Journal of proteome research

Article

pubs.acs.org/jpr

Proteogenomics Dashboard for the Human Proteome Project

Daniel Tabas-Madrid,[†] Joao Alves-Cruzeiro,[†] Victor Segura,[‡] Elizabeth Guruceaga,[‡] Vital Vialas,[†] Gorka Prieto,[§] Carlos García,[∥,⊥] Fernando J. Corrales,[‡,⊥] Juan Pablo Albar,[†,⊥] and Alberto Pascual-Montano*[,†,⊥]

[†]ProteoRed-ISCIII, National Center for Biotechnology-CSIC (CNB), C/Darwin 3, Madrid 28049, Spain

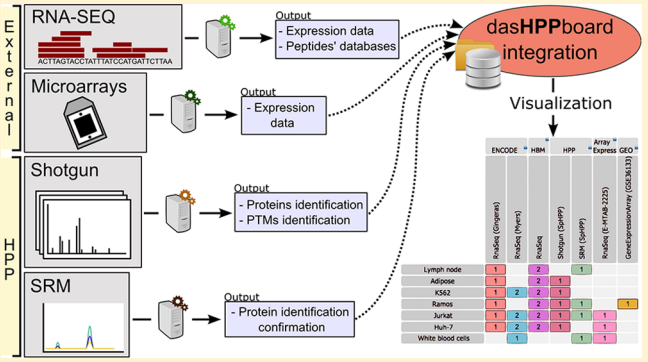
[‡]ProteoRed-ISCIII, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pío XII, 55, Pamplona E-31008, Spain

[§]Department of Communication Engineering E.T.S. Ingeniería de Bilbao, University of the Basque Country (UPV/EHU), Alda. Urquijo, s/n, Bilbao 48013, Spain

[∥]Computer Science Faculty, Complutense University of Madrid (UCM), C/ Jose Garcíá Santesmases 9, Madrid 28040, Spain

Supporting Information

ABSTRACT: *dasHPPboard* is a novel proteomics-based dashboard that collects and reports the experiments produced by the Spanish Human Proteome Project consortium (SpHPP) and aims to help HPP to map the entire human proteome. We have followed the strategy of analog genomics projects like the Encyclopedia of DNA Elements (ENCODE), which provides a vast amount of data on human cell lines experiments. The dashboard includes results of shotgun and selected reaction monitoring proteomics experiments, post-translational modifications information, as well as proteogenomics studies. We have also processed the transcriptomics data from the ENCODE and Human Body Map (HBM) projects for the identification of specific gene expression patterns in different cell lines and tissues, taking special interest in those genes having little proteomic evidence available (missing proteins). Peptide databases have been built using single nucleotide variants and novel junctions derived from RNA-Seq data that can be used in search engines for sample-specific protein identifications on the same cell lines or tissues. The *dasHPPboard* has been designed as a tool that can be used to share and visualize a combination of proteomic and transcriptomic data, providing at the same time easy access to resources for proteogenomics analyses. The *dasHPPboard* can be freely accessed at: http://sphppdashboard.cnb.csic.es.



KEYWORDS: *Human Proteome Project, chromosome 16, spHPP, proteomics, transcriptomics, RNA-Seq, ENCODE, bioinformatics, proteogenomics*

INTRODUCTION

The Human Genome Project,[1] completed in 2003, successfully sequenced nearly the entire human genome and provided the blueprints for the proteome characterization with the identification of a potential panel of protein coding genes; however, it rapidly became evident that our knowledge of about two-thirds of these genes was considerably limited, including an absence of protein-level evidence for ~35% of them.[2] It was in this scenario that the idea to progress to the next level, mapping the entire human proteome, began to take form. In 2008, the Human Proteome Organization (HUPO) started discussing about the creation the Human Proteome Project (HPP), which was finally launched in 2010 and consisted of a wide-global collaborative effort to map and characterize all of the proteins encoded by human genes. The mission of the HPP is to generate the map of the protein-based molecular architecture of the human body and become a resource to help elucidate biological and molecular function and advance diagnosis in treatment of diseases. Three major pillars support the project to accomplish this task: mass spectrometry (MS), antibody/affinity capture reagents (Ab), and bioinformatics driven Knowledge base (KB).[3] To enhance the understanding of the human proteome, HUPO divided the HPP into two major programs: the Chromosome-based Human Proteome Project (C-HPP), to characterize the human proteome on a chromosome-by-chromosome basis, and the Biology/Disease Human Proteome Project (B/D-HPP), which provides a framework for the coordination of biology- and diseased-based research groups' contributions to the HPP.[4] One of the

Special Issue: The Chromosome-Centric Human Proteome Project 2015

Received: May 27, 2015
Published: July 6, 2015

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

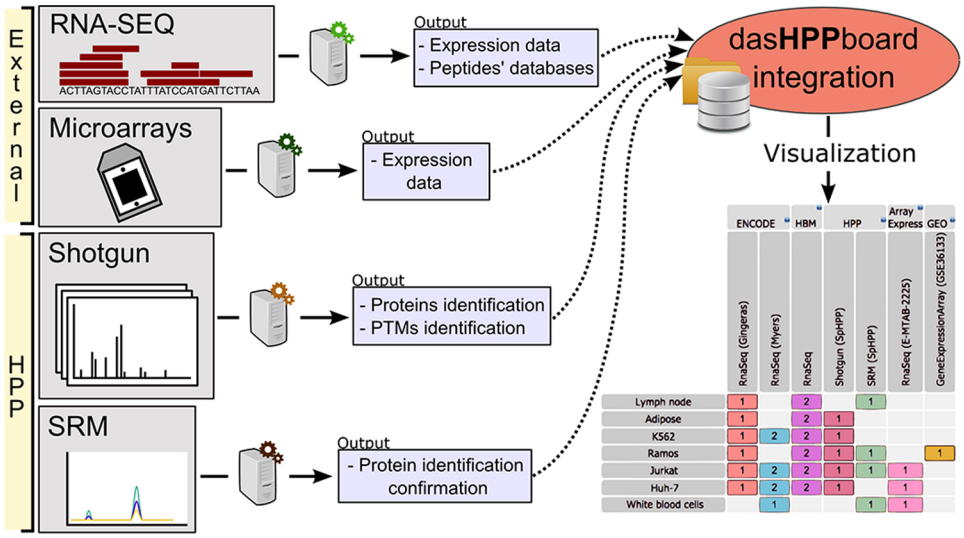


Figure 1. Overall picture of the *dasHPPboard* structure, data sources, and types of results displayed.

C-HPP primary goals is to put the proteomic catalog in the genomic context to improve the understanding of its biological context and to promote more effective collaborations with molecular biologists. To manage these international collaborative efforts in an efficient way, the consortium divided the work into groups by countries representing all of the human chromosomes.

Nowadays we can differentiate three maps of the human proteome: the work of the HPP and the ones presented in Kim et al.[5] and Wilhelm et al.[6] articles. These works are changing the existing panorama into a much richer scenario for research.

Within this context, the Spanish HPP consortium (SpHPP) has been responsible for the characterization of the chromosome 16. The SpHPP consortium, composed of 16 laboratories, is organized into five working groups: Protein expression and Peptide Standard, SRM Protein Platform, Clinical Healthcare and Biobanking, Protein Sequencing, and Bioinformatics.[7] According to Ensembl database (V79),[8] chromosome 16 contains about 90 million base pairs, representing almost 3% of the total DNA in human cells, including 842 protein-coding genes reported by neXtProt (release 2014-09-19),[9] of which 110 are considered to be missing. These missing proteins have only transcriptomic evidence and a predicted sequence (or are inferred by homology) or are partially identified proteins, one in which there is transcript evidence of the existence of the corresponding protein without convincing MS information. The search for experimental evidence of these so-called "missing proteins" constitutes the first phase of the C-HPP and is expected to be finished by 2018.[10]

The great variety of cell types in the human body, which is ~230, translates into an even greater set of different transcriptomes and therefore different proteomes. Unlike other groups that focus on finding the missing proteins in cell lines and tissues of their choice, we firmly believe that a proteogenomics approach is needed, taking advantage of all existing transcriptomics studies to make better selection of samples prior to this search. In this work, we have developed a proteomics-centric dashboard named *dasHPPboard* that collects and reports the experiments produced by the HPP consortium. Not only is this bioinformatic tool a protein browser as the one developed by other chromosomes (the Proteome Browser,[11] CAPE,[12] GenomewidePDB,[13] H-InvDB,[14] Chromosome 18 Knowledgebase,[15] ...) but also it includes the necessary databases to perform proteogenomics studies to detect SAPs and novel junctions. A first logical approximation integrates the data produced by the Spanish contribution to the HPP project.

To cover as much transcriptomic information as possible to make the proteogenomics map a reality, we considered the inclusion of data external to the HPP project. The ENCODE project[16] was selected as a source data set because of the variety of cell lines it hosts and the large number of RNA-Seq experiments from each of these lines. This project began in 2003 with the mission to identify all of the functional elements encoded in the human genome. To fulfill this mission, the institutions involved in the project focused on generating multiple complementary types of genome-wide data, like RNA-Seq, ChIP-Seq, or measurement of DNA methylation. Since the beginning of the HPP project, a collaboration between the HPP and ENCODE is being considered. For example, in Paik Hancock (2012),[17] the authors defend that a combined effort between the two initiatives would help decipher how the interacting genomic elements documented by ENCODE control the families of isoforms generated at the protein level, and although this collaboration has been discussed a great number of times, with specific examples of action such as the search in ENCODE for "missing" proteins that have been difficult to identify in proteomic studies, a robust study of this kind had not yet been made. The SpHPP made in 2014 a first approach by building a transcriptomic map for chromosome 16, based on data from three cell lines extracted from ENCODE and studied by the Spanish consortium (MCF-7, K-562, and HepG2).[2] Additionally, proteogenomics studies of the Illumina Human Body Map (HBM) and the Cancer Cell Line Encyclopedia[18] (CCLE) have also been included. There are other studies, like Human Protein Atlas database,[19] which have recently incorporated many transcriptomic data, including raw files available through platforms like ArrayExpress or GEO, which could be processed by our tool. We intend to add them to our platform in future updates of *dasHPPboard* to improve the cell lines and tissue coverage. Transcriptomics data have been analyzed to identify those cell lines and tissues showing significant expression levels for those genes that are difficult to observe by standard mass spectrometry. The dashboard

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

facilitates proteogenomics analyses by providing peptide databases built using the information on single nucleotide variants (SNVs) and novel junctions obtained from RNA-Seq data. We want to remark with the addition of these databases the importance of detecting alternative splice variants to better detect protein products coming from these transcripts. These databases can be used for the identification of protein variants present in shotgun proteomics experiments on the same cell line or tissue from where those SNVs were retrieved.

Here we present what is, to the best of our knowledge, the only dashboard containing results from transcriptomics and proteomics experiments in a proteogenomics approach that links both fields. We have summarized the structure and aims of the *dasHPPboard* in Figure 1. There are other download centers and dashboards, like the ones developed for the ENCODE project, one of which is hosted in http://genome.crg.es/encode_RNA_dashboard/hg19/ and has been a source of inspiration when defining the graphical part of the overview of our Web site. Comparing with information hosted in ENCODE dashboard, we have several conceptual differences. Our Web site is oriented to experimentalist users, including fully processed end results generated with state-of-the-art methods in order to minimize the time required to understand and apply them to their research. In contrast, the ENCODE dashboard contains, for RNA-Seq data, fastq raw sequencing data, and in most of the cases, bam alignment files and results like bed or bigwig files containing chromosome regions and coverage, which is by far much more complicated to interpret. In the case of microarrays, they only display raw signal, which needs to be further processed to know if a gene is expressed.

Including all of this information, *dasHPPboard* is intended to be a one-stop shop for data produced in the HPP context.

MATERIALS AND METHODS

We present the workflows we have implemented to generate results for the different data types we include in *dasHPPboard*: transcriptomics data, including Microarray and RNA-Seq workflows along with the proteogenomics approach we have used with the NGS data; proteomics data, including Shotgun proteomics with its PTMs pipeline and SRM workflows. Finally, we include details of implementation of the web tool and database that form the *dasHPPboard*.

Analysis of Transcriptomics Data

In this work, we have downloaded and analyzed different publicly available data sets of RNA-Seq and microarray experiments, calculated their gene expression, and uploaded it in an interactive, easy-to-use dashboard designed specifically for this purpose. The RNA-Seq data sets used were extracted from the ENCODE project,[16] the Illumina Human Body Map 2.0 project (HBM), and the Gröschel et al. (2014) study on acute myeloid leukemia cell lines.[20] Three microarray data sets have also been included, with expression data from the Cancer Cell Line Encyclopedia (CCLE) project,[18] tumor samples from IGC's Expression Project for Oncology (expO) and a study of gene expression in tissues across the human body.[21] The gene expression was calculated for all chromosomes and for all samples available in these data sets, resulting in a very complete transcriptomic map for the entire human genome.

Processing RNA-Seq Data. After downloading raw FASTQ files from the different projects, we performed the RNA-Seq analyses following the steps detailed later and summarized in Figure 2. First of all, a quality control step

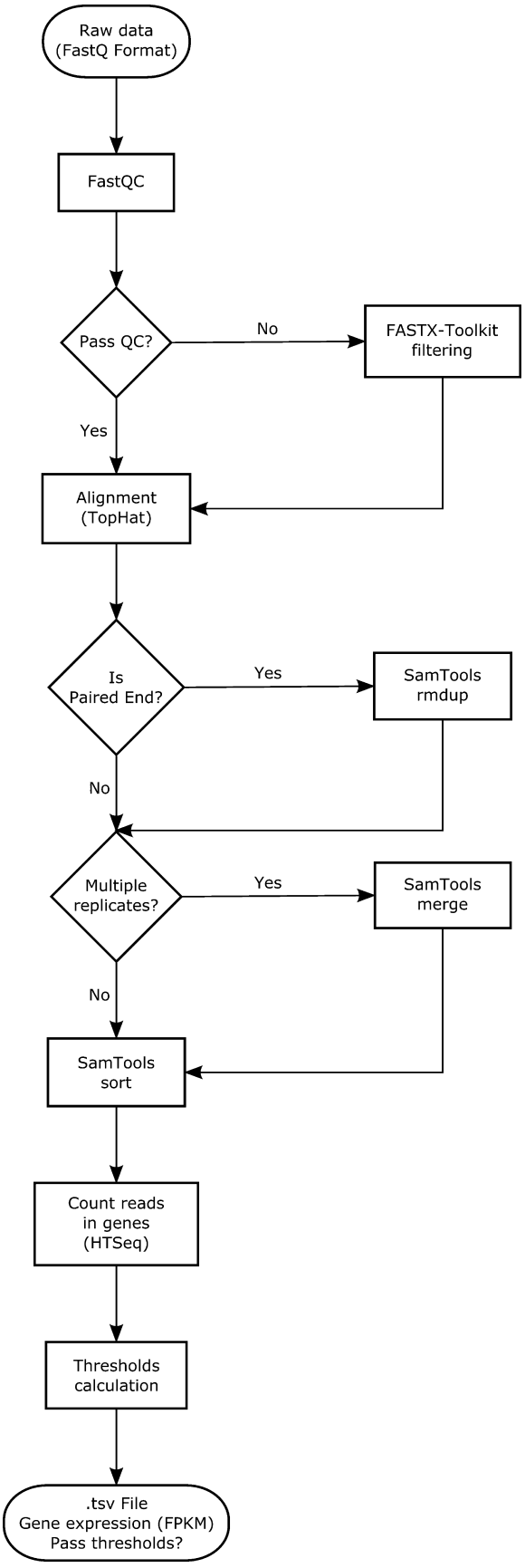


Figure 2. Workflow of transcriptomic analysis from the raw data to the FPKM counting.

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

was performed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) software, and residual adapters and contaminants were removed using the FASTX-toolkit[22] (http://hannonlab.cshl.edu/fastx_toolkit/). High-quality remaining reads were aligned using TopHat[23] v2.0.9 employing the GRCh37 assembly as the human reference genome. SAMtools[24] rmdup tool was used to remove potential PCR duplicates. This software, used specifically with paired-end alignments, in the case of multiple read pairs with identical external coordinates retains the pair with the highest mapping quality. SAMtools sort and SAMtools merge were used to sort the bam files generated in the alignment by read names and to join multiple sorted alignments, respectively. The quantification process was done in two steps: (1) using the htseq-count script (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) from HTSeq[25] package with GENCODE[26] V18 annotations to count the fragments and (2) using a Ruby script to calculate the FPKM (fragments per kilobase of transcript per million mapped reads).

A single FPKM value is not very valuable when trying to determine if a gene is really expressed if we do not have a plausible threshold to compare it with. This is even more problematic when the expression value is low; however, despite the great advances in NGS technology, errors in sequencing are inherent to the technology, and alignment algorithms can map reads to the wrong location. These problems can lead to deceiving FPKM values due to background noise. Some authors propose to address this issue by simply keeping an arbitrary percentage of the highest expressed genes, but this strategy can filter out true expressed genes; therefore, we believe that a more robust method to analyze this data is needed. With this in mind, we have used two very reliable threshold-defining methods available: the one explained in Ramskold et al[27] and the Bgee method (http://bgee.unil.ch/bgee/bgee?page=documentation#RNASeqData) based on the Hebenstreit et al.[28] work.

For each RNA-Seq experiment, the two thresholds were calculated, and the results include graphs of those calculations. Furthermore, for each experiment that includes gene quantification, a new field with threshold information was added to results, getting values of "HIGH QUALITY" when gene expression value is above both thresholds, "LOW QUALITY" when expression value is only above one of the thresholds, and "NO EXPRESSION" when none of the thresholds have been surpassed.

These methods are both based on measuring reads not annotated to genes interpreting them as intergenic noise. GENCODE V18 annotations were used to do that, taking gene ranges and reversing them to obtain intergenic regions. The authors in this study[29] observed that the average read density is dramatically higher in the gene start and end zones, and it is correlated with known genes; therefore, they should not be considered. As a result, the intergenic regions were adjusted by removing 10kb regions from gene starts and ends. The read measurement of the intergenic regions was performed with htseq-count.

Threshold method from Ramskold et al.[27] is defined as follows. First of all, the expression information on genes and intergenic regions is binned using a logarithmic scale from $\log_{10}(0.01)$ to $\log_{10}(100)$ taking steps of 0,1 on this transformed scale, and then converted to cumulative amounts of genes and intergenic regions expressed above different levels. A false discovery rate (FDR) is calculated for each expression level, which is then used to estimate the true number of expressed genes in each bin from gene expression information. Also, a false negative rate (FNR) is calculated as a function of this estimated true number of expressed genes. Finally, the threshold is defined as the intersection between the FDR and the FNR. This method was implemented as a Ruby script, including some R code to generate spline curves from the different FDR and FNR points calculated and to know the exact intersection point between them.

The threshold method explained in Bgee Web site is slightly different but also uses the gene and intergenic regions expression information. Expression levels are binned, and for each point, the value of the proportion of genes and intergenic regions from that point compared with the total proportion is calculated. The minimum point at which this value exceeds a fixed value is set as the threshold. The fixed value we use in our implementation of this threshold calculation is 0.05. This method was developed entirely in Ruby code.

Proteogenomics Analysis. We have used two different approaches for the generation of the databases containing peptides with single amino acid polymorphisms (SAPs) or novel junctions between exons. In both cases the input data are the RNA-Seq data filtered and aligned as explained in Figure 2. All of the processing steps of both analyses have been performed on the CNB or CIMA computer clusters.

In the case of SAP databases, mpileup command of SAMtools (v 0.1.18) was used for the SNV calling. The obtained binary file (BCF file) was converted to the variant call format (VCF file) with BCFtools (v 0.1.17-dev) and filtered with the vcfutils.pl script of SAMtools. SNVs with a read depth lower than 10 or a quality score lower than 10 are not considered for further analysis. The remaining SNVs were annotated with the variant_effect_predictor.pl script of Ensembl that converts SNVs to the corresponding amino acid coordinates and retrieves the SIFT and PolyPhen-2 scores for each SNV.[38] We used this information to filter the synonymous SNVs with R/Bioconductor. Sequences of 80 amino acids around the mutated peptide were extracted from the Ensembl protein FASTA file. The SAP database includes the SAPs whose sequences do not exist in the reference proteome (Ensembl v73). A similar proteogenomic strategy was successfully used by Sheynkman et al.[39]

Databases of the resulting peptides from novel junctions observed in RNA-Seq experiments were generated with functions of customProDB R package.[40] The Tophat BED file with putative splice junctions was read and the novel junctions were three-frame translated to peptide sequences. Those peptide sequences that are identical to existing peptides in the reference database were filtered. In this case, the length of the peptides is not constant and depends on the length of the putative novel junctions defined by TopHat based on each RNA-Seq experiment.

This pipeline has been implemented to use RNA-Seq data at sample level, so in experiments including more than one sample, these databases will be available when expanding the corresponding sample information. The corresponding reference databases including contaminant sequences provided by cRAP (http://www.thegpm.org/crap/) and decoy sequences using pseudoreverse trypsin method are also provided in the *dasHPPboard* for SAPs and novel junctions databases.

Data Sets. ENCODE data. In this work, we have analyzed the gene expression of all chromosomes and all cell lines from ENCODE that have whole-cell Long PolyA+ RNA-Seq data,

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

comprising a total of 23 cell lines and 44 different experiments, out of the 111 cell lines that contain RNA-Seq data in that project. Some of the cell lines had more than one experiment, varying from 1 to 5, and some of the experiments had more than one biological sample with different runs, in a total of 44 experiments (15 single end, and 29 paired end) and 288 sequencing runs. The RNA-Seq FASTQ files were downloaded from the ENCODE RNA dashboard (http://genome.crg.es/encode_RNA_dashboard/hg19/). The list of the cell lines from ENCODE used for this study can be found in Table 1.

Table 1. List of the 23 Cell Lines from ENCODE Used in Our Study

cell line name	tissue of origin	number of experiments included	total number of samples included
BJ	skin	1	2
IMR90	lung	1	2
K562	blood	3	6
SK-N-SH	brain	1	2
NHEK	skin	3	6
HEPG2	liver	3	6
GM12892	blood	1	3
GM12891	blood	1	2
HELA-S3	cervix	3	6
HSMM	muscle	2	4
H1-HESC	embryonic stem cell	5	10
GM12878	blood	4	7
HUVEC	blood vessel	3	6
NHLF	lung	2	4
A549	epithelium	1	2
MONOCYTES-CD14+	monocytes	1	2
HMEC	breast	1	1
SK-N-SH_RA	brain	1	2
LHCN-M2	skeletal muscle myoblast	2	2
CD20+	blood	1	2
MCF-7	breast	2	5
HCT-116	colon	1	2
AG04450	lung	1	2
total		44	86

Acute Myeloid Leukemia Cell Lines.[20] In this study, RNA deep sequencing has been used to determine gene expression patterns and expression genotypes in acute myeloid leukemia (AML) cell lines that harbor a 3q-aberration. The authors wanted to uncover the mechanism by which 3q aberrations result in the overexpression of EVI1, which is involved in leukemogenesis. For this purpose, they performed RNA sequencing of AML cell lines overexpressing this gene and others that do not express it. The raw data from this sequencing study, which includes samples from seven different cell lines, are freely available in ArrayExpress, with the accession number E-MTAB-2225.

Human Body Map 2.0 Data. In 2011, Illumina published a new RNA-Seq data set, as part of the HBM project, generated on HiSeq 2000 instruments. This project provides a transcription profiling by high-throughput sequencing of individual and mixture samples of 16 human tissues: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. Also, in this case, transcriptomic analysis using this data set for the chromosome 16 has been done,[1] which we now completed for the full set of the human chromosomes, and added the results to the *dasHPPboard*. The HBM data are publicly available and the corresponding accession numbers are GSE30611 in GEO[30] database, E-MTAB-513 in ArrayExpress,[31] and ERX011226 in SRA.[32]

The subset of experiments of the HBM project that has been analyzed is detailed in the Supporting Information Table 1 and does not include the experiments with mixture samples. There are two experiments: one single and one paired end experiment for each individual tissue sample that consists of one run each.

Processing Microarray Data. The results obtained by Guruceaga et al.[37] about the cell lines, normal tissues, and tumor samples where each protein is more likely to be expressed, have been included in the *dasHPPboard*. In brief, all of the microarray experiments used in this study were processed using the same analysis pipeline. The statistical environment R/Bioconductor[33] was used for data preprocessing, analysis and visualization. Both background correction and normalization were performed using the fRMA (Frozen Robust Multichip Average) algorithm.[34] After the normalization step, an expression probability threshold was calculated for each sample to distinguish the expressed probe sets from the nonexpressed ones. The Gene Expression Barcode algorithm[35] has been designed to estimate this value taking into account a huge amount of publicly available experiments from different Affymetrix platforms, including the Affymetrix Human Genome U133 Plus 2.0 Array. The output of the algorithm is the z score value under the unexpressed normal distribution for each probe set. A threshold of $z > 2$ was used to identify the expressed genes in a sample. A Naive Bayes Classifier[36] predicted the protein expression probability in a sample based on the 3′UTR, 5′UTR, and CDS gene sequence lengths, the probability of an expressed gene being a coding gene of a missing protein in a certain sample and the gene barcode for each biological sample. The obtained result was a score that ranks the proteins from the most probably expressed to the least probably expressed in each sample.[37]

Data Sets. Three microarray public experiments have been analyzed. The first data set contains expression data from the Cancer Cell Line Encyclopedia (CCLE) project, 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 pharmacological characterization of a large panel of human cancer models. It provides public access to the transcriptome profiling of 917 human cell lines spanning 36 cancer types using the Affymetrix Human Genome U133 Plus 2.0 Array platform. Raw data were downloaded from GEO using the accession number GSE36133. Human normal tissue samples were obtained from GEO database using the accession number GSE3526.[21] This data set was generated using samples from 10 post-mortem donors. Samples were processed to generate total RNA, which was subsequently analyzed for gene expression using the Affymetrix Human Genome U133 Plus 2.0 Array.[21] A total of 352 samples from 65 tissues are available. Finally, 2158 tumor samples representing 156 different tissues that were hybridized on the Affymetrix Human Genome U133 Plus 2.0 Array have been also included in the analysis. This data set, provided by the IGC's Expression Project for Oncology (expO), is publicly available in GEO database with accession number GSE2109.

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

Analysis of Proteomics Data

Analysis of Shotgun Proteomics Data. Information about proteins and peptides identified in the shotgun experiments performed by the SpHPP consortium has been included in the *dasHPPboard*. These experiments include the proteome characterization of MCF7, CCD18, Jurkat, and Ramos cell lines. In brief, raw MS/MS data obtained using several high-performance spectrometers (Orbitrap, Q Exactive, MaXis Impact, and 5600 triple TOF) were converted to mgf (mascot general file) format and used to perform searches against UniprotKB[41] human database (release 2013_06, June 13) using Mascot v2.4. Search parameters were set as follows: carbamidomethyl cysteine as fixed modification and oxidized methionines and acetylation of the peptide amino termini as variable modifications. Peptide mass tolerance was set to 50 ppm, in both MS and MS/MS modes, and two missed cleavages were allowed. Typically, an accuracy of ±10 ppm was found for both MS and MS/MS spectra. Using these search parameters, including filter peptides with fewer than seven amino acids, the target-decoy approach to calculate the FDR, and employing MIAPE Extractor Software v. 2.92 (http://www.proteored.org/miape-extractor), we considered protein identifications with FDR < 1% at both peptide and protein level.Finally, protein inference was performed using PAnalyzer[42] algorithm.

Analysis of Postranslational Modifications. One of the goals for the first phase of the C-HPP is to characterize three major postranslational modifications (PTMs) (i.e., phosphoryl-, glycosyl-, and acetyl-) for each protein.[43] In the SpHPP consortium we have selected phosphorylation as the first PTM to study, and we will include acetylation in a next step.

The experiments previously submitted by the SpHPP consortium to proteomeXchange (accessions PXD000442, PXD000443, PXD000447, PXD000449) have been reprocessed again but now including phosphorylation of S, T, and Y as variable modifications. The input to the processing pipeline consists of the mgf files associated with the different fractions of each experiment, and the search process has been carried out running X!Tandem in a cluster. The X!Tandem output files have been converted to the HUPO-PSI standard format mzIdentML[44] using the mzidLibrary[45] version 1.6.10. The resulting mzid files have been processed using a new version of the PAnalyzer software (2.0-alpha1), which carries out the protein grouping and also identifications filtering. The PSM filter selects only the best PSM for each peptide and precursor charge combination. Then, using these PSMs we have calculated a p value for the peptides, which are subsequently filtered according to this p value using a 1% FDR at peptide-level criterion. Finally, we have calculated a p value for protein groups and applied a second FDR filter of 1% at protein group level.

Analysis of Selected Reaction Monitoring. The selected reaction monitoring (SRM) data was generated in a collaborative manner in proteomics laboratories belonging to the Spanish consortium of the HPP project. The implemented strategy consists of the development of SRM methods and cross-validation assays across the different laboratories of the consortium. In this way, tested, reliable methods are delivered.

The experiments were performed using at least three cell lines that are used in common by all the laboratories in the SpHPP consortium: MCF7, CCD18, and Ramos. Other SRM assays were generated using the additional cell lines: TC28, Jurkat, and HUH7; and tissues: HUAEC (Human Umbilical Artery Endothelial Cells), plasma, and serum. As previously described,[2] these results include solid tested SRM methods for 43 proteins. In brief, proteotypic peptides were obtained from PeptideAtlas[46] and GPMDB[47] databases, and optimal transitions for every target peptide from each of the proteins were selected on the basis of previous MS/MS spectra acquired from the same original samples and also from predictions and data in online databases. The SRM analysis was performed on ABSciex 4000 and 5500 QTRAP instruments. The chromatographic setup consisted of a 90 min linear gradient (5 to 40% ACN in 0.1% formic acid; 300 nL/min) in C18 nanocolumns (75 μm id, 15 cm, 3 μm particle size) (LC Packings, Netherlands; Eksigent, USA; Thermo Scientific, USA).

Dwell time was set to 20 ms per transition, and collision energies were optimized using SRM Pilot Software.

In addition to peptide confirmation by SRM-triggered MS2 spectrum sequence searching (MIDAS), the proteins in this set have been detected with at least 3 transitions per peptide, with 4.16 transitions per peptide on average. The method has been cross-validated by at least two other laboratories different from the original laboratory where the method was developed.

As a result, we have generated SRM results files for chromosome 16 including MCF7, CCD18, HUH7, Jurkat, Ramos, and TC28 cell lines and HUAEC, serum, and plasma tissues.

Dashboard Implementation

The *dasHPPboard* has been built by joining several components of different technologies to classify, access, and display results in a modular and independent way.

First of all, a library for organizing and accessing data is needed. Data need to be linked to a set of metadata to be organized and displayed correctly. For this purpose, a directory tree in which each level represents a different data feature has been used. Having this metadata embedded in directory names has the important advantage of the considerable savings in space needed for storing it. The inclusion of new data in the directory tree can be performed through a script written in Ruby. A library to perform queries on the data and their associated metadata has also been developed.

The hierarchical structure of the directory tree, depicted in Figure 3, first divides data by the project they belong to, which is in turn associated with the type of samples they contain (tissues, cell lines, cancer tissues). Information is subsequently stored in folders according to the following hierarchy: the version of the genome that has been used to analyze the data, the name of the tissue or cell line, and finally folders with information on the types of experiments on which results are present. On the basis of this library to perform queries and using the Nanoc framework, which generates programmatically an html view with Ruby, we have developed the web visualization of the dashboard overview. This web contains all of the metadata information, descriptions, and download links of the different file types that our Web site contains. All of the information in this hierarchical structure corresponds to final and offline-generated results, which are chromosome-oriented and have been splitted to download the minimum amount of information possible for each query, not exceeding the order of a few kilobytes in each download.

The gene and protein search section has been developed with Ruby on Rails framework, which contains a module that enables the association between database tables and programmatic objects. We have populated the database using a

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

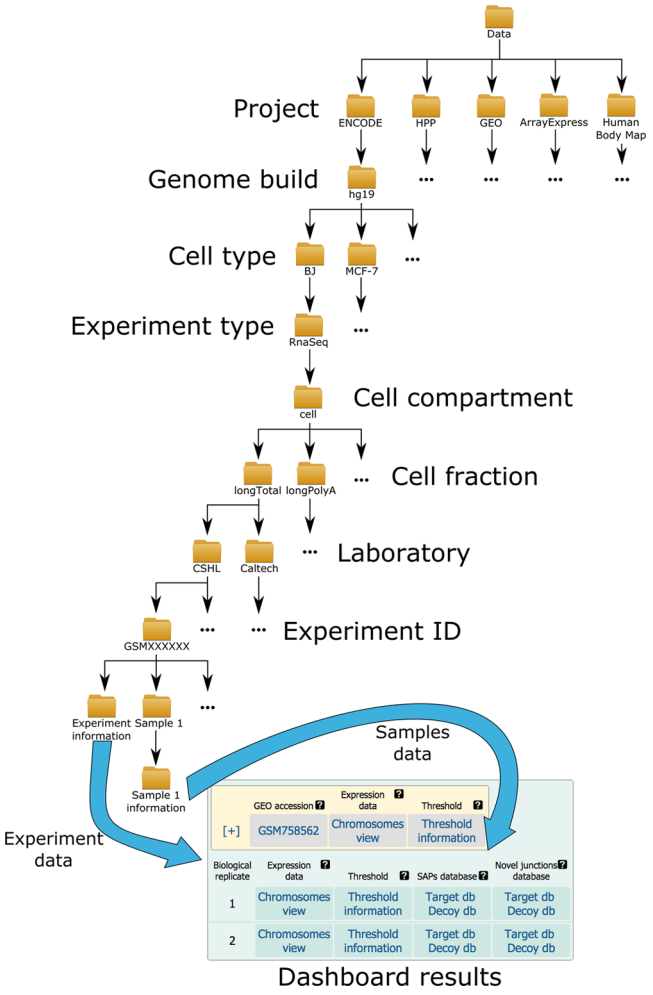


Figure 3. Graphical description of the internal data structure of the *dasHPPboard*.

seeding script written in Ruby. This script takes the different result files in the data directory tree of the dashboard, parses them, and inserts them into the MySQL database, taking into account whether they are transcriptomic results, in which expressed genes are tagged, or proteomic results, in which all data are preserved. The search interface admits input in gene names, Ensembl Gene IDs, Uniprot IDs, and neXtProt accessions. The results of the query include: (a) brief information on the input gene or protein and (b) a table of the experiments, providing links to the data in the dashboard overview.

All of the source codes of the library and Ruby on Rails project are available through https://github.com/danieloop/dasHPPboard.

RESULTS AND DISCUSSION

Up to now, we have generated about 24 gigabytes of results in text and images, so we needed an interface to display it in a compact and easy to browse way, which may result a challenging task. During the *dasHPPboard* development we have taken this into account, making a visual overview in tabular form showing the cell and experiment types and the number of samples for each of them. At present we display 3523 experiments from different samples, separated in tabs corresponding to 948 cell lines, 77 normal, and 156 cancer tissues. All of these numbers are summarized in Table 3. For

Table 2. List of Features Included in the Experiments' Results Tables That Are Available for Download in *dasHPPboard*

	RNA-Seq	microarrays	shotgun	PTMs	SRM
Ensembl Gene ID	×	×	×	×	×
gene name	×	×	×	×	×
gene description	×	×	×	×	×
FPKM	×				
expression quality	×				
neXtProt ID	×	×	×	×	×
thresholds information	×				
Z score		×			
protein missing	×	×	×	×	×
group of proteins peptide belongs to			×		
Uniprot protein isoforms			×	×	
Uniprot canonical protein ID			×	×	×
type of protein group (unique, group of isoforms, protein family)			×		
neXtProt protein evidence			×	×	×
protein inference category (Panalyzer)			×	×	
Mascot score			×		
number of peptides that support the identification			×		
peptide type				×	
peptide sequence				×	×
PTM sequence				×	
peptide *p* value				×	
peptide *q* value				×	
PTM type				×	
PTM position				×	
PTM reported by Uniprot				×	
PTM reported by Phosphositeplus				×	
retention time					×
collision energy					×
experiment setup					×

each of these tabs, experiments are organized first by data source or project and then subclassified by laboratory or data series to which they belong. To simplify the search and display of the data, we only show at first the number of experiments for each experiment type and group of results, adding different colors to the cells. Clicking on a cell of results, users can visualize more information from that specific experiment of certain cell type. That cell is expanded, and a table with different type of information depending on the experiment type is displayed. As a header of the expanded cell, redundant information about cell and experiment type and data about the compartment and cell fraction to which the sample belongs is displayed. If there is information about the laboratory where the experiment has been done, it is also displayed. Although there are multiple types of experiments, they share a chromosome's view where users can access to download the results of a particular chromosome separately, saving download time and space when interested only in a single chromosome. Internally, we have distinguished several types of experiments when displaying the results: RNA-Seq, gene expression arrays (microarrays), and shotgun proteomics including PTMs information and SRM. This internal structure and the results

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

Table 3. Summary of the Number and Type of Experiments Hosted in the *dasHPPboard*

			number of lines/tissues/cancer tissues	number of experiments
cell lines	Encode	RnaSeq (Gingeras)	cell lines = 19 tissues = 0 cancer tissues = 0 total = 14	20
		RnaSeq (Myers)	cell lines = 14 tissues = 0 cancer tissues = 0 total = 14	24
	HPP	Shotgun (SpHPP)	cell lines = 4 tissues = 0 cancer tissues = 0 total = 4	4
		SRM (SpHPP)	cell lines = 6 tissues = 0 cancer tissues = 0 total = 6	6
	ArrayExpress	RnaSeq (E-MTAB-2225)	cell lines = 7 tissues = 0 cancer tissues = 0 total = 7	7
	GEO	GeneExpressionArray (GSE6133)	cell lines = 917 tissues = 0 cancer tissues = 0 total = 917	917
tissues	HBM	RnaSeq	cell lines = 0 tissues = 16 cancer tissues = 0 total = 16	32
	GEO	GeneExpressionArray (GSE3526)	cell lines = 0 tissues = 65 cancer tissues = 0 total = 65	352
	HPP	SRM (SpHPP)	cell lines = 0 tissues = 3 cancer tissues = 0 total = 3	3
cancer tissues	GEO	GeneExpressionArray (GSE2109)	cell lines = 0 tissues = 0 cancer tissues = 156 total = 156	2158
total			cell lines = 948 tissues = 77 cancer tissues = 156	3523

available for each of the types of experiments are depicted in Figure 1. The resulting webpage can be accessed at http://sphppdashboard.cnb.csic.es.

When expanding RNA-Seq results, rows with the different samples that belong to the specific type of cell and experiment are displayed. For each sample the following information is shown: GEO or ArrayExpress accession if available; a chromosomes' view, which is expanded when clicked, where users can select the chromosome for which they want to download a results table containing information described in Table 2. We also show details of calculated expression thresholds, including Ramskold et al.[27] method threshold value and the corresponding graph including FDR and FNR, and if there are no biological replicates, also the Bgee method threshold value and a graph with the expression level proportions of genes and intergenic noise and also links to download target and decoy peptide databases generated from SAPs and novel junctions found in the sample through previously commented proteogenomic methods. If biological replicates are present, users are allowed to click on the row from the corresponding sample to expand it and obtain information about each of the replicates, which is the same as displayed in samples without replicates.

In microarray results, available information when expanding certain cell type and experiment is also organized by rows corresponding to the different samples belonging to it. The GEO accession of the selected sample is provided as well as a chromosomes' view to download expression data from a specific chromosome, whose tables contain information depicted in Table 2, including a Z score calculated as explained in Materials and Methods section, giving "NA" values to that marked as nonexpressed ($z < 2$).

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

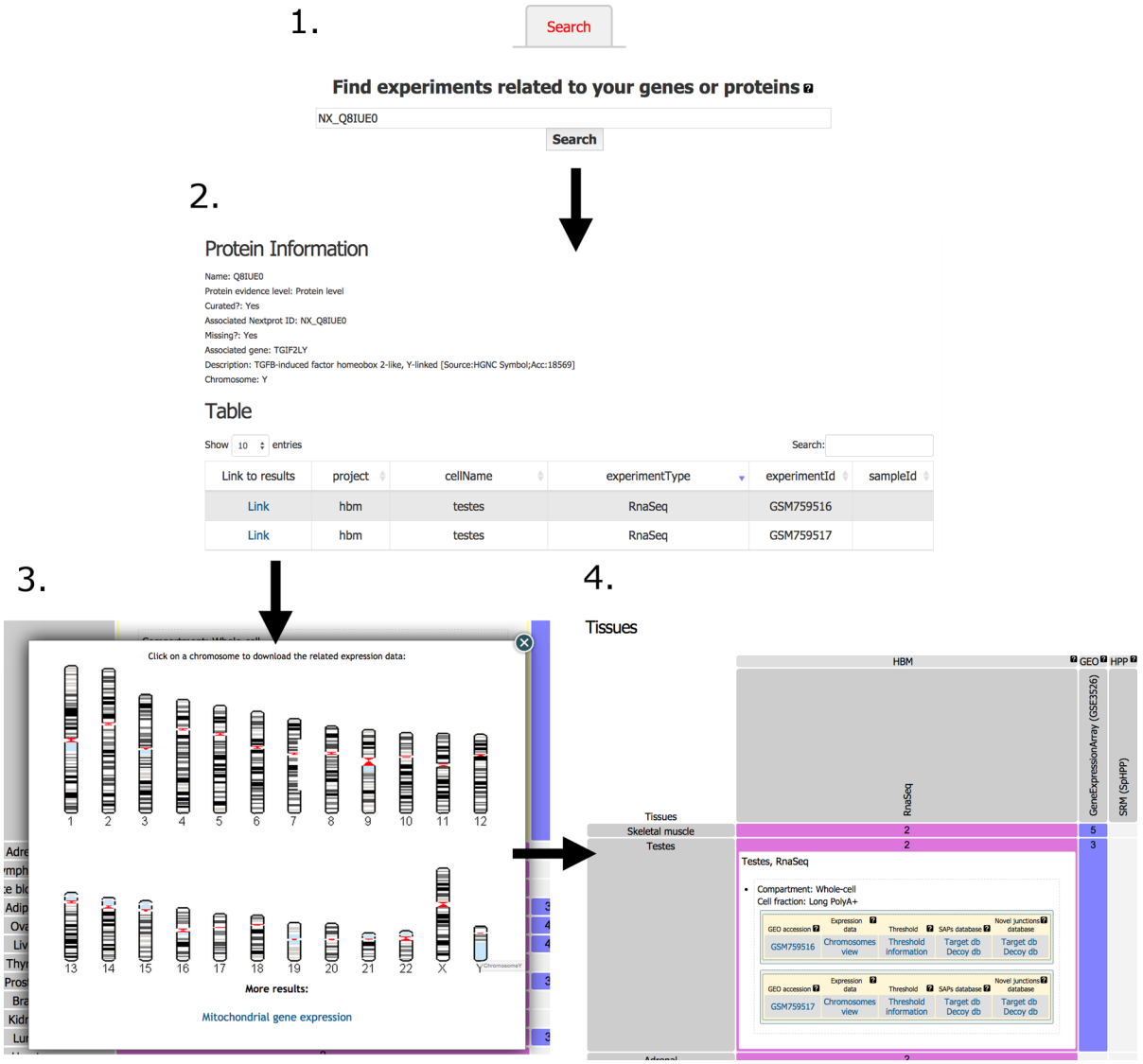


Figure 4. Example of use of the "Search" section of the *dasHPPboard*.

Expanded shotgun proteomics results include a ProteomeXchange[48] accession for each cell type and a chromosomes' view to download results in a separate way. These chromosome results are tabular files that contain experiment information at peptide and protein level, whose characteristics are described in Table 2. These results may be further expanded, displaying PTMs information for each of the experiments that have been performed with the sample. For each experiment, information about experimental setup and a chromosomes' view to download PTMs reports is displayed. These reports consist of TSV files with an entry for each protein and PTM combination indicating multiple characteristics described in Table 2 including a list indicating whether each of these positions has already been reported by UniProt and PhosphoSitePlus.[49]

Each expanded SRM results table includes a chromosomes' view in order to download SRM result files from the corresponding experiment. The tab-separated values include multiple columns that are listed in Table 2.

For the time being, of all 3523 experiments, 3510 are classified as transcriptomics type. Because of this, the greatest value of our *dasHPPboard* consists of showing information about genes corresponding to missing proteins in different cell lines and tissues. From this information, it is possible to determine which type of cell would be the best candidate when making proteomics experiments to detect those missing proteins and contribute to complete the human proteome map, which is the ultimate goal of the HPP. As shown in previous work,[2] transcriptomics and shotgun proteomics experiments can be complementary when searching missing proteins. As it has been demonstrated in the study of Guruceaga et al.,[37] expression distribution of genes that should produce missing proteins in transcriptomics experiments is very different across cell lines and specially tissues. If someone is interested in finding a certain missing protein, the ideal would be first searching it in the whole set of transcriptomics experiments of the *dasHPPboard* and then choosing a cell line or tissue where it is significantly expressed. This is no guarantee of finding the missing proteins when doing shotgun proteomics experiments, but the probability will always be greater than making a blind search. The query section of the *dasHPPboard* is very useful in this sense. Users can directly search the protein of interest and look into transcriptomics studies where it is present. As an example of the use of this query section, we have

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

provided an example displayed in Figure 4. The query section is accessible through the tab titled "Search". Next, users have to provide a gene or protein of interest. In our example, we have chosen the protein with neXtProt identifier NX_Q8IUE0 marked as missing in the release of 2014-09-19 (Figure 4.1). When clicking on the "Search" button a query is performed and a table with results is generated. Then, ordering the table by experiment type, the top two rows correspond to RNA-Seq results, in which *dasHPPboard* has found significant expression levels for the NX_Q8IUE0 protein (Figure 4.2). The cell type corresponds to testis tissue, which agrees with the description present in neXtProt database. Finally, these results can be downloaded by clicking in "Link to results", also redirecting us to the main panel of the dashboard (Figure 4.3). In this panel, we can download the related SAPs or novel junctions database if we are interested in performing a targeted proteomics analysis, or we can look at the threshold information on the experiment (Figure 4.4). We have to comment that this protein has been already found, and it has been flagged with proteomics based evidence in the latest version of neXtProt (May 2015).

CONCLUSIONS

Focusing on the characterization of the proteins with no or inadequate experimental evidence is one of the key points that may lead to advances in the comprehension of our biology and the diagnosis and treatment of diseases. The HPP consortium has concentrated efforts on this but mostly focusing on proteomic approaches in most cases, and this should change. Genomics field goes a step ahead, and all of their available tools and methods should be employed to generate new advances in proteomics. In this work, we have used novel proteogenomic methods that allow us to move in this direction, allowing better protein matches when using peptide search engines. Missing proteins could be expressed only in certain types of tissues; therefore, the more accurate information we have about this will be crucial to progress in the field. A deep analysis of transcriptomics existing data should also be done for this purpose, not only looking for mutations, new transcripts, or gene expression but also measuring the differences in gene expression over the entire set of cell types in the human body, which will facilitate the task of finding those missing proteins in a more accurate way. The work presented here means a step in this direction, having processed a large amount of RNA-Seq and microarrays data that is ready to be used by experimentalists and with the possibility to add more data sets in a very automated way, increasing the available number of studies with different cell types. Our *dasHPPboard* is also ready to be a central place to keep different final reports generated from experiments of the HPP consortium, facilitating their exploration and allowing reaching more global findings. Furthermore, the analysis reported here might encourage other groups of the HPP consortium to suggest new studies, preferably of different cell lines or tissues, and add them to the dashboard, in a collaborative effort that we hope will result in a very complete and unprecedented proteogenomic map of the human.

Most of the data included in our work come from transcriptomics experiments. At the moment we have included proteomics experiments performed within the Spanish HPP consortium, which is focused on Chromosome 16. Including proteomics experiments from all of the groups of the C-HPP consortium is necessary to learn more about the whole proteome, and it would allow us to generate new types of reports in which proteomic and genomic information would be mixed. Experiments coming from other studies like Human Protein Atlas should be also taken into account to complete our panel of transcriptomics and proteomics experiments. We are currently generating tabulated text result files, a versatile solution when exploring and analyzing them, but proteomics experiments can become enormous, with a lot of related fields, and standardized solutions could be a good option to minimize the time spent in understanding the data and the effort required to convert the files to serve as input for other programs. In this direction, as a future work, we are planning to include proteomics results in mzTab[50] format. This is a standard format developed by members of the Proteomics Informatics working group of the HUPO-PSI, which is stored in a simple tabulated way in line with the results we currently display in *dasHPPboard* and will facilitate us to include proteomics results in an easier way because of the increasingly better integration of standards in protein-related programs and the simple conversion between these formats through programs such as PRIDE Converter 2.[51] Also, in this direction, we want to add more metadata to the panel of experiments and to the results of the search section, including information coming from databases like PeptideAtlas[46] or the GPMDB[47] to display a more complete panel, giving them the possibility to easily access to full data sets with raw data.

ASSOCIATED CONTENT

Supporting Information

Supporting Information Table 1: List of the HBM experiments included in the *dasHPPboard*. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00466.

AUTHOR INFORMATION

Corresponding Author

*Tel: +34915854617. E-mail: pascual@cnb.csic.es.

Author Contributions

⊥C.G., F.J.C., J.P.A., and A.P.-M. share senior authorship.

Funding

CIMA, CNB, and UPV/EHU laboratories are member of Proteored, PRB2-ISCIII and are supported by grant PT13/0001 funded by ISCIII and FEDER. The Spanish Minister of Science and Innovation with grant BIO2013−48028-R, the Government of Madrid (CAM) with grant P2010/BMD-2305 and the Children Tumor Foundation also supported this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank IGC and expO for the efforts to procure tissue samples under standard conditions and perform gene expression analyses on a clinically annotated set of deidentified tumor samples. D.T.-M. acknowledges Javier Setoain, Marta Martínez, and Mònica Franch. We dedicate this contribution to Juan Pablo Albar, leader of the Spanish HPP consortium for Chromosome 16, who passed away during the writing of this contribution (http://sphpp.proteored.org/juan_pablo_albar/).

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

ABBREVIATIONS

HPP, Human Proteome Project; SpHPP, Spanish Consortium of the HPP; ENCODE, ENCyclopedia Of DNA Elements; HBM, Human Body Map; HUPO, Human Proteome Organization; C-HPP, Chromosome-Based HPP; B/D-HPP, Biology/Disease HPP; SRM, selected reaction monitoring; CCLE, Cancer Cell Line Encyclopedia; SNV, single nucleotide variant; expO, Expression Project for Oncology; FPKM, fragments per kilobase of transcript per million mapped reads; GEO, gene expression omnibus; AML, acute myeloid leukemia; fRMA, frozen robust multiarray analysis; SAP, single amino acid polymorphism; HUAEC, human umbilical artery endothelial cells; TSV, tab-separated values

REFERENCES

(1) Collins, F. S.; Patrinos, A.; Jordan, E.; Chakravarti, A.; Gesteland, R.; Walters, L. New goals for the U.S. Human Genome Project: 1998−2003. *Science* **1998**, *282*, 682−689.

(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) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10*, M111.009993 DOI: 10.1074/mcp.M111.009993.

(4) Paik, Y.-K.; Jeong, S.-K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H.-J.; Na, K.; Choi, E.-Y.; Yan, F.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30*, 221−223.

(5) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509*, 575−581.

(6) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582−587.

(7) Segura, V.; Medina-Aunon, J. A.; Guruceaga, E.; Gharbi, S. I.; Gonzälez-Tejedo, C.; San Chez Del Pino, M. M.; Canals, F.; Fuentes, M.; Ignacio Casal, J.; Martínez-Bartolomé, S.; et al. Spanish human proteome project: Dissection of chromosome 16. *J. Proteome Res.* **2013**, *12*, 112−122.

(8) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42*, D749.

(9) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; et al. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12*, 293−298.

(10) Paik, Y.-K.; Omenn, G. S.; Thongboonkerd, V.; Marko-Varga, G.; Hancock, W. S. Genome-wide proteomics, Chromosome-Centric Human Proteome Project (C-HPP), part II. *J. Proteome Res.* **2014**, *13*, 1−4.

(11) Goode, R. J. A.; Yu, S.; Kannan, A.; Christiansen, J. H.; Beitz, A.; Hancock, W. S.; Nice, E.; Smith, A. I. The proteome browser web portal. *J. Proteome Res.* **2013**, *12*, 172−178.

(12) Guo, F.; Wang, D.; Liu, Z.; Lu, L.; Zhang, W.; Sun, H.; Zhang, H.; Ma, J.; Wu, S.; Li, N.; et al. CAPER: a chromosome-assembled human proteome browsER. *J. Proteome Res.* **2013**, *12*, 179−186.

(13) Jeong, S.-K.; Lee, H.-J.; Na, K.; Cho, J.-Y.; Lee, M. J.; Kwon, J.-Y.; Kim, H.; Park, Y.-M.; Yoo, J. S.; Hancock, W. S.; et al. GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes. *J. Proteome Res.* **2013**, *12*, 106−111.

(14) Yamasaki, C.; Murakami, K.; Fujii, Y.; Sato, Y.; Harada, E.; Takeda, J.; Taniya, T.; Sakate, R.; Kikugawa, S.; Shimada, M.; et al. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.* **2008**, *36*, D793−D799.

(15) Zgoda, V. G.; Kopylov, A. T.; Tikhonova, O. V.; Moisa, A. A.; Pyndyk, N. V.; Farafonova, T. E.; Novikova, S. E.; Lisitsa, A. V.; Ponomarenko, E. A.; Poverennaya, E. V.; et al. Chromosome 18 Transcriptome Profiling and Targeted Proteome Mapping in Depleted Plasma, Liver Tissue and HepG2 Cells. *J. Proteome Res.* **2013**, *12*, 123−134.

(16) Myers, R. M.; Stamatoyannopoulos, J.; Snyder, M.; Dunham, I.; Hardison, R. C.; Bernstein, B. E.; Gingeras, T. R.; Kent, W. J.; Birney, E.; Wold, B. A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **2011**, *9*, e1001046.

(17) Paik, Y.-K.; Hancock, W. S. Uniting ENCODE with genome-wide proteomics. *Nat. Biotechnol.* **2012**, *30*, 1065−1067.

(18) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603−607.

(19) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Tissue-based map of the human proteome. *Science (Washington, DC, U. S.)* **2015**, *347*, 1260419−1260419.

(20) Gröschel, S.; Sanders, M. A.; Hoogenboezem, R.; De Wit, E.; Bouwman, B. A. M.; Erpelinck, C.; Van Der Velden, V. H. J.; Havermans, M.; Avellino, R.; Van Lom, K.; et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell* **2014**, *157*, 369−381.

(21) Roth, R. B.; Hevezi, P.; Lee, J.; Willhite, D.; Lechner, S. M.; Foster, A. C.; Zlotnik, A. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **2006**, *7*, 67−80.

(22) Pearson, W. R.; Wood, T.; Zhang, Z.; Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **1997**, *46*, 24−36.

(23) Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14*, R36.

(24) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078−2079.

(25) Anders, S.; Pyl, P. T.; Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, btu638.

(26) Harrow, J.; Frankish, A.; Gonzalez, J. M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B. L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760−1774.

(27) Ramsköld, D.; Wang, E. T.; Burge, C. B.; Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **2009**, *5*, e1000598.

(28) Hebenstreit, D.; Fang, M.; Gu, M.; Charoensawan, V.; van Oudenaarden, A.; Teichmann, S. A. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **2011**, *7*, 497.

(29) Van Bakel, H.; Nislow, C.; Blencowe, B. J.; Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **2010**, *8*, e1000371.

(30) Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **2013**, *41*, D991−D995.

(31) Kolesnikov, N.; Hastings, E.; Keays, M.; Melnichuk, O.; Tang, Y. A.; Williams, E.; Dylag, M.; Kurbatova, N.; Brandizi, M.; Burdett, T.;

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749

et al. ArrayExpress update-simplifying data submissions. *Nucleic Acids Res.* **2015**, *43*, gku1057.
(32) Leinonen, R.; Sugawara, H.; Shumway, M. The sequence read archive. *Nucleic Acids Res.* **2011**, *39*, D19−D21.
(33) 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.
(34) McCall, M. N.; Bolstad, B. M.; Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **2010**, *11*, 242−253.
(35) McCall, M. N.; Jaffee, H. A.; Zelisko, S. J.; Sinha, N.; Hooiveld, G.; Irizarry, R. A.; Zilliox, M. J. The Gene Expression Barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.* **2014**, *42*, D938−D943.
(36) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley: New York, 2012.
(37) Guruceaga, E.; Sanchez Del Pino, 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.
(38) McLaren, W.; Pritchard, B.; Rios, D.; Chen, Y.; Flicek, P.; Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **2010**, *26*, 2069−2070.
(39) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **2014**, *13*, 228−240.
(40) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29*, 3235−3237.
(41) Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, *2011*, bar009.
(42) 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.
(43) 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.
(44) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **2012**, *11*, M111.014381 DOI: 10.1074/mcp.M111.014381.
(45) Ghali, F.; Krishna, R.; Lukasse, P.; Martínez-Bartolomé, S.; Reisinger, F.; Hermjakob, H.; Vizcaíno, J. A.; Jones, A. R. Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. *Mol. Cell. Proteomics* **2013**, *12*, 3026−3035.
(46) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34*, D655−D658.
(47) 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.
(48) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32*, 223−226.
(49) Hornbeck, P. V.; Kornhauser, J. M.; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **2012**, *40*, D261−D270.
(50) Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics* **2014**, *13*, 2765−2775.
(51) Côté, R. G.; Griss, J.; Dianes, J. A.; Wang, R.; Wright, J. C.; van den Toorn, H. W. P.; van Breukelen, B.; Heck, A. J. R.; Hulstaert, N.; Martens, L.; et al. The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol. Cell. Proteomics* **2012**, *11*, 1682−1689.

DOI: 10.1021/acs.jproteome.5b00466
J. Proteome Res. 2015, 14, 3738−3749