Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JULY 2015

Impact Factor: 4.25 · DOI: 10.1021/acs.jproteome.5b00474 · Source: PubMed

Journal of proteome research

Article

pubs.acs.org/jpr

Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project

Paula Díez,[†,‡] Conrad Droste,[§] Rosa M. Dégano,[‡] María González-Muñoz,[†] Nieves Ibarrola,[‡] Martín Pérez-Andrés,[†] Alba Garin-Muga,[∥] Víctor Segura,[∥] Gyorgy Marko-Varga,[⊥] Joshua LaBaer,[#] Alberto Orfao,[†] Fernando J. Corrales,[∥] Javier De Las Rivas,[*,§] and Manuel Fuentes[*,†,‡]

[†]Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[‡]Proteomics Unit. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[§]Bioinformatics and Functional Genomics Research Group, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

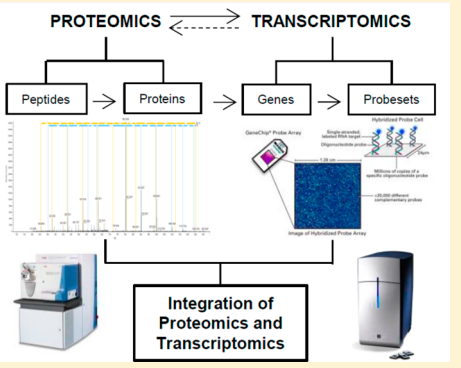
[∥]Division of Hepatology and Gene Therapy, Proteomics and Bioinformatics Unit, Centre for Applied Medical Research (CIMA), University of Navarra, 31008 Pamplona, Spain

[⊥]Clinical Protein Science and Imaging, Biomedical Centre, Department of Biomedical Engineering, Lund University, BMC D13, 221 84 Lund, Sweden

[#]Biodesign Institute, Arizona State University, 1001 South McAllister Avenue, Tempe, Arizona 85287, United States

Supporting Information

ABSTRACT: A comprehensive study of the molecular active landscape of human cells can be undertaken to integrate two different but complementary perspectives: transcriptomics, and proteomics. After the genome era, proteomics has emerged as a powerful tool to simultaneously identify and characterize the compendium of thousands of different proteins active in a cell. Thus, the Chromosome-centric Human Proteome Project (C-HPP) is promoting a full characterization of the human proteome combining high-throughput proteomics with the data derived from genome-wide expression profiling of protein-coding genes. Here we present a full proteomic profiling of a human lymphoma B-cell line (*Ramos*) performed using a nanoUPLC-LTQ-Orbitrap Velos proteomic platform, combined to an in-depth transcriptomic profiling of the same cell type. Data are available via ProteomeXchange with identifier PXD001933. Integration of the proteomic and transcriptomic data sets revealed a 94% overlap in the proteins identified by both -omics approaches. Moreover, functional enrichment analysis of the proteomic profiles showed an enrichment of several functions directly related to the biological and morphological characteristics of B-cells. In turn, about 30% of all protein-coding genes present in the whole human genome were identified as being expressed by the *Ramos* cells (stable average of 30% genes along all the chromosomes), revealing the size of the protein expression-set present in one specific human cell type. Additionally, the identification of missing proteins in our data sets has been reported, highlighting the power of the approach. Also, a comparison between neXtProt and UniProt database searches has been performed. In summary, our transcriptomic and proteomic experimental profiling provided a high coverage report of the expressed proteome from a human lymphoma B-cell type with a clear insight into the biological processes that characterized these cells. In this way, we demonstrated the usefulness of combining -omics for a comprehensive characterization of specific biological systems.



KEYWORDS: *C-HPP, lymphoma B-cell line, protein expression profile, transcriptomics, subcellular fractionation*

INTRODUCTION

Upon successful completion of the Human Genome Project in 2003—approximately 20 055 protein-coding genes have been reported according to neXtProt version of September 19, 2014—a major challenge remains as regards the understanding of how gene expression levels relate to the regulatory behavior

Special Issue: The Chromosome-Centric Human Proteome Project 2015

Received: May 28, 2015

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

of cells.[1,2] After the genomics era, the scientific community realized that DNA coding sequences are not by themselves sufficient to provide an overview of cellular biological processes.[3] The genome remains nearly constant throughout the lifetime of a cell and there are no significant changes regarding the cell type once it achieves its differentiated specific state. However, both the transcriptome and the proteome are much more dynamic and they can vary with the functional state of the cell or in response to intra- and extra-cellular environmental signals. Henceforth, studying changes in mRNA and protein levels can provide a clear and accurate readout of the cell state.[1] Proteins are usually the final regulatory and effector molecules of cells coded by genes, and proteomics allows a comprehensive and integrative study of all proteins in a cellular system.[4] The main goal of proteomics is often to generate a complete and quantitative map of proteins, including cellular localization of proteins, identification of protein complexes, protein isoforms, and post-translational protein modifications (PTM). In fact, proteomes are characterized by large protein-abundance differences, cell-type, time-dependent expression patterns, and PTMs, all of which carry biological information that is not commonly accessible by genomics or transcriptomics data.

In turn, transcriptomics methodologies are the most used to determine the active expression of predicted protein-coding genes. In this respect, RNA deep sequencing technology has emerged as a promising strategy that provides a whole transcriptome shotgun approach to quantifying detailed genome-wide expression.[5,6] Previous to the development of high-throughput RNA and DNA sequencing technologies, microarray technologies have been considered powerful large-scale methods that have been applied for gene expression profiling of multiple samples in many different biological studies. Nevertheless, the detection and quantification of expressed mRNAs do not unequivocally determine the presence of the corresponding translated proteins. Regulatory mechanisms, such as PTMs or silencing processes, can result in an imbalance between the transcribed and the translated portions, as well as the half-live differences between transcripts and proteins.[3,7−11]

Proteomics measures proteins directly, providing information about the active genes at the translational level, and can be used as verification of gene expression.[3,12] Major advances which have occurred in proteomics over the last years have allowed detection and validation of putative genes, together with the added benefit for genome annotation.[13,14] However, many challenges still remain in these approaches due to proteome complexity.[15] Among others, these are related to (i) the sample preparation procedures, because to get maximum coverage of the proteome it is necessary to use multiple sample fractionation methods, either through protein extraction protocols or processing techniques; (ii) the peptide separation optimized via the usage of combinations of different proteases (e.g., trypsin, Lys-C, proteinase K, etc.) to increase peptide recovery leading to increase protein sequence coverage; (iii) the precision and accuracy of mass spectrometry (MS) measurements, because to reduce the occurrence of false positive hits it is recommended to use instruments with high resolution, high sensitivity, fast scanning speed, and high mass accuracy (ppm) such as the Fourier Transform Ion Cyclotron Resonance (FT-ICR) and the Linear Trap Quadrupole-Orbitrap (LTQ-Orbitrap) (also affected by a different ionization efficiency of different peptides); (iv) the data processing methods and database search engines (i.e., Mascot, Sequest, OMSSA), as well as the scoring and validation parameters (false discovery rate − FDR, precursor mass tolerance), which determine the robustness of the results about the identified peptides and proteins[9,13−13].

All these challenges are still open in proteomics; therefore, the integration of transcriptomics and proteomics data, working with adequate bioinformatics strategies, can offer new insights in the field[1,16] and provide reliable information about how genes and proteins are regulated and integrated at the molecular, cellular, and organismal levels to control a set of biological responses.[17]

The Human Proteome Organization (HUPO) has coordinated the efforts of the international community promoting several initiatives to describe the human proteome through a well-planned working scheme. The project is partially organized according to a chromosome-based strategy (C-HPP) where scientific groups from different nationalities agreed to characterize the proteome of a specific chromosome.[18] All 24 chromosomes plus the mitochondrial DNA have been already assigned to many teams from 21 different countries. The vast heterogeneity, wide dynamic range, and different ionization efficiencies of peptides are causing a restriction in detection and quantification capacities of large-scale proteomics studies. Hence, C-HPP groups are now integrating transcriptomics and proteomics data sets in order to better guide the genome-wide proteomics analysis.[19] Specifically, the Spanish team of the Human Proteome Project (SpHPP) addresses the protein mapping of chromosome 16, and that it is the reason why a lymphoma cell line has been used in this research, because a great number of proteins from B-cells are encoded in chromosome 16.[2,20] The development of studies integrating proteomics and transcriptomics may lead to the full characterization of chromosomes and also to determine the relationship between the transcripts and their products (i.e., proteins). Additionally, the usage of shotgun MS approaches generates huge amounts of data providing lots of information in which it might be possible to detect missing proteins. Also, the C-HPP initiative is expected to improve the knowledge about diseases and their biology contributing to the Biology and Disease (B/D)-HPP initiative.

Here, we present a workflow integrating transcriptomics and proteomics data sets about a specific biological sample. As a model, we selected the *Ramos* human B-cell lymphoma cell line, which is well-characterized at the gene expression level. In addition, the proposed approach has reported the identification of missing proteins that correspond to certain subset of protein-coding genes well-annotated in the human genome that have not been yet detected by any proteomic MS-based experimental approach. Additionally, a further comparison between neXtProt and UniProt database searches has been accomplished. On balance, all these points may be of great interest for the scientific community and, specifically, for the C-HPP consortium.

MATERIALS AND METHODS

Reagents

Protease inhibitor cocktail, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), Tween 20, tris (2-carboxyethyl) phosphine hydrochloride (TCEP), phenylmethanesulfonyl fluoride (PMSF), digitonin, octylphenoxypolyethoxyethanol (IGEPAL), RPMI-1640 media, potassium ferrocyanide, sodium

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

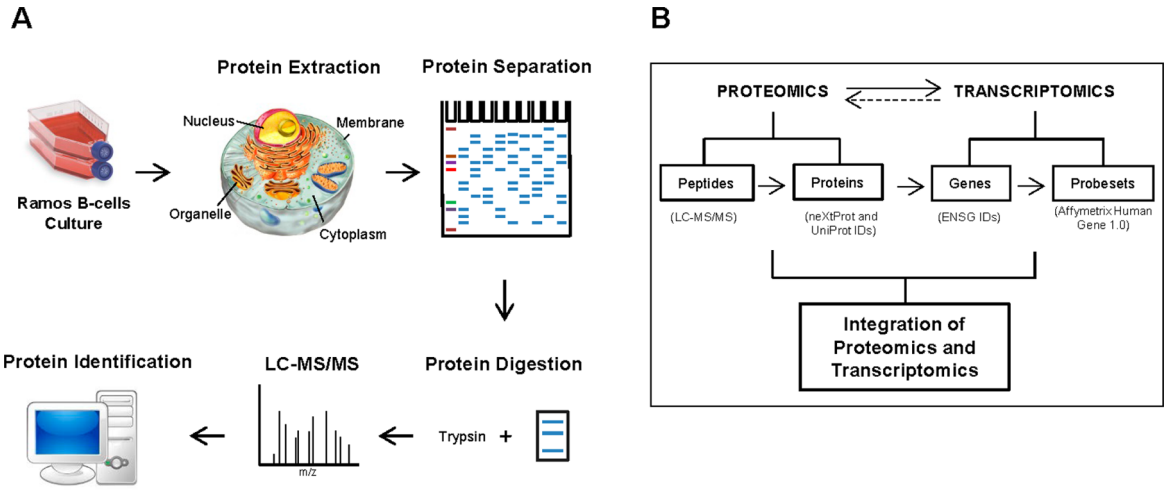


Figure 1. Schematic representation of the strategies followed for the integration of proteomics and transcriptomics analysis. (A) Overview of the experimental workflow. Subcellular protein extraction was performed from *Ramos* B-cells. After protein separation in an SDS-PAGE gradient gel (4−20%), proteins were digested with trypsin. The digests were analyzed using an nUPLC-LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). SEQUEST and MASCOT database search algorithms were used for protein identification. (B) Integration of proteomics and transcriptomics workflow. Comparison of proteomics and transcriptomics data was made via mapping from peptides (obtained by an LC-MS/MS strategy) to DNA probes (Affymetrix Human Gene 1.0 platform).

thiosulfate, dithiothreitol (DTT), iodoacetamide (IAA), formic acid (FA), and acetonitrile (ACN) were purchased from Sigma (St. Louis/MO, U.S.A.). Heat-inactivated fetal bovine serum (FBS), L-glutamine, penicillin, and streptomycin were obtained from Gibco (Scotland, U.K.). n-Dodecyl-β-D-maltopyranoside was purchased from Affymetrix (Maumee, OH, U.S.A.), Coomassie Brilliant Blue was from Merck (Kenilworth, NJ, U.S.A.), and trypsin was from Promega (Madison, WI, U.S.A.). Monoclonal antihuman antibodies conjugated with phycoerythrin (PE): CD20 (clone L27), CD22 (clone HIB22), CD11a (clone HI111), CD45 (clone HI30), CD19 (clone HIB19), and CD3 (clone SK7) were from BD Biosciences (Pharmingen, San Diego, CA, U.S.A.); CD79b (clone CB3−1), and IgG1 mouse isotypic control antibody was from BioLegend (San Diego, CA, U.S.A.). Polyclonal antihuman antibodies conjugated with PE: kappa-Ig and lambda-Ig light chains were purchased from Cytognos SL (Salamanca, Spain).

The hypotonic lysis buffer used contained 30 mM HEPES pH = 8, 20% (v/v) glycerol, 15 mM KCl, 1 mM EDTA, 2 mM $MgCl_2$, 1 mM PMSF, 1 mM TCEP, and 1% (v/v) protease inhibitor cocktail. The hypertonic lysis buffer contains the same components as the hypotonic lysis buffer, except glycerol.

Cell Culture

The *Ramos* Burkitt's lymphoma-derived B-cell line (ATCC CRL 1596) was cultured at 37 °C in a humidified CO_2 incubator (5% CO_2) in complete RPMI media (RPMI-1640 medium supplemented with 10% (v/v) FBS, 200 mM L-glutamine, 10 000U/mL penicillin, and 10 000 μg/mL streptomycin).

Cell Harvest and Cell Lysis Methods

Cellular proteins were harvested by washing the cells (30×10^6 cells/experiment) twice with PBS, and cells were pelleted for 5 min at 1000*g*. The lysis buffer was then added at a volume equal to 5 times that of the cell pellet; all steps were performed at 4 °C (Figure 1A). The cell lysis methods A and B (Supplementary Figure 1) were performed in triplicate.

Method A. The hypotonic lysis buffer supplemented with 0.015% (w/v) digitonin was added to pelleted cells. After 30 min of rotation, the sample was centrifuged at 1500*g* for 5 min, and the cytoplasmic proteins (CYT) were collected in the supernatant. The remaining fractions were processed stepwise in an identical manner. For organelle proteins (ORG), the hypotonic lysis buffer with 0.5% (v/v) Tween 20 was used; for membrane proteins (MB), the hypotonic lysis buffer supplemented with 0.5% (v/v) IGEPAL detergent was employed, and lastly, nuclear proteins (NUC) were extracted by adding the hypertonic lysis buffer supplemented with 1% (w/v) n-dodecyl-β-D-maltopyranoside after 10 min incubation. Two washing steps were performed with nonsupplemented hypotonic lysis buffer between the distinct fractionation steps.

Method B. The CYT and ORG fractions were obtained as described in Method A. The NUC fraction was extracted with hypertonic lysis buffer supplemented with 140 mM NaCl. Lastly, the MB fraction was obtained after incubating for 5 min with hypotonic lysis buffer plus 1% (w/v) n-dodecyl-β-D-maltopyranoside.

Protein Quantification and SDS-PAGE Separation

After protein quantification by the Lowry-DC-Protein Assay as recommended by the manufacturer (Bio-Rad Laboratories, CA, U.S.A.), each sample was separated in a 4−20% gradient SDS-PAGE gel under reducing conditions. The same amount of protein (15 μg) was run for each fraction (CYT, ORG, NUC, MB). After electrophoresis, gels were stained in a solution of 0.5% (w/v) Coomassie Brilliant Blue. Gels were stored at 4 °C in an aqueous solution containing 1% (v/v) acetic acid, until analysis.

In-Gel Digestion and LC-MS/MS Analysis

Each gel lane was cut into five fragments and digested with trypsin following the method of Shevchenko et al.[21] with slight modifications. Briefly, gel pieces were destained with 15 mM potassium ferrocyanide and 50 mM sodium thiosulfate. Protein reduction and alkylation were performed with 10 mM DTT at 56 °C for 45 min, and with 55 mM IAA at room temperature for 30 min, respectively. Proteins were digested with trypsin (6.25 ng/mL) at 37 °C for 18 h. The peptide solution was acidified with FA and desalted by using C18-Stage-Tips

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

columns.[22] The samples were partially dried and stored at −20 °C until they were analyzed by LC-MS/MS.

A nanoUPLC system (nanoAcquity, Waters Corp., Milford, MA, U.S.A.) coupled to a LTQ-Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose,CA, U.S.A.) via a nanoelectrospray ion source (NanoSpray flex, Proxeon, Thermo) was used for reversed-phase LC-MS/MS analysis. Peptides were dissolved in 0.5% FA/3% ACN and loaded onto a trapping column (nanoACQUITY UPLC 2G-V/M Trap Symmetry 5 μm particle size, 180 μm × 20 mm C18 column, Waters Corp., Milford,MA, U.S.A.). Peptides were separated on a nanoACQUITY UPLC BEH 1.7 μm, 130 Å, 75 μm × 250 mm C18 column (Waters Corp., Milford,MA, U.S.A.) with a linear gradient from 7% to 35% solvent B (ACN/0.1% FA) at a flow rate of 250 nL/min over 120 min.

The nUPLC- LTQ-Orbitrap Velos was operated in the positive ion mode by applying a data-dependent automatic switch between survey MS scan and tandem mass spectra (MS/MS) acquisition. Survey scans were acquired in the mass range of m/z 400 to 1600 with a 60 000 resolution at m/z 400 with lock mass option enabled for the 445.120025 ion.[23]

The 20 most intense peaks having ≥2 charge state and above the 500 intensity threshold were selected in the ion trap for fragmentation by collision-induced dissociation with 35% normalized energy, 10 ms activation time, q = 0.25, ± 2 m/z precursor isolation width and wideband activation. Maximum injection time was 1000 and 50 ms for survey and MS/MS scans, respectively. AGC was 1×10^6 for MS and 5×10^3 for MS/MS scans. Dynamic exclusion was enabled for 90 s.

Database Search

Raw data were translated to mascot general file (mgf) format and searched against the neXtProt database (release September 19, 2014) using the target-decoy strategy with an in-house MASCOT Server v. 2.3 (Matrix Science, London, U.K.). Decoy database was created using the peptide pseudoreversed method, and separate searches were performed for target and decoy databases. Search parameters were set as follows: carbamidomethylation of cysteine as a fixed modification, oxidation of methionine and acetylation of the protein n-terminys as variable ones, precursor and fragment mass tolerance were set to 10 ppm and 0.8 Da, respectively, and fully tryptic digestion with up to two missed cleavages. FDR at PSM level (psmFDR) and protein level (protFDR) were calculated using MAYU.[24] Using C-HPP guidelines, protein identifications were obtained using the criteria psmFDR < 1% and protFDR < 1%. Lastly, protein inference was performed using the PAnalyzer algorithm,[25] and nonconclusive protein groups were discarded.

For the search against UniProt database, the MASCOT[26] and SEQUEST HT[27] algorithms were used to search for the acquired MS/MS spectra, using Thermo Scientific Proteome Discoverer software (v. 1.4.1.14) against a custom database of all human reviewed sequences downloaded from the UniProt database (February, 2014) and common contaminant sequences (e.g., human keratins, trypsin, BSA). Search parameters were the same as for the search against neXtProt database. Peptides having MASCOT ion scores of <20 were not considered for analysis. A 1% FDR using Percolator[28] was employed for peptide validation as well as for PSM level.

Supporting Information 1 contains raw data about the results obtained for representative samples.

Transcriptomic Analysis

To perform the gene expression profiling and analysis of the *Ramos* B-cells, we used a raw data set of three samples of mRNA from biological replicates of these cells hybridized on Affymetrix Human Gene ST 1.0 high-density oligonucleotide microarrays. The data are available at GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) database: series number GSE40168; samples GSM987747, GSM987748, GSM987749; platform GPL6244 ([HuGene-1_0-st]). The preprocessing, normalization, and signal calculation of these data were done using the Bioconductor packages oligo[29] and pd.hugene.1.0.st.v1.[30]

Integration of Transcriptomics and Proteomics Data Sets

In order to map the neXtProt IDs of the proteins identified in the LC-MS/MS proteomic assays to the corresponding genes and probesets detected by the gene expression transcriptomics assays, an ID-mapping procedure was run in two steps (Figure 1B): (1st) from protein neXtProt IDs to gene Ensembl IDs, using the mapping to genes provided by neXtProt database (release September 19, 2014); and (2nd) from the gene Ensembl IDs to gene Affymetrix probesets IDs, using the Brainarray tool[31] with the mapping table hugene10st_Hs_ENSG_mapping 18.0.0 corresponding to the arrays used. Before this ID-mapping, we unified all the neXtProt IDs obtained from the different isolated subcellular fractions (i.e., cytoplasm, organelles, membranes, and nucleus). Following these unification and mapping procedures, the neXtProt IDs and UniProt ACs were used to create different data sets with different levels of coverage and confidence based on the combination of the results obtained with three replicates (i.e., all proteomic experiments consisted of three independent biological replicates of the *Ramos* B-cells). In this way, the following data sets were generated (Table 1): (i) an *intersection* data set, including those proteins which were systematically identified in the three replicated experiments, with at least 2 proteotypic peptides at protFDR < 0.01; (ii) an *union* data set, including all proteins identified in any of the replicated experiments with at least 2 proteotypic peptides and protFDR < 0.01, and (iii) a *maximum* data set, including all proteins identified in any of the replicated experiments with at least 1 proteotypic peptide and protFDR < 0.01. These three data sets are included in each other, the third one being the one with the largest coverage (i.e., *maximum* data set) and therefore the one that provides the largest list of proteins identified.

To map proteins to genes in chromosomes, the Biomart[32] managed with R and Bioconductor tools were used. A brief R-script is provided as Supporting Information (Supporting Information 2) to show the details of the mapping protocol and the comparison of proteomic and transcriptomic data. Functional enrichment analysis (FEA) and clustering of the gene lists were done using the DAVID[33] and GeneTerm-Linker[34] tools. The main biological databases selected to find genes with annotated enriched terms were the following: (i) Gene Ontology (GO) using annotations spaces GOTERM_BP, GOTERM_CC and GOTERM_MF; (ii) the pathways database KEGG_PATHWAY; (iii) the INTERPRO and PFAM protein structural domain database; and (iv) the UNIGENE_EST and GNF_U133A_QUARTILE tissues-specific expression databases. To generate the functional clusters in DAVID, we used classification stringency *medium*.[33] All statistical analyses of data distributions, the comparisons, and most of the mapping were done working in the R/Bioconductor environment.[35]

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 1. Number of Proteins and Genes Included in the Datasets Produced in the Analyses of *Ramos* B-Cells[a]

	no. of proteins (neXtProt)	no. of genes (Ensembl)	
	total neXtProt IDs	ENSG IDs	Affymetrix probeset IDs
intersection[b,c]	3383	3433	4088
union[b,d]	5494	5540	6175
maximum[d,e]	8931	8976	9494
"exclusive identifications" in transcriptomics[f]	-	1290	-
"exclusive identifications" in proteomics[g]	516	-	-

[a]The table indicates the number of protein and gene distinct IDs found in the proteomic and genomic assays, respectively: Row 1, *intersection* dataset; row 2, *union* dataset; row 3, *maximum* dataset (these datasets are defined in Materials and Methods). Row 4 ["Exclusive identifications" in Transcriptomics] includes the proteins that were not detected in proteomics but detected in the genomic data in the 25% highest expression quartile of the Affymetrix Human Gene ST 1.0 microarrays (calculating the expression signal average for the 3 arrays). Row 5 ["Exclusive identifications" in Proteomics] includes the genes that were not detectable by the genomic platform (i.e. genes not present in the microarray) but were detected by the proteomic approach. The columns in the table correspond to (i) all the human neXtProt IDs detected, (ii) the mapped Ensembl IDs, and (iii) the mapped Affymetrix probeset IDs. [b]Proteins detected by at least 2 unique peptides in the MS proteomic experiments. [c]Proteins detected in all the 3 experimental biological replicates. [d]Proteins detected in any replicate. [e]Proteins detected by at least 1 unique peptide in the MS proteomic experiments. [f]Genes detected in the 25% higher expression quartile of the microarrays but not present in the MS data. [g]Proteins detected in the MS/MS data but not present in the expression microarrays.

The same analysis strategy was performed for mapping UniProt IDs (using UniProtKB database release February 2014).

Immunophenotypic Analysis

Surface membrane expression of some proteins was validated by flow cytometry using a direct immunofluorescence technique with antihuman phycoerythrin (PE)-conjugated antibodies for the following proteins: CD3, CD11a, CD19, CD20, CD22, CD45, CD79b, kappa-Ig, and lambda-Ig light chains, together with a PE-conjugated isotype control antibody. All the antibodies were purchased from BDBiosciences (San José, CA, U.S.A.). Briefly, *Ramos* B-cells (0.5×10^6) were washed in 0.5% BSA in PBS and incubated with the antibodies for 15 min at room temperature, washed, and acquired on a FACSCanto II flow cytometer (BD Biosciences, San José, CA, U.S.A.) using the FACSDiva software (version 6.1, BD Biosciences). For data analysis, the Infinicyt software (Cytognos SL, Salamanca, Spain) was used.

Statistical Methods

For all continuous variables, mean values and their standard deviation (SD) were calculated. To evaluate the statistical significance of differences observed between groups, the two independent sample Student's *t* test was used for continuous variables displaying a normal distribution (SPSS software 18.0 package; SPSS, Inc., Chicago, IL). Statistical significance was set at a *P* value of <0.01.

RESULTS AND DISCUSSION

Subcellular Fractionation, Key To Increase Proteome Coverage

Mass-spectrometry techniques can increase the coverage of the proteome by using subcellular fractionation strategies for protein extraction.[20] This approach enables the in-depth analysis of biomolecules by reducing the sample complexity through isolation of different subcellular fractions. Here, we performed sequential extraction of proteins (from three biological replicates) by using a combination of different detergents specific for protein profiling from distinct subcellular

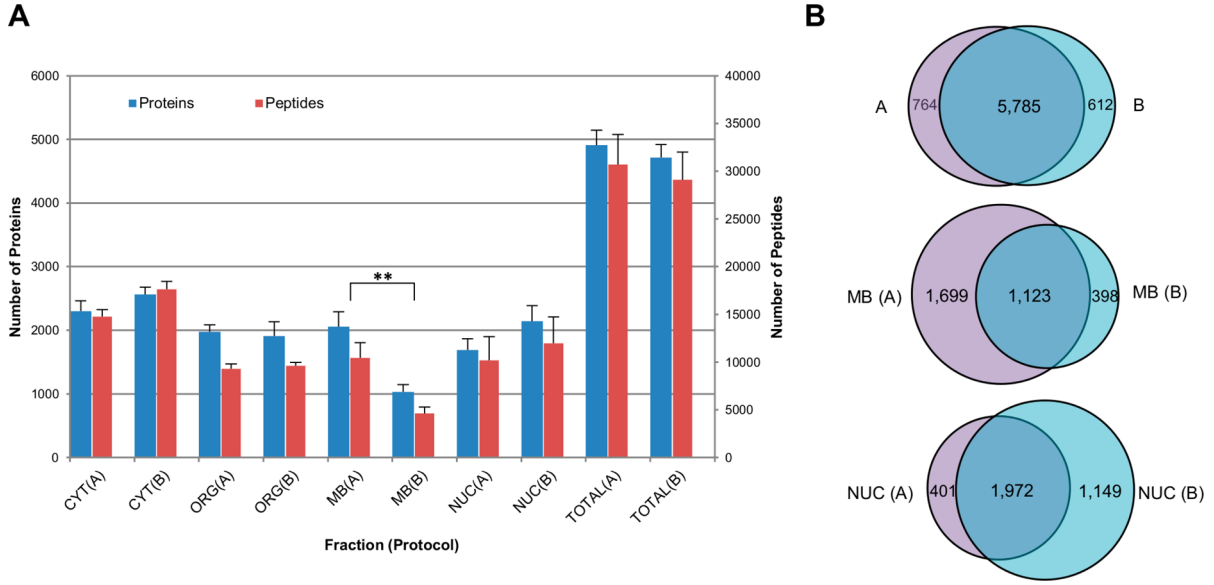


Figure 2. Comparison of protein extraction methods. (A) Total number of proteins and peptides (without duplicates) identified by LC-MS/MS assays for each subcellular fraction (CYT, cytoplasmic; ORG, organelle; MB, membrane; NUC, nuclear; and total protein extract, TOTAL) and protein extraction method (A or B). Each value was calculated as an average using the three replicates. (B) Robustness of LC-MS/MS assays for protein extraction methods A and B. Union of proteins from the three replicates (without duplicates) were considered for this comparison. Upper panel: comparison of total protein extracts. Middle and lower panels: comparisons of the membrane and nuclear proteins, respectively. ** $p < 0.01$.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

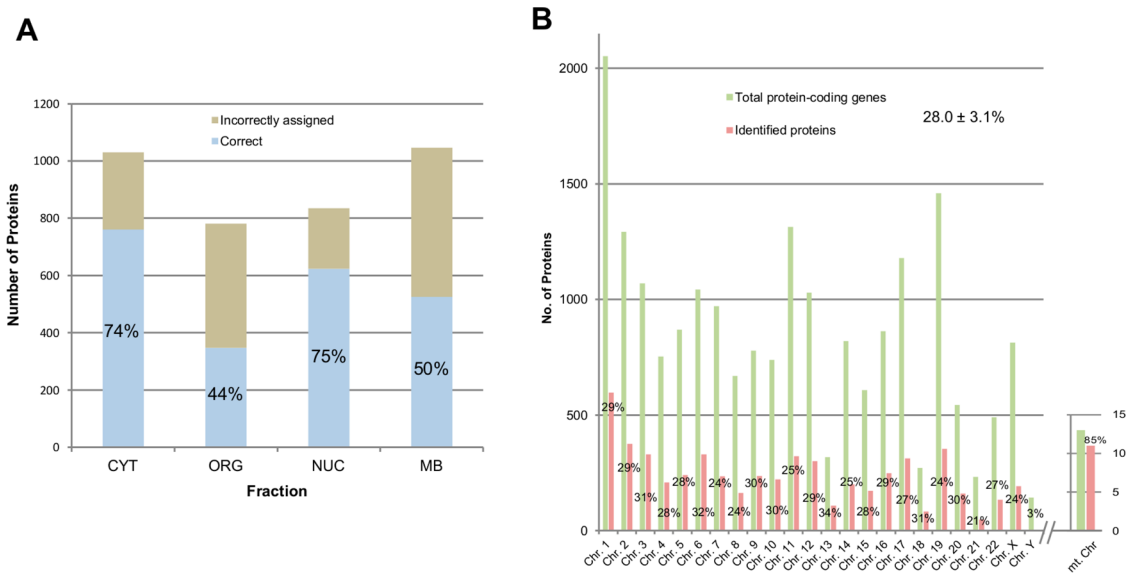


Figure 3. Subcellular protein localization. (A) Considering protein extraction method A and proteins identified in common in all three replicates (*intersection*), we validated our protein identifications for each subcellular fraction comparing with the subcellular location information given in the protein database. Correctly assigned proteins (in light blue) are the proteins whose subcellular location matches both in our MS/MS results and in the literature. Incorrectly assigned (in light brown) proteins are referred to not matching. (B) Chromosome mapping of proteins identified by at least one unique peptide (*maximum*). Green bars show the total protein-coding genes identified per chromosome (data from Ensembl release 78), red bars show the number of proteins identified by our MS/MS strategy. Corresponding percentages are noted within each bar.

compartments of *Ramos* lymphoma B-cells. The efficiency of solubilization and the maintenance of protein structure are directly dependent on the detergent choice, salt concentration, and pH. Thus, two similar approaches (methods A and B) were tested in parallel to isolate four different subcellular fractions of proteins (cytoplasmic, CYT; organelle, ORG; nuclear, NUC; and membrane, MB). These methods only differ in the extraction of membrane and nuclear proteins, which can be isolated by using IGEPAL and n-dodecyl-β-D-maltopyranoside (for method A) or NaCl and n-dodecyl-β-D-maltopyranoside (for method B), respectively. However, in both methods, the cytoplasmic and organelle fractions were similarly isolated using digitonin and Tween 20 (both are nonionic detergents), respectively. Digitonin effectively water-solubilizes membrane proteins, but not the nuclear ones which remain structurally intact. Thus, cytosolic proteins can be recovered. Second, we used Tween 20, which does not affect protein activity, to extract organelle-related proteins.

Supplementary Figure 2A displays the total amount of protein extracted from each subcellular fraction. By SDS-PAGE analysis (Supplementary Figure 2B), a homogeneous distribution of proteins independently of the molecular weight is observed.

The MS/MS data analysis of the four subcellular fractions described revealed a high number of identifications at the protein and peptide levels. In these assays, a very tight precursor mass tolerance (0.8 Da) and an FDR lower than 1% for both PSM, peptide and protein, were considered to reduce the chance of false positive identifications. The results obtained suggest that the differences in the number of identified proteins, between the protein extraction methods A and B, correlate with the detergent selected/used (Figure 2A). Specifically, the difference between the number of membrane proteins recovered with IGEPAL (method A) and n-dodecyl-β-D-maltopyranoside (method B) is statistically significant (p-value of 0.007). In case of method A, IGEPAL solubilizes membrane proteins, whereas it is not strong enough to lyse the nuclear membrane, which allows this subcellular compartment to remain intact for the effects of the n-dodecyl-β-D-maltopyranoside detergent (for extraction of nuclear proteins). In turn, in method B, the membrane fraction was isolated in the last step by using n-dodecyl-β-D-maltopyranoside.

Regarding nuclear proteins, we detected that usage of high salt concentrations (method B) was more efficient than n-dodecyl-β-D-maltopyranoside (method A) for nuclear protein extraction. In summary, these preliminary analyses indicate a slight increase in proteome coverage with method A versus method B. Therefore, the differences between both methods provide an explanation to the variations observed in protein recovery. In fact, with method A, a better recovery of membrane proteins was achieved, and we were particularly interested in the membrane fraction because it is the cellular compartment where more missing proteins are estimated to be located in this specific cell line.

For a more in-depth comparison between the two protein extraction methods, the robustness through the 3 replicates was analyzed with an overall overlap of 81%, 35%, and 56% between methods A and B for total, membrane, and nuclear proteins, respectively (Figure 2B). Of note, again method A identified a greater number of proteins than method B (up to 4 times). Therefore, on the basis of these preliminary analyses, we selected method A for further transcriptomics–proteomics comparisons owing to the higher proteome coverage not only at the total protein level but also at the membrane fraction of proteins.

Characterization of the Proteome of *Ramos* B-cells

As described above, *Ramos* B-cells derive from a Burkitt lymphoma carrying the *MYC* gene rearrangements (i.e., t(8;14)[36]), and they are often used as a model for proteomics of B lymphocytes. *MYC* gene was first discovered in Burkitt lymphoma patients as a proto-oncogene whose activation leads to the induction of cellular proliferation. Although our

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

proteomics strategy has not detected the Myc protein—mainly due to the fact that this protein is located on the nucleus, as it is further explained in the section entitled Exclusive Identifications from Comparison of both Proteomics and Transcriptomics)—a large number of proteins interacting with Myc have been identified (Actl6a, Bcl2, Chd8, Gtf2I, Mapk1, Max, Mlh1, Mycbp2, Mycbp, Nmi, Nyfc, Pfdn5, Ruvb, Sap130, Smad2, Smad3, Smarca4, Smarcb1, Taf9, Wdr5, Yyi, among others). Thus, this -omic platform allows the characterization of cell signaling pathways by identifying the components of the interactions leading to changes in cellular responses.

In order to evaluate the robustness of the MS/MS data set and its reliability for integration with transcriptomics data sets, the presence of cross-contamination between the subcellular fractions obtained with method A was determined according to the protein database. Thus, the percentage of the proteins identified in each subcellular fraction that were correctly isolated by the protein extraction method is presented in Figure 3A. In total, 74−75% was the proportion found for the fraction of proteins assigned by the protein database to cytoplasm (CYT) and nucleus (NUC). The lowest percentages for correct protein location corresponded to the organelle (ORG) and the membrane protein (MB) fractions. Both fractions were sequentially and consecutively extracted, and the nature of their protein mixture is highly related, hindering their correct localization. In addition, the literature-based protein database assignment can be ambiguous because many times there is more than one unique location for each protein, and additionally, there are membrane-associated organelle proteins that could be included in both fractions. Because our main goal was to profile the *Ramos* cell proteome, all proteins identified were grouped into a unique data set independently of their correct or incorrect location. In this regard, it is important to keep in mind that subcellular fractionation was specifically used to increase protein recovery, not for the independent analysis of the fractions.

On the basis of chromosome mapping (Figure 3B) of the proteins identified, overall we identified about 30% of all protein-coding genes (~5000−6000) present in the human genome (without considering the X and Y chromosomes, and mitochondrial DNA). This confirms that this approach is proteome-wide and unbiased, proving that selecting for a specific cell type is a useful approach for total proteome coverage. Interestingly, the proteins identified also covered about an 85% of the total protein-coding genes of the mitochondrial DNA, using the LC-MS/MS approach here employed. This good coverage of the mitochondrial proteins indicates a great improvement with respect to other proteomic procedures and platforms (e.g., GFP-tagging analysis[37]). This is mainly due to the development of proteomic strategies with high sensitivity and accuracy in combination with the usage of subcellular fractionation approaches that allow the enrichment of the organelle proteomes such as the mitochondrial one.[38]

Other studies have recently published proteomic data related to B-cells lymphomas by using SILAC approaches. Mann's lab has developed a super-SILAC to classify large B-cell lymphomas subtypes by their protein profiles.[39] In turn, Rüetschi and colleagues studied the same disease by a SILAC-based quantitative approach.[40] A comparison of our results revealed an overlap with 1696 and 2141 proteins identified in Mann's and Rüetschi's studies, respectively, constituting 54% and 60% of proteins in common (data not shown). The differences in identification might be due to the absence of subcellular fractionation steps in these studies that improves the recovery and protein identification. Even so, B-cell specific proteins have been detected by both approaches, demonstrating the feasibility of proteomics strategies to characterize and model the disease.

Integration of Proteomics and Transcriptomics Data Sets

The characterized MS/MS proteomic data was compared to transcriptomic measurements performed on *Ramos* B-cells (3 biological replicates) with genome-wide expression high-density oligonucleotide microarrays (Figure 4). To accomplish the proteomics-transcriptomics integration, the neXtProt IDs (corresponding to identified proteins by the LC-MS/MS approach) were mapped into the Ensembl IDs (for the genes), and these were mapped into the Affymetrix probesets IDs that identify the gene-specific DNA oligo probes which are present in the microarrays (Supporting Information 3,4 and 5). Such comparison was addressed from different ways termed *intersection*, *union*, and *maximum*, as described above in the Materials and Methods section. On the basis of this strategy, the following goals were pursued: (i) to identify as many proteins as possible using this approach present in *Ramos* B-cells (*maximum* and *union*), and (ii) to characterize those proteins that are unequivocally identified with this proteomic approach (*intersection*). Briefly, 3383, 5494, and 8931 known human proteins (i.e., neXtProt IDs) were found within the *intersection*, *union*, and *maximum* data sets respectively (Table 1). These data sets corresponded to 3433, 5540, and 8976

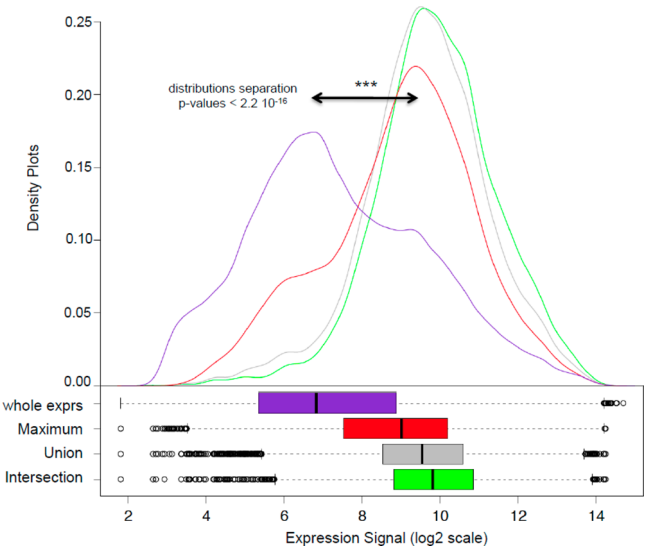


Figure 4. Density plots and boxplots showing the distributions of the whole gene expression signal versus the signal corresponding to the genes detected in the MS proteomic profiles for *Ramos* B-cells. The expression was measured with *Affymetrix* high-density oligonucleotide expression arrays type Human Gene 1.0. (Purple) Whole expression signal corresponding to all the 33 297 probesets included in the arrays [whole exprs]. (Red) 9494 probesets corresponding to the 8976 distinct human genes (ENSG IDs) and 8391 proteins (neXtProt IDs), which showed at least one unique identifying peptide in the MS proteomic assays [*maximum*]. (Gray) 6175 probesets corresponding to the 5540 distinct human genes (ENSG IDs) and 5494 proteins (neXtProt IDs), which showed at least two identifying peptides in the proteomic assays [*union*]. (Green) 4088 probesets corresponding to the 3433 distinct human genes (ENSG IDs) and 3383 proteins (neXtProt IDs), which had at least two identifying peptides in the proteomic assays and were found in all the replicates of the proteomic isolations [*intersection*].

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

genes (i.e., Ensembl IDs). These results indicate that proteomics and transcriptomics studies, once applied to the global molecular characterization of a cell type, display a high accuracy and overlap within each analytical level. Such overlap can be estimated considering the proportion of proteins that are detected by proteomics and transcriptomics, because—as we explain below—there were only 516 proteins exclusive of proteomics (i.e., not detected by the transcriptomic platform) out of a total of 8931, and this corresponds to a 94% of overlap of the expression data over the proteomic data.

FEA of those 3383 proteins detected in a coherent and steady manner (*intersection* data set) by the proteomics assays (Supporting Information 6) revealed the expression of many essential proteins for general cell functions and house-keeping processes (e.g., anabolism and synthesis processes, together with catabolism and cellular respiration) as well as the activity and regulation of major biological macromolecules (DNA, RNA, and proteins) involved in key maintenance processes like cell cycle, cell growth, cell proliferation, and so forth.

On the other side, mapping of proteins to tissue-type and cell-type databases provided a clear enrichment in B-cell specific genes and proteins that was in agreement with the character and properties of a lymphocytic cell type. For example, regarding B-cell receptor (BCR) cross-linking, many intracellular signaling cascades appended to be activated leading to regulation of gene expression. Moreover, synthesis initiation proteins (eIF3a/h proteins), proteins for cellular adhesion and costimulatory signaling (CD11a), accessory signal transduction components (such as CD20, CD19, CD79a or CD79b), protein tyrosine kinases (Fyn, Lyn, Syk and Btk), and proteins related to the activation of Ras signaling pathway (N-, K-, and H-Ras) were identified. In turn, proteins belonging to the B-cell receptor signaling pathways have been widely detected including those related to BCR internalization (SHP-1, CD19, Bam32), cytoskeletal rearrangements and integrin activation (Bam32, PLCγ2, DAG), transcription (Blnk, Grb2, SHP-1, Erk1/2, Raf), proteasomal degradation (PKC, Carma1, Bcl10, MALT1, IKK, NF-κB) and growth arrest and apoptosis (Akt, FoxO), among others pathways that follow BCR activation.

Exclusive Identifications from Comparison of both Proteomics and Transcriptomics

As mentioned before, it is notorious that the overlap observed between both the proteomic and the transcriptomic methodologies applied to the same cell type, although these two experimental technologies had also a complementary part. In fact, comparing the results of both approaches, each one enables to cover the failures in identification ("exclusive identifications") due to its technical limits with respect to the other strategy.

Performing the FEA for the exclusively identified proteins in proteomics (Supporting Information 7), we identified a gain of proteins related to the mitochondrial and ribosomal organelles, as well as cytoplasmic ones. This is a quite expected result because the genomic microarrays employed do not include probes for mitochondrial DNA. In addition, we could infer that our subcellular extraction method in the proteomic procedure was effective on isolating organelle proteins, as it is shown for mitochondria. A loss of identifications associated with immunoglobulins (Ig) and major histocompatibility complex (MHC) proteins was also detected in the transcriptomic data probably due to Ig gene rearrangements and hypersomatic mutations of Ig genes that hamper the design of adequate array probes for these genes.

Regarding exclusively identified proteins in transcriptomics, the functional enrichment analysis (Supporting Information 7) revealed identifications related to nuclear and DNA-binding proteins. This is probably due to the fact that isolating the nuclear fraction in the last step of the proteomics approach decreases the recovery for nuclear proteins. Upon comparing the method chosen for this study (method A) with method B (Figure 2B), it seems clear that extracting the nuclear fraction in a previous step and with another detergent improves isolation of these proteins. Therefore, it could be appropriate to previously establish the protein fraction of interest to perform the best protein extraction method for such fraction. In our case, we select method A as the overall most suitable approach combining all fractions and also because we were interested in enriching for membrane proteins since these are the most commonly lost in many proteome-wide studies. Additionally, around 300 of these transcripts exclusively identified in transcriptomics correspond to noncoding RNAs. Thus, it is obvious that their corresponding proteins have not been detected by the proteomics platform.

Missing Proteins

Since thousands of human proteins have not been detected yet (as it is noted in the last release of missing proteins from the neXtProt database), the exploration of proteome data sets has become an indispensable exercise that may be accomplished to reduce the number of existing missing proteins. With this purpose, we have performed the mapping of our results into the neXtProt database for missing proteins (current release of April 28, 2015) obtaining the results shown in Table 2. Specifically, the searching has been carried out across the three data sets generated (*maximum, union*, and *intersection*) identifying up to 370 missing proteins from our *maximum* data set (containing 8931 neXtProt identifications). As the number of neXtProt IDs decreases from one data set to another (from *maximum* to *union* and *intersection*), the number of identified missing proteins decreases as expected (from 370 to 32 and 4, respectively). However, this reduction in number involves an increase in quality because these missing proteins have been identified in 3 different biological replicates and with, at least, 2 peptides per protein. In Supporting Information 8 is shown the information related to missing proteins identified for each data set and the related data (PE group, chromosome location).

Table 2. Missing Proteins Identified Across the Three Datasets Obtained after Searching against the neXtProt Database[a]

	PE group				
data set	PE2	PE3	PE4	PE5	total
maximum	273	37	5	55	370
union	18	3	-	11	32
intersection	-	-	-	4	4

[a]Proteins from each dataset (*maximum, union*, and *intersection*) were mapped into the neXtProt missing proteins database (April 28, 2015)—used as reference of missing proteins—to identify missing proteins. The missing proteins have been classified accordingly to their protein existence (PE) level (i.e., PE2 for experimental evidence at transcript level; PE3 for protein inferred from homology; PE4 for predicted protein; and PE5 for uncertain proteins).

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

Comparative Analysis of the Results Obtained with neXtProt and UniProt Database Searches

In a further analysis, the effect of database search on identification was evaluated. With this purpose, we additionally performed a search against the UniProt database and compared the results (Supporting Information 9, 10, 11, 12) with those obtained with neXtProt. In general, neXtProt database search generated a greater number of identified proteins what determined, consequently, an increased number of missing protein identifications compared to UniProt (Supporting Information 13 and 14). The integration of information from different databases (Swiss-Prot, Ensembl, Human Protein Atlas, PeptideAtlas....) makes possible a better characterization of the proteomes and justifies this increase in identifications (more spectrum and peptide information in the database leads to a possible increasing in identifications). In addition, this supposes an improvement in the characterization of the specific functions of identified proteins. In Supporting Information 15 are reported the FEA of proteins exclusively identified in proteomics and transcriptomics for searches performed against neXtProt and UniProt, respectively, reporting an enriched annotation of the functions for neXtProt search.

Immunophenotypic Characterization of the *Ramos* B-cells

In order to give support and provide some external validation of the characterization of the cell type studied in our proteomic and transcriptomic studies, we use an independent cell-oriented platform to characterize the immunophenotype of the *Ramos* B-cells. With this purpose, several proteins were selected and screened by multiparametric flow cytometry (FCM) (Supplementary Figure 3) showing high expression levels of the B-cell associated antigen—CD19, CD20, CD22 and CD45—as well as the coreceptor of the B-cell receptor, CD79b. Expression of the lambda-Ig light chain was high as well, confirming the available data for the *Ramos* cell line (ATCC CRL 1596, www.atcc.org). Additionally, CD11a, an integrin involved in cellular adhesion and costimulatory signaling was evaluated showing a dim expression. As negative controls for FCM assays, T-cell marker CD3 and kappa-Ig light chain were also tested.

PERSPECTIVES AND CONCLUSIONS

Integration of proteomics and transcriptomics technologies has emerged as a potent strategy for the mapping of the biomolecules that define a cell type or a cellular state. The searches derived from both -omics methodologies are complementary, and once performed in conjunction one with the other, they allow a mutual validation and increase the total coverage of genome-wide protein-coding genes identified. With this goal in mind, the C-HPP initiative has promoted an effective combination of proteomics data into a genomic framework to achieve a full map of the human proteome that will provide a better understanding the studied biological systems.

Based on deep characterization of the proteome of the *Ramos* lymphoma B-cells by LC-MS/MS, a total of up to 6000 proteins and ∼30 000 different peptides were identified. This data resulted in about 30% coverage of all human gene-coded proteins per chromosome found in the proteome of these lymphoma B-cells (such percentage being as high as 85% for the mitochondrial DNA). Integration of this proteomics data set with transcriptomics data sets (derived from high-density oligonucleotide microarrays technology) showed an 82% overlap between both technologies. Despite this, several gaps were identified with each of the two technologies. Regarding proteomics, special attention must be given to the protein extraction method that could affect correct extraction of proteins from specific subcellular compartments; moreover, proteins expressed at very low concentrations and in highly complex multimers could also have problems in being detected. For this reason, extensive fractionation of the proteome of interest might contribute to improve the resolution of proteomics via detection of thousands of proteins simultaneously, improving current knowledge about the functional and biological cellular processes via higher coverage of the proteins involved in these mechanisms.

Finally, the characterization of PTMs within the lymphoma proteome would be of great interest to model the disease. In this sense, further studies will be performed to determine the PTMs that may influence the normal response behavior of these cells.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00474.

Additional data as noted in the text (ZIP)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jrivas@usal.es. Phone: +34 923294819. Fax: +34923294743.
*E-mail: mfuentes@usal.es. Phone: +34 923294811. Fax: +34 923294743.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Carlos III Health Institute of Spain (ISCIII, FIS PI11/02114, FIS PI14/01538, and FIS PI12/00624), Fondos FEDER (EU) and Junta Castilla-León SA198A12-2. The Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported by grant PT13/0001. P.D. and C.D. are supported by a JCYL-EDU/346/2013 Ph.D. scholarship. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium[41] via the PRIDE partner repository with the dataset identifier PXD001933. We thank Peter Horvatovich for providing us the list of missing proteins (release 2015-04-28).

ABBREVIATIONS

ACN, acetonitrile; BCR, B-cell receptor; C-HPP, Chromosome-human Proteome Project; CYT, cytoplasmic proteins; DTT, dithiothreitol; FA, formic acid; FBS, fetal bovine serum; FCM, flow cytometry; FDR, False Discovery Rate; FEA, Functional Enrichment Analysis; FT-ICR, Fourier Transform-Ion Cyclotron Resonance; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; HUPO, Human Proteome Organization; IAA, iodoacetamide; Ig, immunoglobulin; IGEPAL, octylphenoxypolyethoxyethanol; LTQ, Linear Trap Quadrupole; MB, membrane proteins; MHC, major histocompatibility complex; MS, mass spectrometry; NUC, nuclear proteins; ORG, organelle proteins; protFDR, FDR at protein level; psmFDR, FDR at PSM level; PMSF, phenylmethane-

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

sulfonyl fluoride; PTM, post-translational modification; TCEP, tris (2-carboxyethyl) phosphine hydrochloride

REFERENCES

(1) Muñoz, J.; Heck, A. J. R. From the human genome to the human proteome. *Angew. Chem., Int. Ed.* **2014**, *53*, 10864−10866.
(2) Segura, V.; Medina-Aunon, J. A.; Mora, M. I.; Martínez-Bartolomé, S.; Abian, J.; Aloria, K.; Antúnez, O.; Arizmendi, J. M.; Azkargorta, M.; Barceló-Batllori, S.; et al. Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *J. Proteome Res.* **2014**, *13*, 158−172.
(3) Ansong, C.; Purvine, S. O.; Adkins, J. N.; Lipton, M. S.; Smith, R. D. Proteogenomics: Needs and roles to be filled by proteomics in genome annotation. *Briefings Funct. Genomics Proteomics* **2008**, *7*, 50−62.
(4) Jensen, O. N. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 391−403.
(5) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57−63.
(6) Toung, J. M.; Morley, M.; Li, M.; Cheung, V. G. RNA-sequence analysis of human B-cells. *Genome Res.* **2011**, *21*, 991−998.
(7) Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M. F.; Kellis, M.; Lindblad-Toh, K.; Lander, E. S. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 19428−19433.
(8) Woo, S.; Cha, S. W.; Merrihew, G.; He, Y.; Castellana, N.; Guest, C.; Maccoss, M.; Bafna, V. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **2014**, *13*, 21−28.
(9) Renuse, S.; Chaerkady, R.; Pandey, A. Proteogenomics. *Proteomics* **2011**, *11*, 620−630.
(10) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andrén, P. E.; et al. Chromosome 19 annotations with disease speciation: A first report from the global research consortium. *J. Proteome Res.* **2013**, *12*, 135−150.
(11) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473*, 337−342.
(12) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114−1125.
(13) Castellana, N.; Bafna, V. Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* **2010**, *73*, 2124−2135.
(14) Krug, K.; Nahnsen, S.; Macek, B. Mass spectrometry at the interface of proteomics and genomics. *Mol. BioSyst.* **2011**, *7*, 284−291.
(15) Jacob, F.; Goldstein, D. R.; Fink, D.; Heinzelmann-Schwarz, V. Proteogenomic studies in epithelial ovarian cancer: established knowledge and future needs. *Biomarkers Med.* **2009**, *3*, 743−756.
(16) Haider, S.; Pal, R. Integrated analysis of transcriptomic and proteomic data. *Curr. Genomics* **2013**, *14*, 91−110.
(17) Cox, B.; Kislinger, T.; Emili, A. Integrating gene and protein expression data: Pattern analysis and profile mining. *Methods* **2005**, *35*, 303−314.
(18) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.et al. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10*, 10.1074/mcp.M111.009993.
(19) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; et al. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11*, 2005−2013.
(20) Segura, V.; Medina-Aunon, J. A.; Guruceaga, E.; Gharbi, S. I.; Gonzälez-Tejedo, C.; San Chez Del Pino, M. M.; Canals, F.; Fuentes, M.; Casal, J. I.; Martínez-Bartolomé, S.; Elortza, F.; Mato, J. M.; Arizmendi, J. M.; Abian, J.; Oliveira, E.; Gil, C.; Vivanco, F.; Blanco, F.; Albar, J. P.; Corrales, F. J.; et al. Spanish human proteome project: Dissection of chromosome 16. *J. Proteome Res.* **2013**, *12*, 112−122.
(21) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1*, 2856−2860.
(22) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2*, 1896−1906.
(23) Olsen, J. V.; de Godoy, L. M. F.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4*, 2010−2021.
(24) 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*, 2405−2417.
(25) Prieto, G.; Aloria, K.; Osinalde, N.; Fullaondo, A.; Arizmendi, J. M.; Matthiesen, R. PAnalyzer: a software tool for protein inference in shotgun proteomics. *BMC Bioinf.* **2012**, *13*, 288.
(26) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.
(27) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(28) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and sensitive peptide identification with mascot percolator. *J. Proteome Res.* **2009**, *8*, 3176−3181.
(29) Carvalho, B. S.; Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **2010**, *26*, 2363−2367.
(30) Carvalho, B. Platform design info for Affymetrix HuGene-1_0-st-v1. R package version 3.8.0.
(31) Dai, M.; Wang, P.; Boyd, A. D.; Kostov, G.; Athey, B.; Jones, E. G.; Bunney, W. E.; Myers, R. M.; Speed, T. P.; Akil, H.; Watson, S. J.; Meng, F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **2005**, *33*, E175.
(32) Durinck, S.; Spellman, P. T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184−1191.
(33) Dennis, G.; Sherman, B. T.; Hosack, D. A.; Yang, J.; Gao, W.; Lane, H. C.; Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **2003**, *4*, P3.
(34) Fontanillo, C.; Nogales-Cadenas, R.; Pascual-Montano, A.; de Las Rivas, J. Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. *PLoS One* **2011**, *6*, e24289.
(35) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.
(36) Williams, S. C.; Winter, G. Cloning and sequencing of human immunoglobulin V lambda gene segments. *Eur. J. Immunol.* **1993**, *23*, 1456−1461.
(37) Gabaldón, T.; Huynen, M. A. Shaping the mitochondrial proteome. *Biochim. Biophys. Acta, Bioenerg.* **2004**, *1659*, 212−220.
(38) Huber, L. A.; Pfaller, K.; Vietor, I. Organelle proteomics: Implications for subcellular fractionation in proteomics. *Circ. Res.* **2003**, *92*, 962−968.
(39) Deeb, S. J.; D'Souza, R. C. J.; Cox, J.; Schmidt-Supprian, M.; Mann, M. Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles. *Mol. Cell. Proteomics* **2012**, *11*, 77−89.
(40) Coiffier, B.; Lepage, E.; Briere, J.; Herbrecht, R.; Tilly, H.; Bouabdallah, R.; Morel, P.; Van Den Neste, E.; Salles, G.; Gaulard, P. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **2002**, *346*, 235−242.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX

(41) Vizcaíno, J.; Deutsch, E.; Wang, R.; Csordas, F.; Reisinger, F.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32*, 223−226.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. XXXX, XXX, XXX−XXX