, and PE4). Then the neXtProt peptide uniqueness checker tool[15] was used to remove unique peptides that were considered ambiguous taking into account SNPs and disease mutations. After applying this filter, we obtained a set of proteins with two or more different unique peptides detected by at least one of the search engines. In addition to these missing proteins with mass-spectrometry based evidence of their presence in a biological matrix, we also found a set of missing proteins with only one detected unique peptide that must be validated using another proteomic technology such as targeted proteomic experiments (MRM or PRM).

A functional analysis of the detected missing proteins was performed using DAVID 6.8[42] and Ingenuity Pathways Analysis (QIAGEN IPA Spring 2017). Ingenuity functional categories with $p < 0.05$ were considered enriched. In the case of DAVID, analysis of GO terms, INTERPRO domains, KEGG pathways, and UNIGENE quantile expression level gene sets was performed with default parameters and, although the *p*-value was corrected using multiple hypothesis methods (including FDR), the selection of enriched categories was based on a criterion of EASE score (modified Fisher exact *p*-value) < 0.1. In addition, we complement the biological information about these proteins using the Protein MissingPedia[43] and GeneCards.[44]

Validation Experiment with Heavy Peptides

Heavy peptides for the 10 unique peptides detected for the five missing proteins identified, labeled with either $^{13}C_6$ $^{14}N_4$−Arg or $^{13}C_6$ $^{14}N_2$-Lys, were synthesized on an automated Multipep peptide synthesizer (Intavis) by standard F-moc chemistry. After synthesis, they were pooled together and desalted with a C18 OMIX tip (Agilent), speed-vac dried and redissolved in 0.5% formic acid, 20% acetonitrile. This peptide mixture was directly infused at a flow rate of 0.5 μL/min into a 5600 Triple-TOF mass spectrometer (Sciex) through a nanospray III ionsource (Sciex) equipped with a fused silica PicoTip emitter (10 μm × 12 cm, New Objective). MS/MS spectra of each precursor ion were acquired for 0.25 to 1 min with accumulation times of 100 to 500 ms.

We compared the fragmentation spectra of the endogenous peptides obtained in NCI60 data set with the corresponding synthetic peptide spectra. The Supporting Information includes all the annotated spectra and the obtained spectral dot product (SDP) scores[45] as a measure of spectral matching. The method used for the calculation of SDP scores is also described in the Supporting Information.

It is important to highlight that in this validation experiment we used a 5600 Triple-TOF mass spectrometer instead of an Orbitrap that was the instrument used in the shotgun experiments. This fact complicated the comparison of the endogenous and synthetic spectra.

Analysis of Transcriptomic Data Set

The BAM files corresponding to the cell lines available in both the NCI60 data set and the CCLE project were downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov). The reference genome used for the alignment of the reads was hg19. The annotation of the transcript structures of the human transcriptome considered in this study was derived from MiTranscriptome.[46] This assembly, based on 7256 RNA-Seq experiments from human normal tissues and cancer samples, contains 384 066 predicted transcripts, 165 020 of them corresponding to protein coding genes of Gencode version 19. The *ab initio* transcriptome assembly was performed using Cufflinks.[47] The quantification of these transcripts for each RNA-Seq experiment to obtain the matrix of expression levels of the 43 cell lines was performed using the software featureCounts.[48] Finally, a global normalization method using the mean size of the libraries was applied to make the samples comparable.

A multiomic bioinformatic analysis was used to highlight the samples in which the probability of detection of missing proteins was higher. For this purpose, we used the expression profiles of all the gene structures in the 43 cell lines of the NCI60 for which we had RNA-Seq experiments in the CCLE project. We considered a gene to be expressed when at least one of its corresponding transcripts was expressed. The difference between expressed and highly expressed genes was defined based on the histogram of the normalized counts for all the gene structures in all the cell lines: a gene was considered expressed in a cell line when its expression value was greater than the first quartile (Q1) or highly expressed when its

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

expression exceeded the third quartile (Q3). Using these thresholds as reference, it was possible to identify which of the analyzed samples had an over representation of missing proteins at transcript level. These cell lines would be considered as good candidates for validation of missing proteins, especially those ones that expressed a higher number of the one-hit wonders detected in the shotgun experiments (see the Supporting Information for more details and R code).

Study of Peptide Detectability Using a Machine Learning Approach

In the peptide detectability study, all the tryptic peptides of the human proteome and their detection frequency in proteomic experiments were the input data. Tryptic peptides were obtained from neXtProt database using Proteogest software,[41] and detection frequencies for each peptide were downloaded from GPMDB database (http://peptides.thegpm.org/~/peptides_by_species/). The total number of observations for each peptide was defined considering all the observations independently of the parent ion charge. Then more than 550 physicochemical and biochemical properties were calculated for each tryptic peptide using *seqinr* R package. These properties were: peptide length, peptide molecular weight, theoretical isoelectric point, percentage of different classes of amino acids (tiny, small, aliphatic, aromatic, nonpolar, polar, charged, positive or negative amino acids), and the mean value of the characteristics stored in the AAindex database (release 9.1).[49]

We sorted tryptic peptides based on the number of observations in proteomic experiments and compared the properties of the most observed peptides with the less observed ones. We randomly sampled 5000 peptides from the 50 000 most observed peptides and 5000 peptides from the 50 000 less observed peptides 500 times. In this way, we performed 500 *t* tests for each feature, and we corrected the obtained *p*-values using FDR. There were 302 properties with FDR < 0.05 in the 500 tests, but some of them were redundant. For each group of correlated properties described by the AAindex database, we chose the feature with the best mean FDR.

A final selection of 106 nonredundant properties was used to distinguished between the most and the less observed peptides in GPMDB database. For this purpose, the 100 000 tryptic peptides used for the selection of the differential peptide properties were divided in a set of training peptides (75% of the peptides) and a testing set (the remaining 25%). Different classification methods were trained and their performance was evaluated using Receiver Operator Characteristics Curve (ROC) analysis. Some methods included built-in feature selection, such as RPART, C5, JRIP, Random Forest (RF), and PART, while others do not (Partial Least Squares (PLS), Generalized Linear Model (GLM), Naïve Bayes (NB), Neural Network (NNET), and Support Vector Machine (SVM.R)). This machine learning approach was performed with *caret* R package,[50] and the RF classifier resulted to be the best option for the prediction of detectable peptides (see the Supporting Information for more details and R code).

RESULTS AND DISCUSSION

NCI60 Proteomic Experiments

We analyzed the shotgun and deep proteome experiments of the 59 cell lines from the NCI60.[36] In this study, there were 61 shotgun experiments with 12 fractions for each experiment and 9 deep proteomes with 24 fractions each. Overall, we obtained 948 raw files, which were converted to 948 MGF files prior to their analysis.

A previous study that compared the results of most of the search engines we have used in our analysis (X!Tandem, OMSSA, and Mascot)[28] showed that the decoy database approach for FDR filtering resulted in a similar number of identified peptides by each search engine. They did not find a great difference between the performances of Mascot and X! Tandem search engines, but each search engine gives a number of unique identifications due to the difference in the underlying search algorithms. The different identifications between search engines become especially important when we are analyzing low quality mass spectra (high signal-to-noise ratio, lower dissociation efficiency, etc.), as in the case of missing proteins. This is the reason why we decided to use the union of the results with all the search engines as integrative approach.

The total number of spectra in the complete data set was 14 275 503 and the percentage of them assigned to a peptide for each search engine was 24.21% by Comet, 27.38% by Mascot, 26.94% by OMSSA, and 25.58% by X!Tandem. The summary of the number of peptides, unique peptides, and proteins detected by each search engine can be seen in Tables 2, 3, 4, and 5. Numbers of proteins following C-HPP guidelines per sample are represented in Supplementary Figures 1−4, one for each search engine (Comet, Mascot, OMSSA, and X! Tandem, respectively). In the Supporting Tables 1, 2, 3, and 4, the PSM, peptide, and protein FDR, and the number of estimated false positives per sample are calculated, one table for each search engine. Interestingly, summarizing the PSMs

Table 2. Summary of Results Obtained from Analysis of NCI60 Proteomic Dataset Using Comet Search Engine by Tissue Type

cancer type	PSMs	peptides	unique peptides	proteins (≥1 unique peptides)	proteins (≥2 unique peptides)
BREAST	390 486	48 683	46 290	5502	4481
CNS	373 302	46 323	43 979	5147	4073
COLON	388 323	46 843	44 501	5062	4104
MELAN	506 468	53 165	50 622	5780	4721
NSCLC	459 102	43 488	41 244	4981	3865
PROSTATE	181 427	42 809	40 641	4935	4045
RENAL	470 663	52 973	50 346	5281	4327
LEUK	341 972	42 950	40 667	5035	4041
OVAR	343 730	44 362	42 051	5047	4126

Table 3. Summary of Results Obtained from Analysis of NCI60 Proteomic Dataset Using Mascot Search Engine by Tissue Type

cancer type	PSMs	peptides	unique peptides	proteins (≥1 unique peptides)	proteins (≥2 unique peptides)
BREAST	428 280	50 283	47 748	5726	4622
CNS	432 001	49 635	47 075	5421	4274
COLON	439 008	49 692	47 135	5317	4271
MELAN	556 346	54 707	52 022	5949	4822
NSCLC	532 187	47 233	44 739	5444	4174
PROSTATE	198 401	45 042	42 687	5231	4199
RENAL	538 219	55 014	52 163	5568	4528
LEUK	391 846	45 187	42 738	5335	4259
OVAR	392 124	46 253	43 824	5338	4272

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

Table 4. Summary of Results Obtained from Analysis of NCI60 Proteomic Dataset Using OMSSA Search Engine by Tissue Type

cancer type	PSMs	peptides	unique peptides	proteins (≥1 unique peptides)	proteins (≥2 unique peptides)
BREAST	429 479	48 173	45 701	5645	4486
CNS	432 684	47 294	44 816	5293	4049
COLON	418 875	48 430	45 943	5221	4151
MELAN	571 977	53 223	50 578	5875	4632
NSCLC	533 040	47 572	45 060	5389	4152
PROSTATE	186 361	41 214	38 988	5074	4008
RENAL	533 934	52 833	50 096	5440	4360
LEUK	393 310	43 128	40 733	5272	4118
OVAR	345 953	43 449	41 166	5243	4127

Table 5. Summary of Results Obtained from Analysis of NCI60 Proteomic Dataset Using X!Tandem Search Engine by Tissue Type

cancer type	PSMs	peptides	unique peptides	proteins (≥1 unique peptides)	proteins (≥2 unique peptides)
BREAST	421 176	51 135	48 623	5541	4634
CNS	393 296	48 889	46 391	5219	4225
COLON	408 798	49 316	46 846	5077	4227
MELAN	541 641	56 136	53 433	5855	4859
NSCLC	484 025	45 109	42 764	5017	3990
PROSTATE	193 642	45 429	43 118	5104	4180
RENAL	481 916	55 880	53 115	5480	4530
LEUK	355 932	43 046	40 721	4940	4083
OVAR	371 722	47 155	44 708	5127	4258

obtained across different tissues with each one of the search engines results in a very similar total number of PSMs (Figure 2).

When Comet search engine was used, the number of total PSMs obtained was 3 455 473, representing a mean number of 58 567 PSMs per cell line. The mean number of peptides identified per sample was 13 605 and the total number of unique peptides 94 135. Following the C-HPP guidelines for each experiment, a total number of 6867 proteins were identified.

In the analysis with Mascot, 3 908 412 PSMs were assigned in all the cell lines, and the mean number of PSMs per cell line was 66 244. The mean number of peptides per sample was 16 726, and the number of unique peptides identified considering all the cell lines under study was 93 903. This number of unique peptides allowed the identification of 6999 proteins following the C-HPP guidelines.

The number of PSMs and the mean number of PSMs per cell line were 3 845 613 and 65 180 when OMSSA was used as search engine. The mean number of peptides detected per sample was 16 746, and the total number of unique peptides was 91 215. At protein level, using the C-HPP guidelines, we identified 6822 proteins.

The last search engine used in this study was X!Tandem, which obtained 3 652 148 total PSMs with a mean number of 61 901 PSMs per cell line. In terms of peptides, we identified a mean number of 13 731 peptides per sample. The total number of unique peptides using the results of all the samples was 98 056, which achieved the identification of 7041 proteins using the C-HPP guidelines.

Finally, Table 6 summarizes all the results obtained with the different search engines per tissue type. Peptides and proteins were considered identified if at least one of the search engines detected them. Using this approach, we found 107 237 unique peptides, and 7452 proteins were identified following the C-HPP guidelines.

Table 6. Summary of Results Obtained from Analysis of NCI60 Proteomic Dataset Using the Four Search Engines

cancer type	PSMs	peptides	unique peptides	proteins (≥1 unique peptides)	proteins (≥2 unique peptides)
BREAST	1 669 421	60 488	57 565	6416	5002
CNS	1 631 283	59 113	56 188	6075	4604
COLON	1 655 004	59 526	56 603	5975	4636
MELAN	2 176 432	65 528	62 458	6703	5174
NSCLC	2 008 354	56 601	53 736	6104	4515
PROSTATE	759 831	53 176	50 502	5761	4541
RENAL	2 024 732	65 726	62 505	6259	4911
LEUK	1 483 060	54 334	51 520	5968	4614
OVAR	1 453 529	54 887	52 105	5950	4603

The numbers of unique peptides and proteins detected for each cell line with any of the search engines are shown in Figure 3. In the figure, we distinguished between the deep proteome experiments and the shotgun experiments. As expected, the number of detections at peptide and protein level is higher in the first ones. The value of the mean number

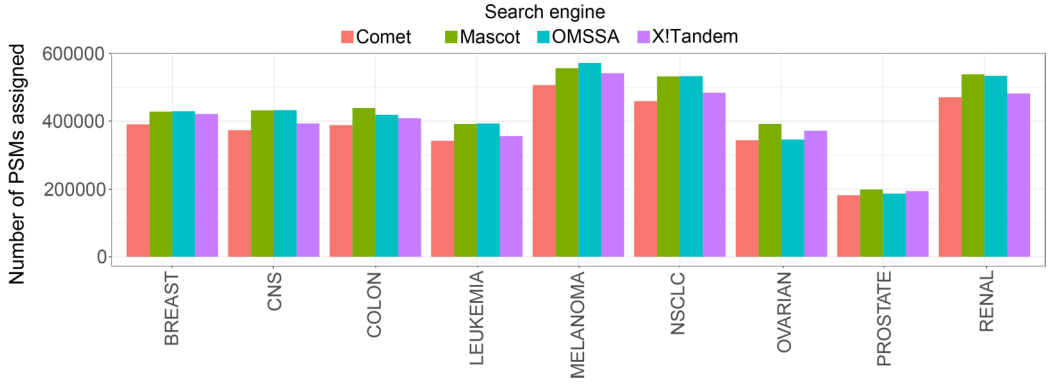


Figure 2. Number of total PSMs obtained from the analysis of the NCI60 proteomic data set summarizing the results per search engine used and tissue of origin of the cell lines.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

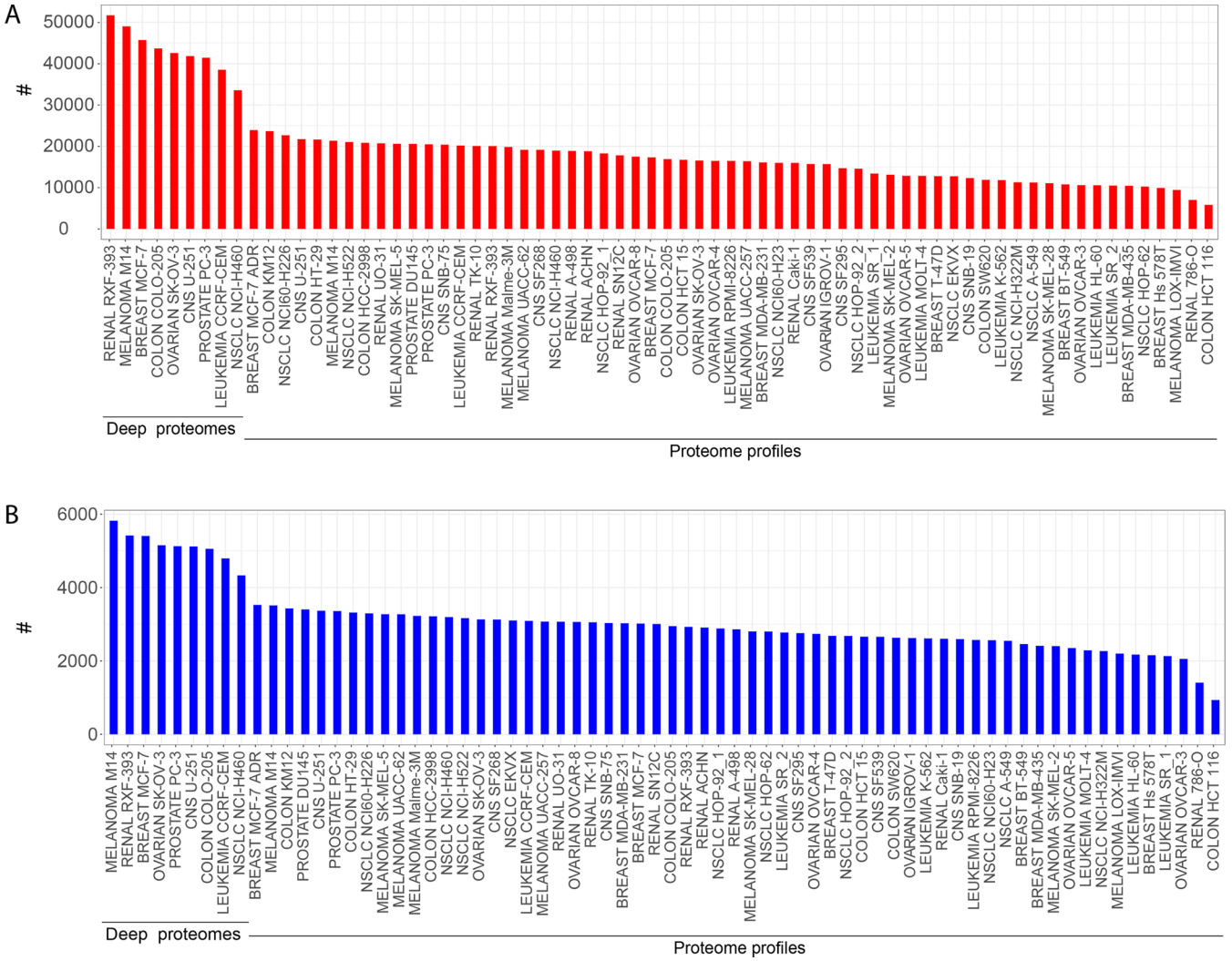


Figure 3. (A) Number of unique peptides detected with any of the four search engines. (B) Number of proteins detected following the C-HPP guidelines. For each cell line and experiment type (deep proteome or proteome profile), all the results obtained with the four search engines are represented.

of unique peptides (42 166 for deep proteomes and 15 151 for proteome profiles) and proteins detected (4206 for deep proteomes and 1985 for proteome profiles) can be used as a measure of the improvement obtained in the coverage of the proteome using a deeper MS experiment.

The graphical representation of peptides and proteins detected across chromosomes is given in Figure 4A and B. We did not find any significant bias in the number of detections toward a specific search engine. However, although the numbers are similar, differences in the assignment of peptides generate distinct sets of identified proteins that allow to increase the proteome coverage for a certain sample compared to the coverage obtained using only one search engine (Figure 4C,D).

The conclusion of our analysis was that it was possible to increase the number of proteins detected in shotgun experiments using different search engines to perform the analysis. From a total of 111 848 unique peptides detected, 75 580 (67.57%) peptides were assigned by the four search engines considered. Moreover, 10.50% of the peptides (around 3000 peptides with each search engine) were detected only by one of them. This effect could also be seen regarding the number of proteins detected following the C-HPP guidelines. We identified a total of 7452 proteins, 6351 (85.23%) independently of the search engine, while 321 proteins (4.31%) were only detected by one of them. In addition, 780 proteins were identified by two or three different search engines, which contributed to increase also the confidence of these results.

Identification of Missing Proteins

The possibility of detecting missing proteins is one of the main reasons it is relevant to improve the proteome coverage obtained in the analysis of a shotgun experiment in the framework of the goals of the C-HPP project. In the case of missing protein detection, this point of the analysis could be critical. Considering the low number of unique peptides usually detected for the missing proteins, and the fact that these identifications have to be validated using synthetic peptides or with SRM verification, we decided to use all the results obtained with the different search engines. Our objective is an increase of the sensitivity, although we are aware of a consequent decrease in the specificity. In fact, in a previous study,[28] they compared the results of the search engines used in our analysis (X!Tandem, OMSSA, and Mascot), and the observed false positive identifications were unique for each search engine, so the intersection among the results obtained

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

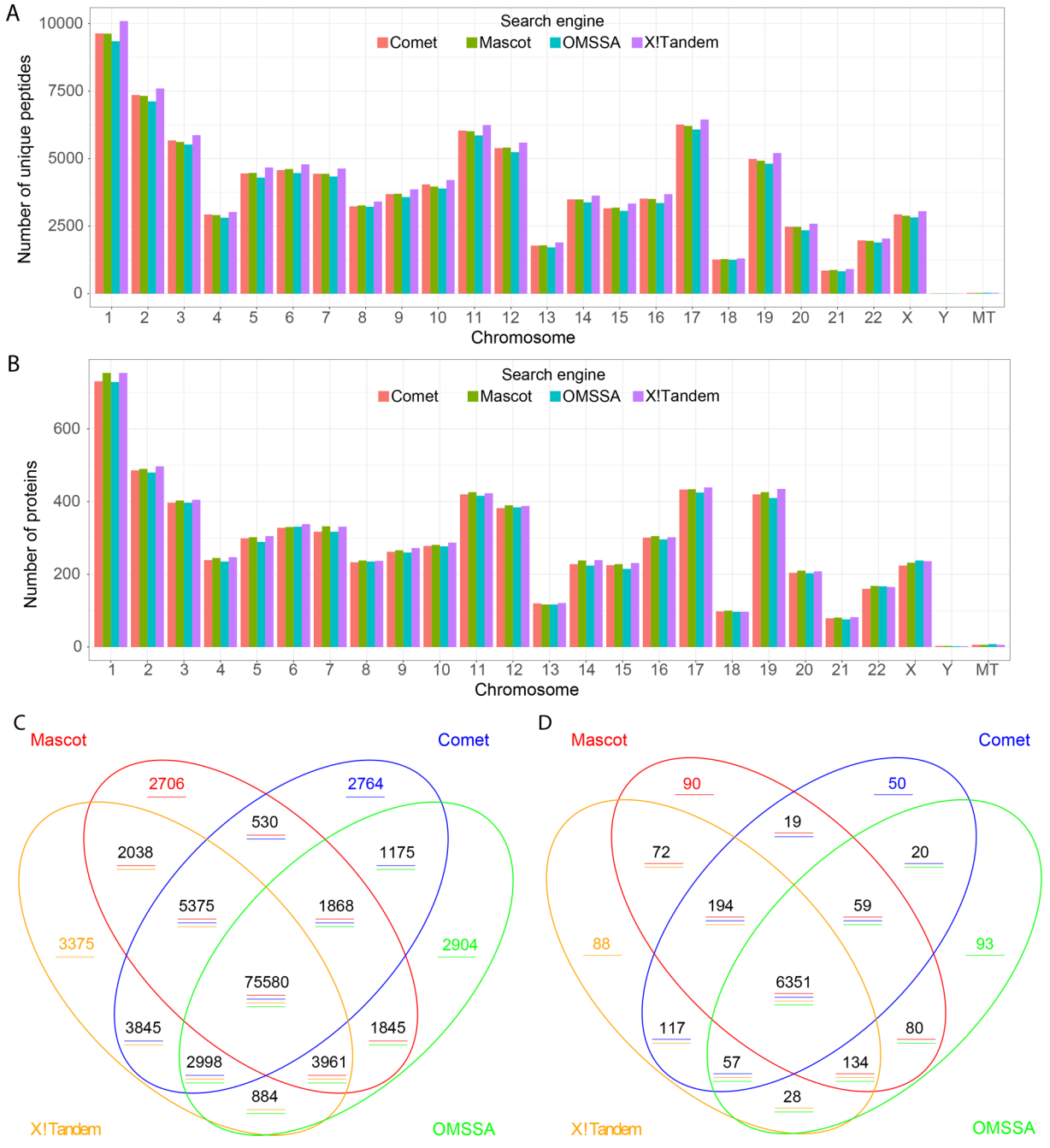


Figure 4. (A) Number of unique peptides detected across chromosomes considering all the experiments analyzed and obtained for each of the four search engines used in the study. (B) Number of proteins identified using the C-HPP guidelines across chromosomes obtained for each of the four search engines used in the study. (C) Venn diagram representation of the unique peptides found per search engine considering all the experiments analyzed. (D) Venn diagram representation of the proteins found per search engine (C-HPP guidelines).

were statistically more confident at the price of a loss of sensitivity. However, the union of the results of the search algorithms applied was able to increase the number of identifications. The differences between search engine performances become especially important when we are analyzing low quality mass spectra (high signal-to-noise ratio, lower dissociation efficiency, etc.), as in the case of missing proteins.

In the selected case study, we identified five missing proteins in five different cell lines with two unique peptides each (Supporting Table 5). The cell lines where the missing proteins were found were the following: MCF7−7 (breast tumor), SF268 (CNS tumor), COLO-205 (colon tumor), CCRF-CEM (Leukemia), and NCI60-H23 (NSCLC). If we applied a less restrictive criteria and we considered those proteins with PSM < 1%, peptide FDR < 1%, and protein FDR < 1% but with one unique peptide detected (one-hit wonders), we included 165 missing proteins identified in 58 cell lines (Supporting Table 5). One-hit wonder missing proteins are represented per sample in Supplementary Figures 5−8, one for each search engine. The unicity of all the peptides from the missing proteins was verified using the peptide uniqueness checker of neXtProt. The distribution of unique peptides and detected proteins across chromosomes is represented in Figure 5.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

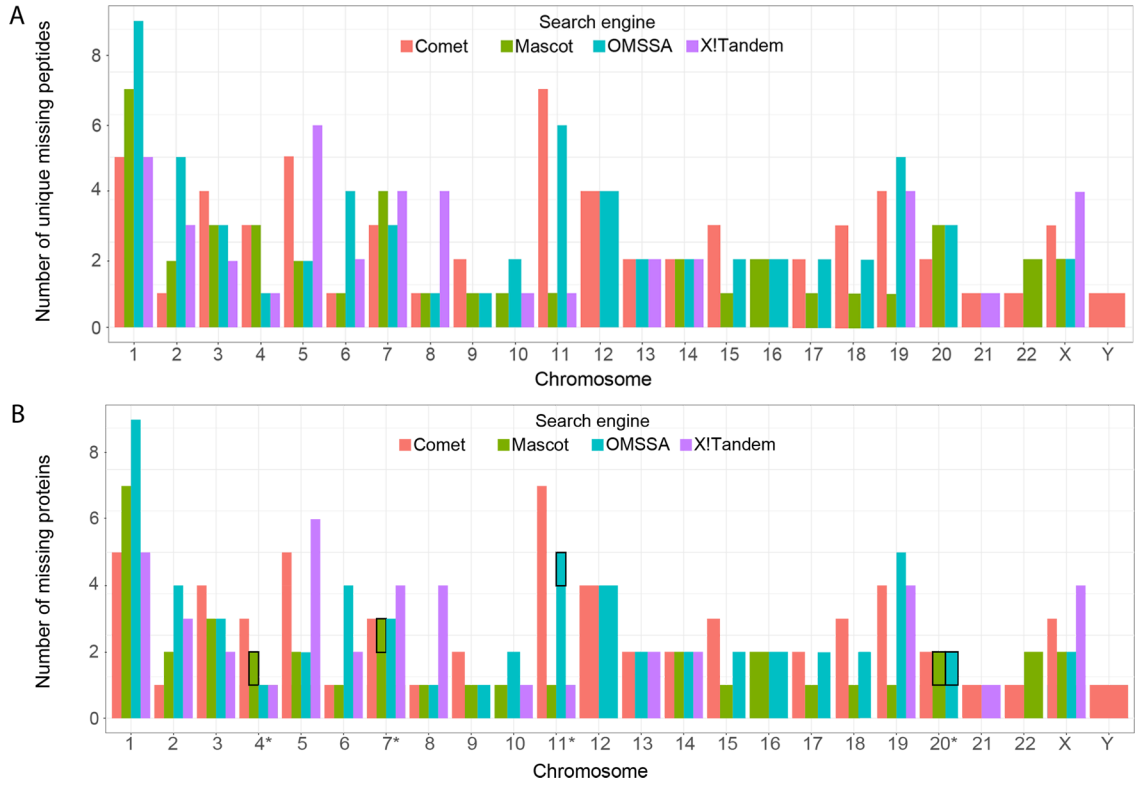


Figure 5. (A) Number of unique peptides associated with missing proteins per chromosome and search engine. (B) Number of missing proteins identified with at least one unique peptide. Highlighted (in black) proteins were identified with two unique peptides, following the C-HPP guidelines.

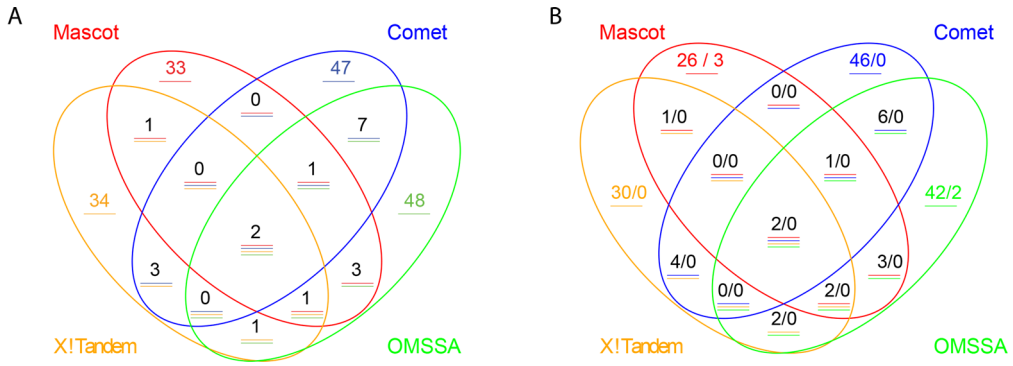


Figure 6. (A) Number of unique peptides associated with missing proteins separated per search engine. (B) Number of missing proteins identified with one (left) and two (right) unique peptides.

These results were achieved with the assignment of 866 spectra from 69 samples (all but one) using four search engines (Comet, Mascot, OMSSA, and X!Tandem), but only two peptides were detected by the four search engines. Most of the peptides were detected only in one of the searches: 47 with Comet, 33 with Mascot, 48 with OMSSA, and 34 with X! Tandem. These results highlighted the importance of the integration of different results in the quest for the missing proteins (Figure 6).

The number of missing proteins that fulfill the C-HPP criteria was five, and they were identified with more than one unique peptides (FREM3(chr 4), LAMB4 (chr 7), MYEOV (chr 11), RAD21L1 (chr 20), and TLDC2 (chr 20)). The peptides for the identified missing proteins were the following: IFITDVDNK and LVDAVGAPLPR (FREM3), LNEEADGAQKLLVK and LAGDTEAKIR (LAMB4), VAGSWLTVVTVEALGGWR and GVSFLTFHLHQSVPLGDR (MYEOV), IWLAAHWEKK and MLFTKCFLSSGFK (RAD21L1), GGSSPCPTFNNEVLAR and DGFSLQSLYR (TLDC2). We performed a validation experiment for these peptides (Supporting Information), and although we obtained good SDP Score values in many of them, the results were dubious. The mass analyzer of the validation experiments was different from the instrument used in the shotgun experiments and this could be one of the causes of the differences between endogenous and synthetic spectra. Consequently, further experiments should be performed to validate this missing proteins.

On the other hand, 165 missing proteins are detected with one unique peptide by at least one of the search engines (46 proteins by COMET, 26 proteins by Mascot, 42 proteins by OMSSA, and 30 proteins by X!Tandem). In addition, 19 one-

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

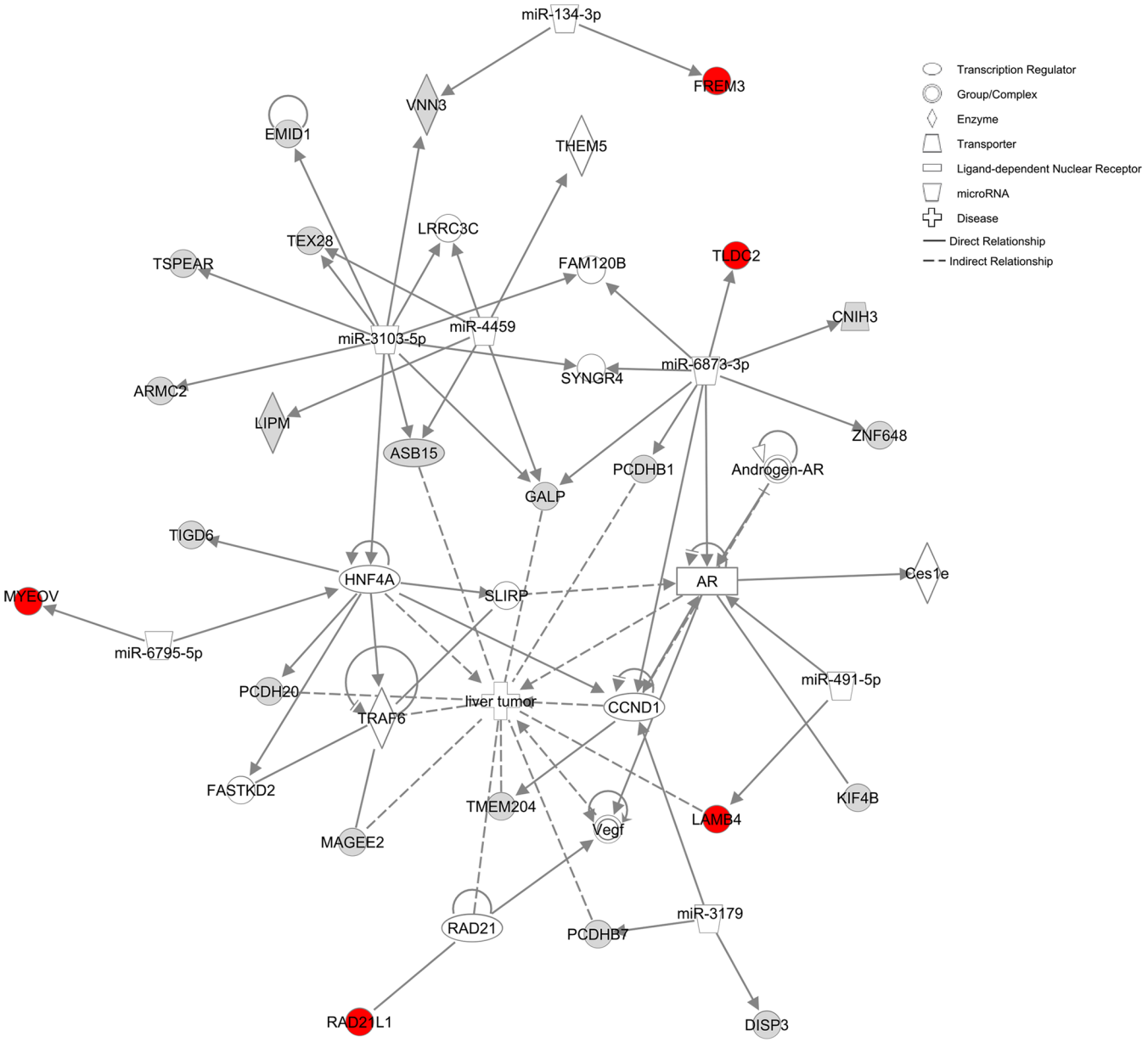


Figure 7. Interaction network of the detected missing proteins with the best score in IPA.

hit wonders were detected with two or three search engines, and all the search engines detected the same two peptides. We cannot assign MS evidence for these proteins using the C-HPP criteria, but they are good candidates for further validation using targeted proteomic experiments. The selection of the proper sample in which design these experiments can be guided by the combination of the proteomic and transcriptomic experiments performed in this study.

Additionally, the information on resources such as PeptideAtlas[51] and SRM Atlas[52] could help prioritizing missing proteins for further validation. From the 165 one-hit wonder proteins identified in our analysis, we have found other additional unique peptides previously reported in PeptideAtlas for 44 of them and SRM Atlas provides natural or synthetic SRM transitions for 159 of the proteins.

The functional analysis results of DAVID for the 170 missing proteins were in line with previous characterizations of the missing proteins.[10,12] As previously described, enriched GO categories (Supporting Table 6) included G-protein coupled receptors (15 proteins), regulation of transcription (17 proteins), olfactory receptor activity (11 proteins), and integral components of membrane (60 proteins). Over-represented Interpro domains were zinc fingers (14 proteins) and G-protein coupled receptors (15 proteins) among others, while enriched KEGG pathways were related to neuroactive ligand−receptor interaction (5 proteins) and olfactory transduction (9 proteins). The latter may result from the fact that we have detected 11 olfactory receptors (OR10J4, OR8G5, OR9K2, OR4C13, OR5M3, OR6N2, OR51F2, OR51H1, OR2 V1, OR51E1, OR2A14) of the total of 165 one-hit wonders. As we have mentioned before, this set of proteins have to be considered for further validation and some of them are expected to be false positives. According to the biological origin of the cell lines from the NCI60 data set, in which none of them are derived from nasal tissue, olfactory receptors need some additional support information for being considered a candidate. To

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

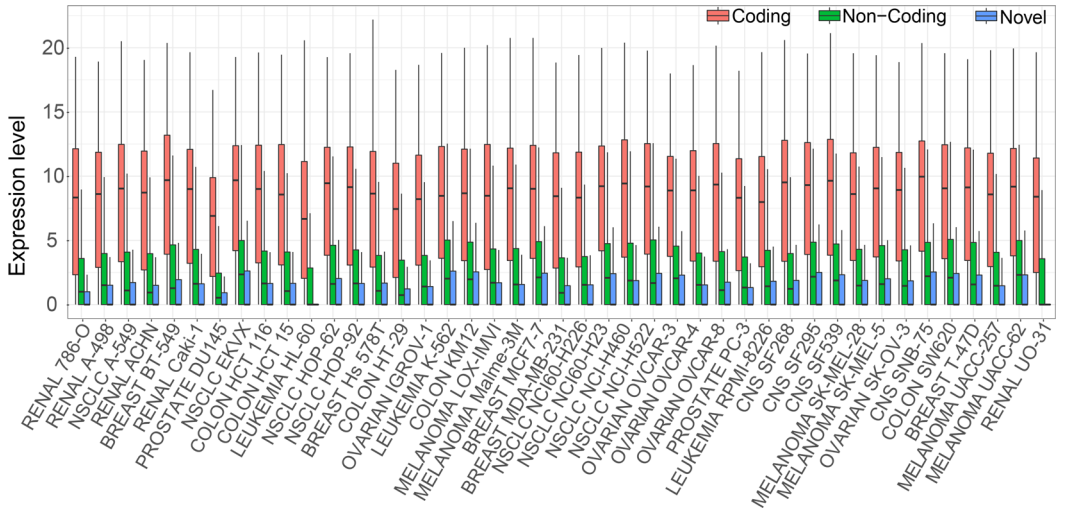


Figure 8. Transcript expression level distributions of protein coding, noncoding, and novel gene categories were compared in each of the 43 cell lines of the CCLE initiative.

determine which ones are more likely to be true positives we have consulted the information available in the Missing ProteinPedia (http://www.missingproteins.org/)[43] and GeneCards[44] about these proteins. Except for OR10J4 and OR8G5, their localization in plasma membrane is clear, and the expression of their transcripts has been detected in several tissues of the nervous, immune, muscle, secretory, reproductive, and internal systems. There is no information about OR8G5, while for OR10J4 protein some low expression of its transcript has been observed in testis, cortex, and thyroid in addition to an association with cardiovascular disease.[53] OR9K2 has also been associated with autism and schizophrenia,[54] OR51E1 with prostate cancer,[55] intestine carcinoma,[56] and lung carcinoids,[57] and OR2A14 with high-altitude pulmonary edema or HAPE.[58]

The tissue specific expression analysis using the "UNIGENE EST QUARTILE" categories highlighted brain (54 proteins), testis (61 proteins), and tissues related to embryo development (58 proteins), confirming the sample specificity of the detected missing proteins and previously published predictions.[10,12]

In the IPA functional and pathway analysis, we found a lack of enrichment of molecular functions or canonical pathways. This is to be expected since IPA is based on a curated database and the missing proteins are proteins without experimental evidence, which, in most of the cases, is linked to scarce bibliographic information about them or their coding genes. However, 167 of the missing proteins had some functional annotation and 137 of them are annotated to cancer category ($p = 1.23 \times 10^{-4}$), coherently with the samples where they have been detected (Supporting Table 7). In particular, the most enriched categories are related to melanoma (93 proteins), pancreatic tumor (39 proteins), and uterine carcinoma (38 proteins). Interestingly, the five missing proteins detected with two discriminant peptides can be related to the IPA network that has the best score. This network is enriched in proteins associated with liver tumor (Figure 7).

Even thought the current biological knowledge about the proteins for which we found MS evidence is not abundant, some relevant information was found. FREM3 (FRAS1 Related Extracellular Matrix 3) is a extracellular matrix protein, which may play a role in cell adhesion, and it has been associated with Fraser Syndrome and Glucosephosphate Dehydrogenase Deficiency diseases.[59] MYEOV (Myeloma Overexpressed) was found deregulated in a subset of t(11;14) positive multiple myelomas,[60] and LAMB4 (Laminin Subunit Beta 4) is a extracellular matrix protein that is involved in different pathways in cancer, and it is also involved in migration and organization of cells into tissues during embryonic development.[61] According to Missing ProteinPedia[43] and GeneCards,[44] LAMB4, which was seen in the NCI60 cell line SF268 (CNS), was previously detected in cerebrospinal fluid and MYEOV, which was detected in MCF7 (breast), has a corresponding transcript expressed in breast. RAD21L1 (AD21 Cohesin Complex Component Like 1) is a meiosis-specific component of a cohesin complex required during the initial steps of prophase I in male meiosis, and its activity is related to synaptonemal complex assembly, synapsis initiation, and crossover recombination between homologous chromosomes during prophase I.[62] Finally, no information is available for TLDC2 (TBC/LysM-Associated Domain Containing 2).

CCLE RNA-Seq Experiments and Enrichment of Missing Proteins

We quantified the transcript structures of the MiTranscritome human assembly[46] using the RNA-Seq experiments corresponding to the cell lines found in both the NCI60 data set and the CCLE project (43 samples). These structures were compared with GENCODE version 19, which resulted in the annotation of 17 136 protein coding genes, 12 986 noncoding genes, and 15 129 novel structures. The expression level distributions of these biotypes (Figure 8) showed statistically significant differences among them ($p < 0.01$), which confirmed the higher expression at transcript level of the protein coding genes.

To select the thresholds to distinguish between nonexpressed, expressed, and highly expressed genes we used the quartiles (Q1 and Q3, respectively) of the expression level distribution corresponding to all the gene structures (Figure 9A). Applying this criterion to each one of the 43 cell lines with RNA-Seq experiments, we determined the number of expressed and highly expressed genes in each sample and the number of proteins identified in the corresponding proteomic experiments (Figure 9B). We combined the proteomic and transcriptomic results at gene level, and we have not considered which of the transcripts of each gene is being expressed. In this way, we used the term gene as a generalization of all the possible structures

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

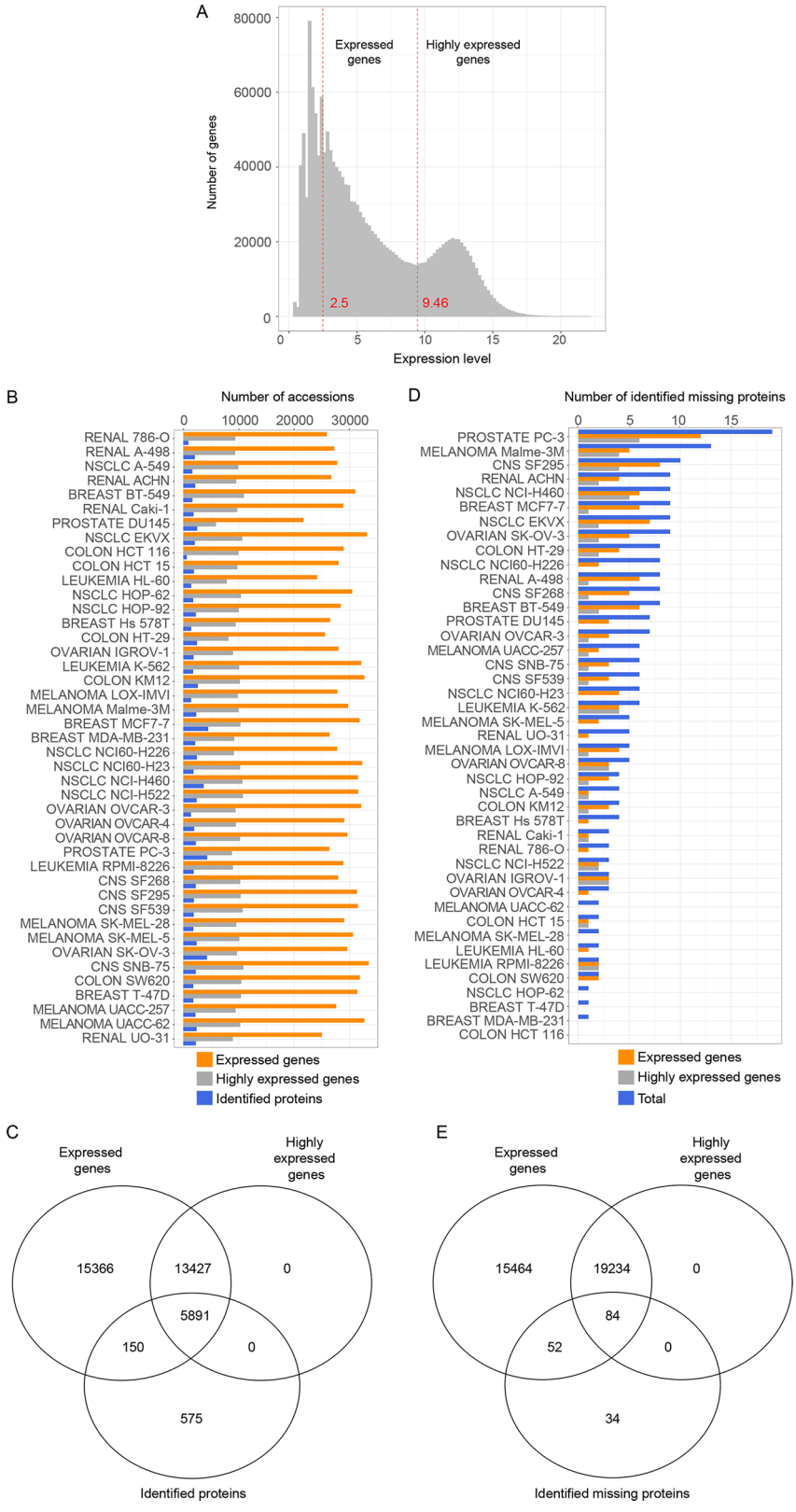


Figure 9. (A) Expression level distribution of all the genes structures in the 43 cell lines analyzed is shown and both quartiles Q1 and Q3 are marked in red. (B) Number of genes expressed or highly expressed for each cell line and proteins identified in the corresponding proteomic experiments for the same cell lines are represented (Number of MiTranscritome accessions are shown for transcriptomics and number of neXtProt accessions for proteomics). (C) Venn diagram with the intersections between expressed genes, highly expressed genes, and detected proteins in the set of 43 cell lines. (D) Number of missing proteins detected in each cell line and how many of their corresponding genes are expressed or highly expressed in the same cell lines. (E) Venn diagram with the intersections between expressed genes, highly expressed genes, and identified missing proteins in the set of 43 cell lines.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

that can be expressed from the transcription of a DNA region. The mean number of expressed genes was 29 170 genes, and the mean number of highly expressed genes was 9727, while the mean number of detected proteins using the C-HPP guidelines in these samples was 2155 proteins. As it is well-known, the number of proteins that we are able to identify in MS experiments is limited. Accordingly, the number of expressed genes in the analyzed cell lines was higher than the number of detected proteins. The intersection between the expressed genes and the detected proteins in at least one of the cell lines (Figure 9C) was 6041 proteins (29.97% of the total number of proteins in neXtProt). Interestingly, the genes of almost 90% of the identified proteins were highly expressed in some of the analyzed cell lines. In the case of missing proteins, Figure 9D summarizes the number of missing proteins that are expressed or highly expressed for each cell line of the total of detected missing proteins in the same cell lines (Supporting Table 8 and Supporting Figure 9). Although only 57% of the detected missing proteins were highly expressed (Figure 9E), this information is a valuable resource to decide in which cell lines should be performed the targeted proteomic validation experiments for the detected one-hit wonders. PC3 and NCI-H460 cell lines are good options due to the high number of one-hit wonders detected and the proportion of these missing proteins that are highly expressed (Supporting Table 8).

Study of Peptide Detectability of Peptides from Missing Proteins Using Several Search Engines

One of the possible causes for the difficulties encountered in the missing proteins detection could be a detectability problem of peptides. Under this scenario, we decided to study the physicochemical and biochemical properties of the tryptic peptides identified in shotgun experiments of the NCI60 initiative and test this hypothesis. We were able to sort the tryptic peptides based on the number of observations in proteomic experiments and compare the properties of the most observed peptides with the less observed ones. We used a final selection of 106 nonredundant properties statistically different between both groups of peptides to train a classifier. As described in the Methods section, different classification methods were applied, and Random Forest (RF) was the best option for the prediction of detectable peptides with a sensitivity of 0.746 and a specificity of 0.719 (Figure 10).

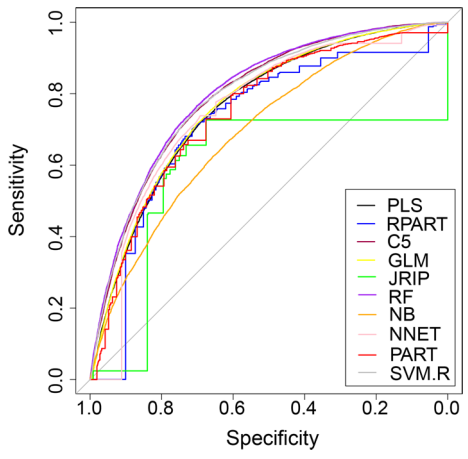


Figure 10. Performance evaluation of the peptide detectability classifiers is shown using ROC analysis with the test data set.

The developed classifier was applied to the detected tryptic peptides of the missing proteins, and 38.67% of them were predicted to be detectable, while the detectability of the nonmissing identified tryptic peptides was significantly higher (73.48%). The low predicted detectability was an expected result considering the number of missing proteins that are membrane proteins or highly insoluble proteins.[13] The tryptic peptides of the identified missing proteins that were not detected had even lower predicted detectability (32.39%) so this information could be considered for prioritizing the peptides of a missing protein for further validation.

In addition, we decided to check the peptide detectability of the results obtained with each one of the four search engines used. Detected peptides of the proteins identified using all the used search engines were peptides with a good predicted detectability (73.48%). On the other hand, the mean peptide detectability for the detected peptides of proteins identified only by one of the search engines was very similar (72.47%), and these additional peptides could make the difference in the detection of the missing proteins (Figure 11). Hence, we can consider that the results are complementary because each search engine identifies additional peptides with a predicted good detectability.

CONCLUSIONS

It is well-known that the existence of a certain type of proteins in a given biological matrix is difficult to prove using mass spectrometry or antibody-based technologies, although bioinformatic evidence of their translation is available in proteomic databases. This type of proteins includes among others, low expression proteins like transcription factors, tissue specific proteins, proteins that are expressed only under certain biological conditions or produced only in certain development stages, or proteins with particular cellular locations, as membrane proteins. However, although currently unknown, the implications of these proteins in biological processes and disease could be of major significance. On the basis of this assumption, since its start in the year 2010, the Human Proteome Project (HPP) is trying to complete the characterization of the human proteome, with a special interest in those proteins with a lack of robust experimental evidence. In the context of this project, this group of proteins is known as "missing proteins", and they are cataloged in the neXtProt database, the central knowledge-based tool of the C-HPP initiative.

Different methodological approaches to detect these proteins have been developed by the research groups involved in the project. These bioinformatic pipelines have made a tremendous contribution both to (1) the advances in the description of the human proteome and (2) the development of new data analysis methods to improve the results obtained from a proteomic shotgun experiment. Furthermore, many innovations in the field of proteogenomics have been introduced in these workflows with the aim of integrating different omics (mainly genomics, transcriptomics, and proteomics) to make a leap forward on the understanding of the complexity of the cellular and molecular machinery.

In this manuscript, we go a step further in the analysis of public proteomic data sets to take greater advantage of the potential of these experiments. We analyzed the complete NCI60 data set, which contains nine deep proteomes and 61 proteome profiles from 59 cell lines. However, we increased the proteome coverage for each sample integrating the results from

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

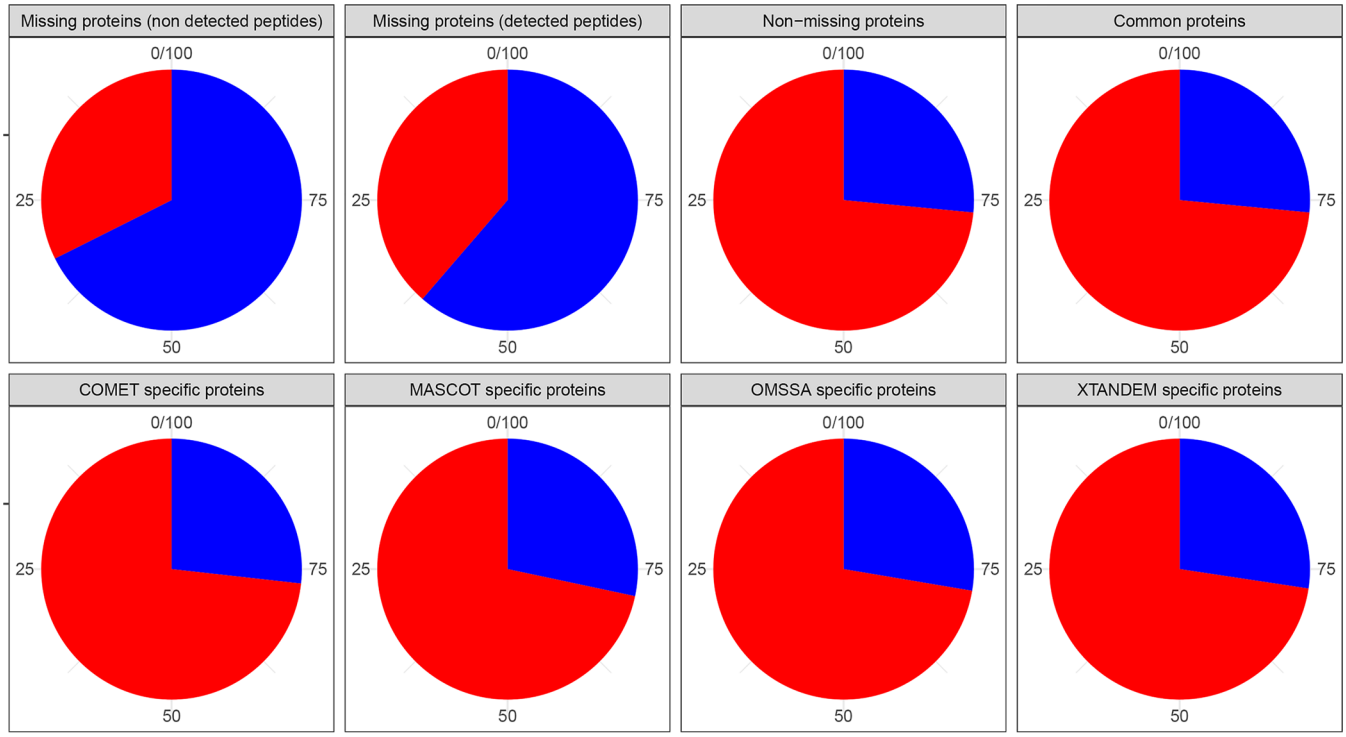


Figure 11. Percentage of predicted peptide detectability for distinct sets of peptides: nondetected peptides of the identified missing proteins, detected peptides of identified missing proteins, detected peptides of nonmissing identified proteins, detected peptides of the proteins identified by the four search engines used (Common proteins), and detected peptides of the proteins identified by only one of the search engines (Comet, Mascot, OMSSA, and X!Tandem specific proteins). In red, predicted to be detectable peptides and in blue, peptides predicted to be not detectable.

four different search engines (Comet, Mascot, OMSSA, and X! Tandem). More than 3600 searches were performed, and the detected peptides were intersected with the unique peptides of neXtProt database to be aligned with the C-HPP guidelines. According to the data obtained, we can safely assume that the search engines are complementary, and their integration is an appropriate method to increase the performance of the analysis of shotgun experiments. The statistical threshold of 1% at PSM, peptide, and protein level were applied. As a result, we found MS evidence for five missing proteins (FREM3 (Chr 4), LAMB4 (Chr 7), MYEOV (Chr 11), RAD21L1 (Chr 20), and TLDC2 (Chr 20)), identified with more than one unique peptide, and we also found 165 missing protein candidates detected with only one unique peptide (one-hit wonders). We performed validation experiments using heavy peptides and a SDP score approach to compare the fragmentation spectra of the endogenous and the synthesized peptides for the five missing proteins identified with two unique peptides. However, we obtained questionable results, and these peptides cannot be considered validated. Further experiments are required.

A machine-learning approach allowed us to study peptide detectability, and we can conclude that unique tryptic peptides of the identified missing proteins have a low predicted detectability. Besides this, additional peptides detected only by one of the used search engines have as high predicted detectability as the peptides detected by all the search engines. In this way, we confirmed the complementarity and quality of the detection results obtained with our integrative bioinformatic approach.

The MS evidence of the one-hit wonder proteins must be verified using an alternative proteomic technology, for example, using targeted proteomic experiments (MRM or PRM). Although we have not carried out these experiments, the integration of the proteomic results with the quantification of the protein coding genes in a subset of the NCI60 cell lines available in the CCLE project allowed us to provide guidance for the selection of the biological matrices in which these proteins are more probable to be detected. The analysis of the 165 missing proteins using DAVID and Ingenuity softwares for tissue specificity, GO, KEGG pathways, and protein domain enrichments provided insight into the biological function of these proteins and supported the ranking of cell lines for validation studies provided by the RNA-Seq experiments from the CCLE.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00388.

Number of proteins following C-HPP guidelines per sample for Comet, Mascot, OMSSA, and X!Tandem results; number of missing proteins with one unique peptide per sample for Comet, Mascot, OMSSA, and X! Tandem results; expression profiles of transcripts corresponding to identified missing proteins in CCLE experiments (PDF)

Spectra of endogenous peptides detected for five missing proteins identified following C-HPP guidelines; corresponding validations with synthetic peptides (PDF)

Supporting methods; R code (PDF)

FDR values for each sample calculated at PSM, peptide, and protein level for Comet results (XLSX)

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

FDR values for each sample calculated at PSM, peptide, and protein level for Mascot results (XLSX)
FDR values for each sample calculated at PSM, peptide, and protein level for OMSSA results (XLSX)
FDR values for each sample calculated at PSM, peptide, and protein level for X!Tandem results (XLSX)
Missing proteins identified with PSM, peptide, and protein FDR < 1% and one or more unique peptides (XLSX)
Functional analysis of identified missing proteins using DAVID (XLSX)
Functional analysis of identified missing proteins using IPA (XLS)
Identified missing proteins per cell line and their corresponding transcript expression level (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: vsegura@unav.es.

ORCID

J. Ignacio Casal: 0000-0003-1085-2840
Fernando J. Corrales: 0000-0002-0231-5159
Victor Segura: 0000-0002-7740-6290

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Laboratories of CIMA, CNB, CIB, and UPV are members of the PRBB-ISCIII platform. This work was supported by PRBB and the Carlos III National Health Institute Agreement, PRBB-ISCIII (PT13/0001/0002); Grant Nos. SAF2014-5478-R from Ministerio de Ciencia e Innovación and ISCIII-RETIC RD06/0020 to F.J.C., Grant No. BIO2015-66489-R to J.I.C., and Grant Nos. 33/2015 from Dpto. de Salud de Gobierno de Navarra and DPI2015-68982-R from Ministerio de Ciencia e Innovación to V.S. J.A.V. and Y.P.R. are supported by the Wellcome Trust [Grant No. WT101477MA].

REFERENCES

(1) Legrain, P.; et al. The human proteome project: Current state and future direction. *Mol. Cell. Proteomics* **2011**, *7*, M111−009993.
(2) Paik, Y.-K.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30*, 221−223.
(3) Paik, Y.-K.; et al. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11*, 2005−2013.
(4) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community. *J. Proteome Res.* **2013**, *12*, 23−27.
(5) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussman, M.; Qin, J.; Omenn, G. S. Highlights of B/D-HPP and HPP Resource Pillar Workshops at 12th Annual HUPO World Congress of Proteomicsf. *Proteomics* **2014**, *14*, 975−988.
(6) Van Eyk, J. E.; Corrales, F. J.; Aebersold, R.; Cerciello, F.; Deutsch, E. W.; Roncada, P.; Sanchez, J.-C.; Yamamoto, T.; Yang, P.; Zhang, H.; Omenn, G. S. Highlights of the biology and disease-driven Human Proteome Project, 2015−2016. *J. Proteome Res.* **2016**, *15*, 3979−3987.
(7) Paik, Y.-K.; Hancock, W. S. Uniting ENCODE with genome-wide proteomics. *Nat. Biotechnol.* **2012**, *30*, 1065−1067.
(8) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guével, B.; Sergeant, N.; Mitchell, V.; Pineau, C. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14*, 3606−3620.
(9) Carapito, C.; et al. Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J. Proteome Res.* **2015**, *14*, 3621−3634.
(10) Guruceaga, E.; Sanchez del Pino, M. M.; Corrales, F. J.; Segura, V. Prediction of a missing protein expression map in the context of the human proteome project. *J. Proteome Res.* **2015**, *14*, 1350−1360.
(11) Park, G. W.; Hwang, H.; Kim, K. H.; Lee, J. Y.; Lee, H. K.; Park, J. Y.; Ji, E. S.; Park, S.-K. R.; Yates, J. R.; Kwon, K.-H.; Park, Y. M.; Lee, H.-J.; Paik, Y.-K.; Kim, J. Y.; Yoo, J. S. Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate. *J. Proteome Res.* **2016**, *15*, 4082−4090.
(12) Garin-Muga, A.; Odriozola, L.; Martínez-Val, A.; Del Toro, N.; Martínez, R.; Molina, M.; Cantero, L.; Rivera, R.; Garrido, N.; Dominguez, F.; Sanchez Del Pino, M. M.; Vizcaíno, J. A.; Corrales, F. J.; Segura, V. Detection of Missing Proteins Using the PRIDE Database as a Source of Mass Spectrometry Evidence. *J. Proteome Res.* **2016**, *15*, 4101−4115.
(13) Horvatovich, P.; et al. Quest for Missing Proteins: Update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14*, 3415−3431.
(14) Gaudet, P.; et al. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D177−D182.
(15) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P.-A.; Zahn-Zabal, M.; Lane, L. The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics* **2017**, *btx318*, 1−2.
(16) Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y.-K.; Weintraub, S. T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; Hancock, W. S.; Hermjakob, H.; Aebersold, R.; Moritz, R. L.; Omenn, G. S. Human proteome project mass spectrometry data interpretation guidelines 2.1. *J. Proteome Res.* **2016**, *15*, 3961−3970.
(17) Duek, P.; Bairoch, A.; Gateau, A.; Vandenbrouck, Y.; Lane, L. Missing Protein Landscape of Human Chromosomes 2 and 14: Progress and Current Status. *J. Proteome Res.* **2016**, *15*, 3971−3978.
(18) Deutsch, E. W.; Sun, Z.; Campbell, D. S.; Binz, P.-A.; Farrah, T.; Shteynberg, D.; Mendoza, L.; Omenn, G. S.; Moritz, R. L. Tiered human integrated sequence search databases for shotgun proteomics. *J. Proteome Res.* **2016**, *15*, 4091−4100.
(19) Wei, W.; Luo, W.; Wu, F.; Peng, X.; Zhang, Y.; Zhang, M.; Zhao, Y.; Su, N.; Qi, Y.; Chen, L.; Zhang, Y.; Wen, B.; He, F.; Xu, P. Deep coverage proteomics identifies more low-abundance missing proteins in human testis tissue with Q-exactive HF mass spectrometer. *J. Proteome Res.* **2016**, *15*, 3988−3997.
(20) Vandenbrouck, Y.; et al. Looking for missing proteins in the proteome of human spermatozoa: an update. *J. Proteome Res.* **2016**, *15*, 3998−4019.
(21) Segura, V.; Garin-Muga, A.; Guruceaga, E.; Corrales, F. J. Progress and pitfalls in finding the 'missing proteins' from the human proteome map. *Expert Rev. Proteomics* **2017**, *14*, 9−14.
(22) Tabas-Madrid, D.; Alves-Cruzeiro, J.; Segura, V.; Guruceaga, E.; Vialas, V.; Prieto, G.; García, C.; Corrales, F. J.; Albar, J. P.; Pascual-Montano, A. Proteogenomics Dashboard for the Human Proteome Project. *J. Proteome Res.* **2015**, *14*, 3738−3749.
(23) Vizcaíno, J. A.; Csordas, A.; Del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44*, D447−D456.
(24) Park, E. S.; et al. Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. *Mol. Cancer Ther.* **2010**, *9*, 257−267.
(25) 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*, 1599−1608.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390

(26) 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*, 1220−1229.
(27) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7*, 245−253.
(28) Audain, E.; Uszkoreit, J.; Sachsenberg, T.; Pfeuffer, J.; Liang, X.; Hermjakob, H.; Sanchez, A.; Eisenacher, M.; Reinert, K.; Tabb, D. L.; Kohlbacher, O.; Perez-Riverol, Y. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics* **2017**, *150*, 170−182.
(29) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234−1242.
(30) Barretina, J.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603−607.
(31) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.
(32) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.
(33) Eng, J. K.; Fischer, B.; Grossmann, J.; MacCoss, M. J. A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **2008**, *7*, 4598−4602.
(34) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.
(35) Schirle, M.; Heurtier, M.-A.; Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2003**, *2*, 1297−1305.
(36) Gholami, A. M.; Hahne, H.; Wu, Z.; Auer, F. J.; Meng, C.; Wilhelm, M.; Kuster, B. Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **2013**, *4*, 609−620.
(37) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.
(38) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958−964.
(39) Fenyö, 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*, 768−774.
(40) Gentleman, R. C.; et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **2004**, *5*, R80.
(41) Cagney, G.; Amiri, S.; Premawaradena, T.; Lindo, M.; Emili, A. In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci.* **2003**, *1*, 5.
(42) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4*, 44−57.
(43) Baker, M. S.; Ahn, S. B.; Mohamedali, A.; Islam, M. T.; Cantor, D.; Verhaert, P. D.; Fanayan, S.; Sharma, S.; Nice, E. C.; Connor, M.; Ranganathan, S. Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **2017**, *8*, 14271.
(44) Stelzer, G.; et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **2016**, 1−30.
(45) Ye, D.; Fu, Y.; Sun, R.-X.; Wang, H.-P.; Yuan, Z.-F.; Chi, H.; He, S.-M. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010**, *26*, i399−i406.
(46) Iyer, M. K.; et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **2015**, *47*, 199−208.
(47) Trapnell, C.; Williams, B. A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511−515.
(48) Liao, Y.; Smyth, G. K.; Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923−930.
(49) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **2007**, *36*, D202−D205.
(50) Kuhn, M. Caret package. *Journal of Statistical Software* **2008**, *28*, 1−26.
(51) 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*, 3461−3473.
(52) Kusebauch, U.; Campbell, D. S.; Deutsch, E. W.; Chu, C. S.; Spicer, D. A.; Brusniak, M.-Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; Shteynberg, D. Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell* **2016**, *166*, 766−778.
(53) Benjamin, E. J.; et al. Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med. Genet.* **2007**, *8*, S11.
(54) Goodbourn, P. T.; Bosten, J. M.; Bargary, G.; Hogg, R. E.; Lawrance-Owen, A. J.; Mollon, J. Variants in the 1q21 risk region are associated with a visual endophenotype of autism and schizophrenia. *Genes, Brain and Behavior* **2014**, *13*, 144−151.
(55) Weng, J.; Wang, J.; Hu, X.; Wang, F.; Ittmann, M.; Liu, M. PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer. *Int. J. Cancer* **2006**, *118*, 1471−1480.
(56) Cui, T.; Tsolakis, A. V.; Li, S.-C.; Cunningham, J. L.; Lind, T.; Öberg, K.; Giandomenico, V. Olfactory receptor 51E1 protein as a potential novel tissue biomarker for small intestine neuroendocrine carcinomas. *Eur. J. Endocrinol.* **2013**, *168*, 253−261.
(57) Giandomenico, V.; Cui, T.; Grimelius, L.; Öberg, K.; Pelosi, G.; Tsolakis, A. V. Olfactory receptor 51E1 as a novel target for diagnosis in somatostatin receptor-negative lung carcinoids. *J. Mol. Endocrinol.* **2013**, *51*, 277−286.
(58) Yingzhong, Y.; Yaping, W.; Jin, X.; Rili, G. The susceptibility gene screening in a Chinese high-altitude pulmonary edema family by whole-exome sequencing. *Hereditas* **2017**, *39*, 135.
(59) Smyth, I.; Du, X.; Taylor, M. S.; Justice, M. J.; Beutler, B.; Jackson, I. J. The extracellular matrix gene Frem1 is essential for the normal adhesion of the embryonic epidermis. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 13560−13565.
(60) Janssen, J. W.; Vaandrager, J.-W.; Heuser, T.; Jauch, A.; Kluin, P. M.; Geelen, E.; Bergsagel, P. L.; Kuehl, W. M.; Drexler, H. G.; Otsuki, T.; Bartram, C. R.; Schuuring, E. Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t (11; 14)(q13; q32). *Blood* **2000**, *95*, 2691−2698.
(61) Choi, M. R.; An, C. H.; Yoo, N. J.; Lee, S. H. Laminin gene LAMB4 is somatically mutated and expressionally altered in gastric and colorectal cancers. *Apmis* **2015**, *123*, 65−71.
(62) Gutierrez-Caballero, C.; Herran, Y.; Sanchez-Martin, M.; Suja, J. A.; Barbero, J. L.; Llano, E.; Pendas, A. M. Identification and molecular characterization of the mammalian α-kleisin RAD21L. *Cell Cycle* **2011**, *10*, 1477−1487.

DOI: 10.1021/acs.jproteome.7b00388
J. Proteome Res. 2017, 16, 4374−4390