

Integrated Bioinformatics Analysis of the Publicly Available Protein Data Shows Evidence for 96% of the Human Proteome

Suresh Mathivanan*

Department of Biochemistry, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia

Abstract

Protein-coding genes are predicted by genome annotation pipelines and are conceptually translated into protein sequences. Several thousands of these protein-coding genes catalogued in publicly-available databases seldom have evidence at the protein level. In this study, we have created a map of the human proteome by integrating publicly-available proteomic studies and resources. With the encompassed data, we are able to map 96% of the human proteome with ample experimental evidence for protein expression. Over 2.2 million annotations are recorded for 19,716 proteins from 63,239 independent studies that utilized more than 800 tissue/cell types/body fluids. Among the mapped human proteome, 96% of the protein expression is supported by two or more independent studies or experimental methods. The collated data (localization, tissue expression, post-translational modifications, protein-protein interactions, enzymes-substrate and 3D structures) is freely accessible through the web-based compendium Human Proteome Browser (<http://www.humanproteomebrowser.info>).

Keywords: Protein-coding genes; Human proteome; Genome annotation; Proteomic databases

Abbreviations: PTMs: Post-Translational Modifications; HGP: Human Genome Project; ESTs: Expressed Sequence Tags; HPP: Human Proteome Project; IHC: Immunohistochemistry; ORF: Open Reading Frame; TM: Transmembrane; HPB: Human Proteome Browser

Introduction

With the completion of the human genome project (HGP) [1,2], attention has now shifted to annotating the estimated 20,000 to 25,000 protein-coding genes [3,4]. The task of fully annotating the human genome is multifaceted involving first deciphering protein-coding regions of the euchromatin, followed by developing the definitive catalogue of encoded proteins (the proteome) and, finally, tying together literature information related to protein abundances, post-translational modifications (PTMs) and interacting partners in order to make the genome data useful to the broader biological research community. To this end, genome annotation pipelines have been developed and gene sequences (protein-coding and otherwise) are continuously catalogued in publicly-available sequence databases [5]. Eventually, all of these genes are conceptually-translated into proteins and are compiled in protein sequence databases. Though many genes have evidence of existence at the transcript level (mRNA, ESTs), relatively few have been identified at the protein level. On the other hand, some have neither transcript nor protein evidence and are merely predicted models. For example, Uni Prot [6], a protein sequence centric database, has protein proof only for 14% of its conceptually-translated protein sequences (at the time of writing). Another 13% have evidence at the transcript level only while a staggering 73% of the protein sequences have neither (<http://au.expasy.org/sprot/relnotes/relstat.html>). However, in the manually curated human Uni Prot KB/Swiss-Prot [6], the evidence of a gene at protein level is higher (~66%) but still not comprehensive.

Genome sequencing rates have drastically increased over the years and more than 3,800 organisms have already been sequenced [7]. In contrast, mapping and functionally annotating even one proteome is no trivial task [8,9]. We are yet to map the complete list of genes that are expressed at the protein level even for one organism [9,10]. Ten

years on from the completion of HGP, there is still debate over the exact number of the estimated 20,000-25,000 protein-coding genes [3]. Unlike the finite genome, the proteome is very complex both temporally and spatially within a cell, depending upon physiological stimuli, epigenetic status, post-transcriptional and post-translational events [10,11]. Hence, functional characterization of each and every human protein based on their subcellular localization, tissue expression, isoforms, protein interaction partners, PTMs and 3D structures is indeed a daunting task.

To address this issue, an internationally-coordinated Human Proteome Project (HPP) was officially launched at the 9th Annual HUPO World Congress in September 2010 (<http://hupo.org/research/hpp>) to systematically characterize human proteins across various tissues and decipher their interaction partners by multiple proteomic methods [12]. Proteomic researchers have already initiated projects to catalog gene products at the protein level and to map the human proteome to its entirety [12-17]. One of the targets of the HPP is to develop a definitive catalogue of human proteins that can be experimentally identified at the protein level in various tissues/cell lines. In this endeavour, it is timely to catalogue extant protein identifications already present in the public domain – that is, to establish a reference starting point against which the HPP can be assessed. Decades of hypothesis-driven research has already been performed on human biological tissues and cell lines. While a bulk of these data are presently scattered across various publicly-available databases including BioGRID [18], Intact [19], Human Proteinpedia [20], HPRD [21], PRIDE [22], Peptide Atlas

*Corresponding author: Suresh Mathivanan, Department of Biochemistry, La Trobe Institute for Molecular Science, La Trobe University, Bundoora, Victoria 3086, Australia, Tel: +61 03 9479 2565; Fax: +61 03 9479 1226; E-mail: S.Mathivanan@latrobe.edu.au

Received December 11, 2013; Accepted February 17, 2014; Published February 20, 2014

Citation: Mathivanan S (2014) Integrated Bioinformatics Analysis of the Publicly Available Protein Data Shows Evidence for 96% of the Human Proteome. J Proteomics Bioinform 7: 041-049. doi:10.4172/jpb.1000301

Copyright: © 2014 Mathivanan S. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

[23], Peptidome [24], HPA [25] Plasma Proteome Database [26], next Prot [27] and GPMdb [28], a substantial amount is hidden, and not queryable, in the supplementary tables of journal websites [20].

Here, we have created a list of 20,515 non-redundant human protein-coding genes (target proteome) by merging manually curated Uni Prot and Ref Seq [29] protein entries. Further, we collated extant human protein identifications to create the first draft of the human proteome. A three-pronged approach was used to accomplish the task: -I, integrating publicly-available proteomic databases, -II, collating published proteomic studies, and -III, undertaking protein-centric manual curation of scientific literature. Employing these strategies, we have assembled 2.2 million protein annotations from >800 tissues/cell types/body fluids and mapped 96% of the human proteome (compared against the target proteome). Among them, 96% were supported by two or more independent studies or experimental methods (e.g., mass spectrometry (MS), immunohistochemistry (IHC)) confirming the expression of the protein. Additionally, we were able to provide functional annotations for 19,716 proteins based on subcellular localization, tissue and cell line expression, PTMs, protein-protein interactions, enzymes-substrate relations and 3-D structures.

Methods

Creation of the human target proteome

To create a human proteome reference set, Ref Seq and Uni Prot human proteome sequences were combined together and the sequence accession numbers were mapped to Entrez Gene identifiers.

NP entries from Ref Seq human proteome database were mapped to Entrez Gene identifiers resulting in 19,129 unique genes. Similarly, Uni Prot human complete proteome was mapped to Entrez Gene identifiers resulting in 20,091 unique genes. However, 226 Uni Prot accession identifiers could not be mapped to Entrez Gene and were still added to the reference set. In total, combining Ref Seq and Uni Prot protein entries, the human target proteome set encompassed 20,515 unique genes.

Integration of proteomic databases

To integrate human protein data that is already available in proteomic databases, publicly accessible proteomic databases were used. Datasets were downloaded from various protein-protein interaction databases (BioGRID [18], Intact [19], Human Proteinpedia [20] and HPRD [21]) and mapped to Entrez Gene or Uni Prot accession identifiers. Customized Perl scripts were written to parse the respective datasets and only human proteins were retained in the final list. Similarly, protein structure (PDB [30]), mass spectrometry (PRIDE [22], Peptide Atlas [23], Peptidome [24] and Human Proteinpedia), immunohistochemistry (HPA [25] and Human Proteinpedia), protein annotation (HPRD, Entrez Gene and UniProt), exosome (Exo Carta [31], Vesiclepedia [32]), colorectal cancer (Colorectal Cancer Database), plasma (Plasma Proteome Database [33]), post-translational modification (Phosphosite Plus [34], HPRD, Human Proteinpedia and UniProt) databases were used to download the protein annotations and parsed with customized Perl scripts. As no two databases had the download files in same format or same accession identifiers, Perl scripts were customized to every single proteomic database file. The protein lists were collated and integrated into Human Proteome Browser.

Database cross reference file

Database cross references files were downloaded from NCBI Entrez

Gene, Ref Seq, UniProt, IPI, Ensembl and GenBank. Protein accession identifiers were mapped to Entrez Gene identifiers and Uni Prot entries (for 226 entries that could be mapped to Entrez Gene ids). Unmapped protein identifiers were mapped again to Entrez Gene identifiers using DAVID and BioMart tools. The cross database mappings resulted in a master mapping file.

Data inclusion criteria

Various data inclusion criteria employed in Human Proteome Browser are as follows:

1. Experiment duplicates in various databases (based on the experimental accession number) were not considered.
2. Bioinformatics predictions were removed and protein identifications should be based on an experiment.
3. Orthologous protein identifications fixed to human proteins were not considered.
4. MS based subcellular localization datasets were used for protein identifications but were not used for subcellular localization unless another orthogonal method is used in the same study to prove the localization.
5. PTMs without amino acid sites were not considered.

Transcriptome

Microarray data for 79 human tissues was downloaded from BioGPS [35]. Averaged mRNA expression values were used to perform hierarchical clustering for genes unidentified at the protein level. Customized R scripts were written to generate heat maps.

Proteotypic peptides

Peptide sequences from proteomic databases and published proteomic studies were searched using BLAST against human non-redundant database. Peptide sequences that uniquely identify a human protein were retrieved and were used in our analysis.

Results

Creation of a list of human protein-coding genes

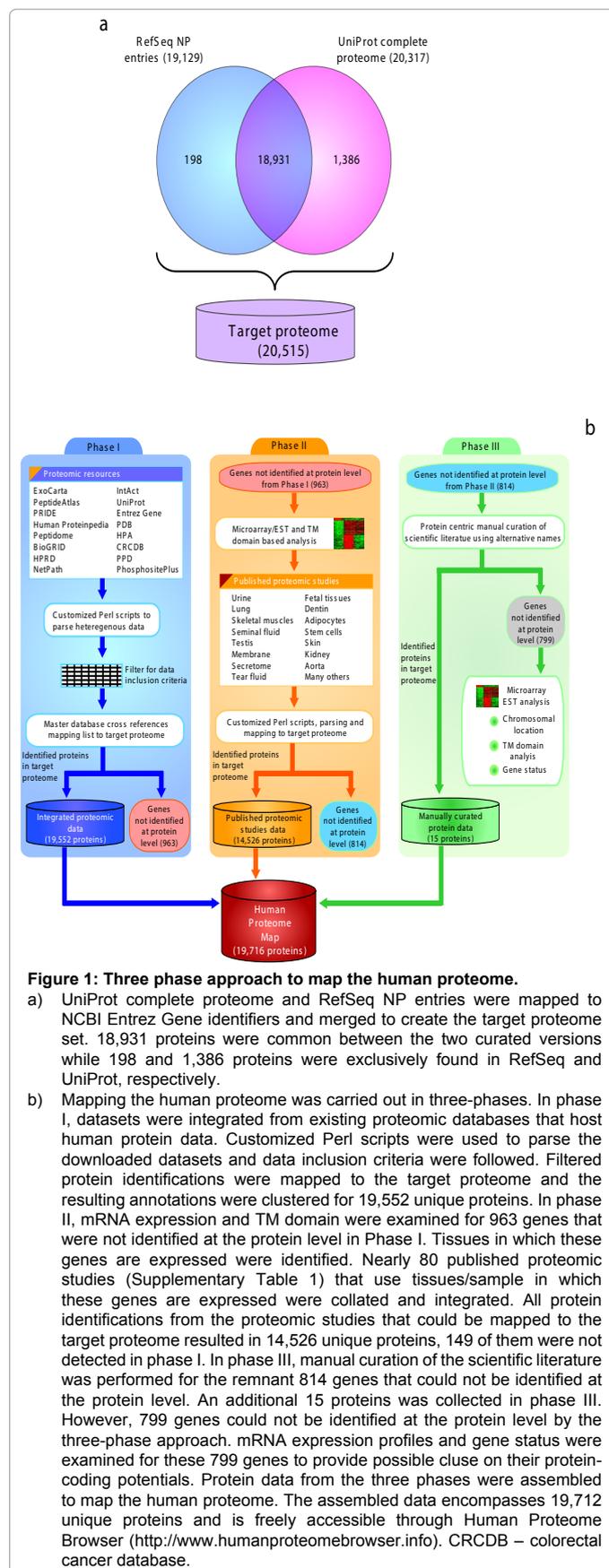
It is not a trivial task to predict genes accurately from genomic data [36]. In *Mycoplasma genitalium* genome, 8% of the 340 genes were considered to be incorrect [37]. When the error rate is extrapolated for complex eukaryotic organisms with the consideration of intron-exon junctions, a significant number of gene predictions can be incorrect. Identification of an open reading frame (ORF) in genomic data does not always imply the existence of a protein-coding gene [36]. Apart from protein-coding genes, genome annotation pipelines predict a variety of gene types such as, pseudogenes, tRNA, rRNA, snRNA, scRNA, miscRNA and ncRNA. Irrespective of their protein-coding efficiency, all the gene types are conceptually-translated into corresponding protein sequences and stored in databases such as TrEMBL [38] (translation of EMBL nucleotide sequence database) resulting in several erroneous proteins [39,40]. Additionally, redundancy is also a concern [41] (TrEMBL currently hosts 76,802 human protein sequences). To provide reliable non-redundant protein-coding sequences, databases such as Ref Seq and Uni Prot adopted manual curation and, as a result, data is curated by highly trained biologists [40]. In this current study, to map the human proteome with extant protein annotations, a target set of non-redundant protein-coding genes that is curated by expert

scientists is needed. Because Ref Seq and Uni Prot use different genome annotation criteria, these curated databases differ, to a significant extent, in their proteins [42]. For example, histone H3.1 is represented by 10 distinct genes (10 unique gene symbols) in Ref Seq (protein sequences are not always 100% identical) but only one entry (P68431) in UniProt. Similarly, human endogenous retroviral-K115 envelope protein (UniProt:Q902F8) is not found in Ref Seq. For this reason, we combined both Ref Seq and Uni Prot protein accession identifiers to create an integrated, non-redundant, list of human protein-coding genes – referred to as the ‘target proteome’. To accomplish this, human complete proteome dataset from Uni Prot and NP entries from Ref Seq were mapped to NCBI Entrez Gene [43] identifiers resulting in 20,515 unique protein-coding genes (226 Uni Prot entries not mapped to Entrez Gene identifiers were retained in the target proteome). In the target proteome set (20,515 genes), 18,931 genes are common to both the curated versions while 198 and 1,386 genes are unique to Ref Seq and UniProt, respectively (Figure 1a).

Mapping the human proteome by three-phase strategy

To interrogate the extant literature for evidence of at least one protein product for every gene in the target proteome set, we used a three-phase approach -I, integrating publicly-available proteomic databases, -II, collating published proteomic studies, and -III, protein-centric manual curation of scientific literature (Figure 1b). In phase I, we integrated multiple proteomic databases as every database has a specific mandate and host data on particular features. PRIDE [22], Peptide Atlas [23] and Peptidome [24] are MS-centric while Bio GRID [18] and IntAct [19] are interaction specific databases. Currently, no compendium integrates heterogeneous data from various proteomic resources and presents it to the biomedical users. To integrate heterogeneous protein annotations, datasets were downloaded and parsed from a variety of proteomic resources including Exo Carta [44], PRIDE [22], Peptide Atlas [23], Human Proteinpedia [45], Peptidome [24], BioGRID, Human Protein Reference Database [46], Int Act [19], Uni Prot [6], Entrez Gene [43], PDB [30], Human Protein Atlas [25], Colorectal Cancer Database (Mathivanan, In preparation), Plasma Proteome Database [33] and Phosphosite Plus [34]. Redundant studies that are available in multiple proteomic databases were included only once, if they could be clearly separated. For example, Peptidome hosts MS experiments that are retrieved from Peptide Atlas and such studies were included once. As the download file formats, protein features annotated and the protein accession identifiers varied for each of the databases, individualized Perl scripts were used to parse the downloaded datasets. Protein accession identifiers were mapped to the target proteome by using a master database accession cross reference file that was created as part of this study. Additionally, tissue, localization and experimental method vocabularies were matched to community ontologies (eVOC [47], Gene Ontology [48] and PSI-MI [49]) and standardized terms were fixed for each entry. The collated protein annotations include subcellular localization, tissue and cell line expression, PTMs, protein-protein interactions, enzymes-substrate relations and 3-D structures. Notably, the dataset assembled here is obtained from multiple proteomic experimental methods and is not biased to any one particular method (e.g., MS or IHC). In phase I, by integrating proteomic databases, we identified 95% (19,552/20,515) of the target proteome while 963 (5%) proteins were not detected in any of the proteomic databases.

In phase II, we examined gene expression patterns of 963 genes (not identified in phase I) by downloading and examining mRNA



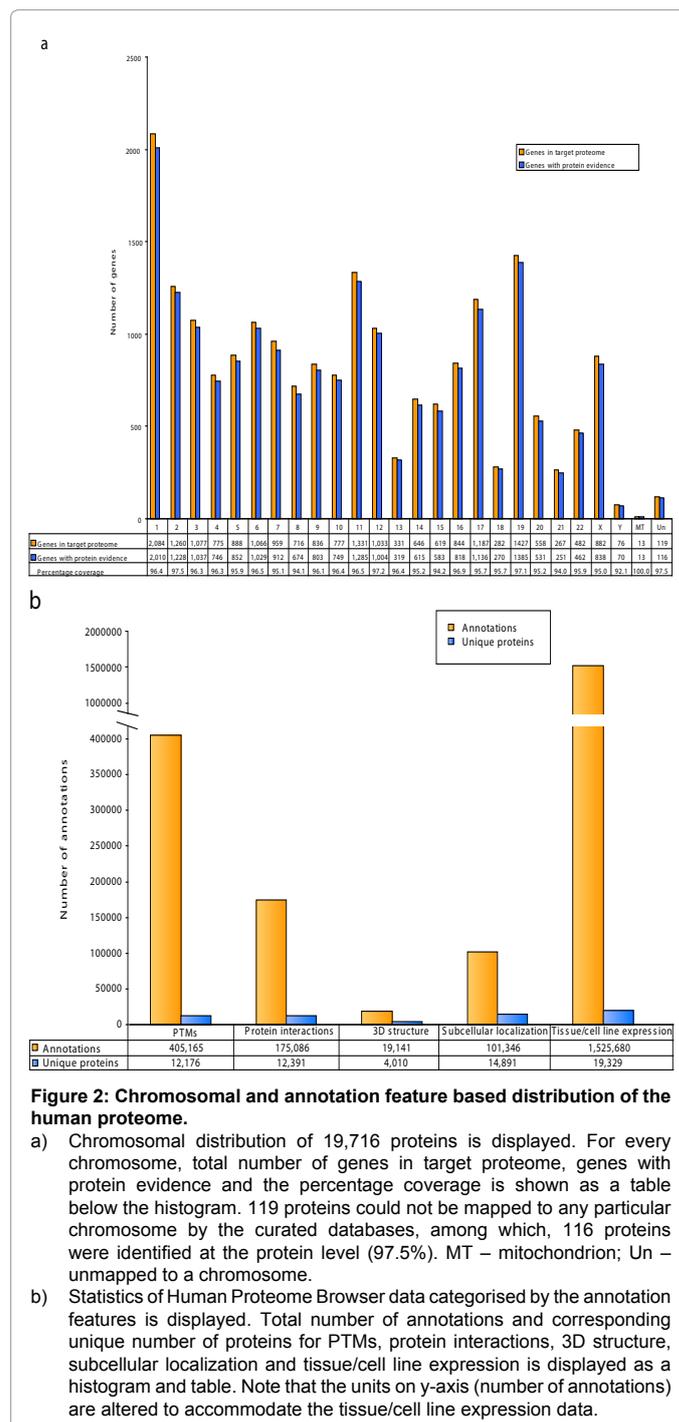
microarray-based expression profiles in 79 human tissues [35,50] and unigene-based EST expression datasets (<http://www.ncbi.nlm.nih.gov/UniGene>). The analysis revealed several tissues (fetal, skeletal muscles, testis, adipocytes, skin, kidney, developmental, lung and aorta) in which some of the 963 mRNA are expressed (data not shown). Proteomic experiments using some (e.g., testis, developmental tissues and skeletal muscles) of the identified tissues are minimal because of ethical reasons. Additionally, transmembrane (TM) domain based analysis confirmed the presence of several (24%) membrane proteins, a protein class that is often under-represented in proteomic studies due to low abundance and hydrophobicity [51]. To capture the 963 genes not identified in phase I, proteomic experiments that used tissues in which some of the 963 genes were expressed and membrane enrichment studies were targeted. As a result, we collated nearly 80 proteomic studies (Supplementary Table 1) that were performed on testis, membrane molecules, skeletal muscles, cancer secretomes and bodily fluids. Supplementary and manuscript inline tables (few cases) were downloaded/copied and customized Perl scripts were used to parse these datasets. In a few instances, when data are not provided as supplementary information or when provided as a PDF file (which renders information retrieval harder), we approached the authors to obtain the protein identifications. Overall, the approach in phase II resulted in the identification of 14,526 proteins among which 149 were not detected in phase I.

A total of 814 proteins in the target proteome were not detected in phase I and II. In phase III, we performed manual curation of the scientific literature for the remaining 814 unidentified proteins by searching with available alternative names obtained from Entrez Gene and UniProt. Among these, 15 could be identified at the protein level by manual curation. Collectively, the three-phase strategy resulted in the identification of 19,716 proteins (i.e., 96% of the target proteome – Figure 1b). Chromosomal coverage of the mapped 19,716 genes with evidence at the protein level ranged from 92-97.5% (Figure 2a), the lowest in chromosome Y (70/76 proteins were detected – 92%).

Human Proteome Browser a compendium hosting heterogeneous proteomic data

Proteomic data integrated as part of this analysis can be accessed freely through the compendium Human Proteome Browser (HPB) (<http://www.humanproteomebrowser.info>). Overall, 2.27 million annotations from 63,296 independent studies using 110 experimental methods are currently available in HPB (Table 1). As shown in Figure 2b, 405,165 annotations (12,176 unique proteins) are recorded for PTMs, 175,086 annotations (12,391 unique proteins) for protein interactions, 19,141 annotations (4,010 unique proteins) with 3D structures, 101,346 annotations (14,891 unique proteins) with subcellular localization and more than 1.5 million annotations (19,329 unique proteins) with tissue/cell line expression. Notably, 98% of the mapped human proteome had at least one tissue/cell line expression. Next, we examined which experimental method covered most of the human proteome. The most number of protein annotations (18,985 unique proteins) were recorded from MS (Figure 3a) where in 93% of the human proteome could be captured because of its high-throughput nature [52]. Among the 18,985 proteins identified by MS, 18,646 (98%) proteins are detected by 2 or more peptides and/or supported by more than one study or experimental method emphasizing the high data quality. Notably, only 339 proteins were detected by single peptide and not supported by another study or experimental method. IHC had the second highest number of annotations largely derived from

the Human Protein Atlas [25]. Among protein interaction methods, yeast-two hybrid based assay identified 8,710 proteins primarily due to proteome-wide interaction studies [53,54]. We further checked for the total number of proteins that are identified in cancerous tissues or cell lines. The cancer proteome assembled as part of this study includes 11,918 unique proteins (Figure 3b). Most number of proteins (10,643) is detected in ovarian cancer models followed by colorectal cancer (9,031). We next checked for the most often identified protein from decades of proteomics experiments. Albumin (ALB) surfaced as the most widely identified protein from 947 experiments followed by



| | Data feature | Number |
|----|--|--------------|
| 1 | Protein annotations | 2.27 million |
| 2 | Experiments | 63,296 |
| 3 | Unique proteins | 19,716 |
| 4 | Percentage of human proteome coverage | 96% |
| 5 | Normal tissue/cell types/body fluids | 704 |
| 6 | Cell lines | 85 |
| 7 | Cancer tissues | 18 |
| 8 | Experimental methods | 110 |
| 9 | Proteins with subcellular localization | 14,885 |
| 10 | Proteins with tissue/cell line expression | 19,329 |
| 11 | Proteins with post-translational modifications | 12,176 |
| 12 | Proteins with 3D structures | 4,010 |
| 13 | Proteins with interactions | 12,391 |
| 14 | Proteins identified from all cancer tissues/cell lines | 11,918 |

Table 1: Statistics of data integrated in Human Proteome Browser.

serpin peptidase inhibitor, clade A (SERPINA1) with 771 experiments. Overall, HPB integrates heterogeneous data under one single resource and biomedical users can access various features of their protein of interest in one website (Figure 4).

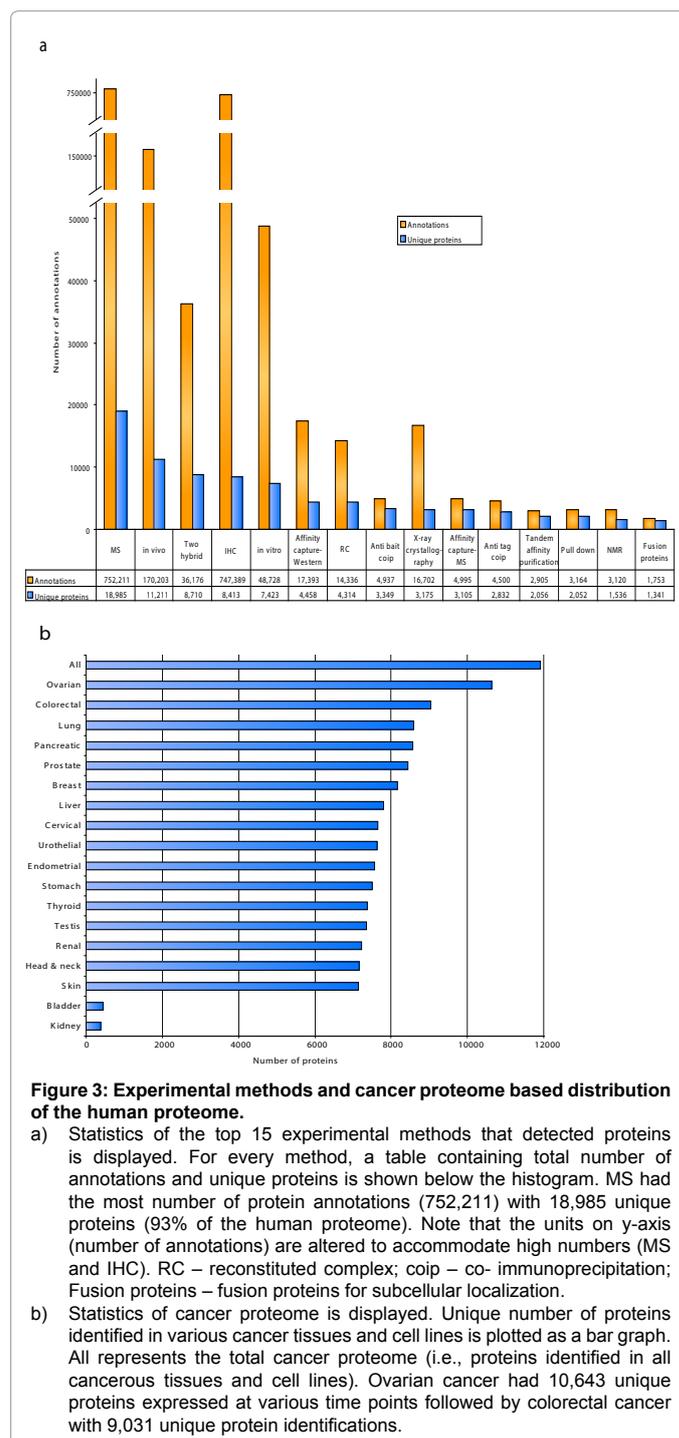
Data inclusion criteria and data quality

Even though some of the MS-based integrated datasets had a false discovery rate, it is hard to assign a false discovery rate for the integrated data as the collated data arises from multiple experiments including immunohistochemistry, Western blotting, confocal microscopy and MS. So the data inclusion criteria employed in our integration is critical to the data quality. To emphasise on high data quality, we had adopted various data inclusion criteria. First and foremost, only human protein data was included. Protein identifications mapped to human orthologs and any form of bioinformatics predictions was omitted. For example, Gene ontology based subcellular localization profiles were downloaded from Entrez Gene and Uni Prot for proteins with experimental codes (e.g., IDA - inferred from direct assay; IPI - inferred from physical interaction; TAS - traceable author statement) and bioinformatics predictions (e.g., ISO - inferred from sequence orthology; IEA - inferred from electronic annotation) were discarded. Similarly, from PDB, 3D structures obtained by NMR and X-ray crystallography were retained while model-based predictions were removed. Similarly, MS based subcellular localization studies were used for tissue/cell line expression but were not considered for localization annotations due to the possible contaminations during fractionation procedures [55]. MS-based subcellular localization annotations were included when the data is validated by orthogonal experiments in the same study. By integrating multiple proteomic databases and studies, the confidence level of the protein identifications has eventually increased. One of the drawbacks of collating heterogeneous datasets is the increase in the individual false positives identifications. Nevertheless, in cases where multiple experiment platforms and independent studies support the protein identification, the integration of multiple resources can be justified. Based on this, 96% of the mapped human proteins were detected by two or more experiments/studies confirming the identification of the protein (Figure 5a). Eight five percentage of the proteins were detected by 5 or more experiments. A staggering 69% of the proteins are detected by 10 or more experiments again emphasising the reliability of the assembled data.

Majority of the proteins that were not identified were also not detected in the transcriptome dataset

In spite of our three-pronged approach to map the human

proteome, 799 genes from the target proteome could not be detected at the protein level (Supplementary Table 2). No chromosomal location bias existed for the unidentified proteins (Supplementary Figure 1a). TM domain based analysis revealed that 22% of them have at least one TM domain (Supplementary Figure 1b). To check how many of the 799 genes have evidence at the transcript level, we examined the mRNA expression profile of these genes by using the downloaded 79 human tissue microarray dataset and UniProt-based transcript evidence. Interestingly, only 210 (26%) genes had expression at the transcript level while 74% (589/799) did not (Supplementary Figure 2). To check



for the protein-coding potentials, gene types of these curated genes were examined (Figure 5b). Among the 799 unidentified genes, 37% (297) are protein-coding with validated/provisional status, 24% (195) are protein-coding predicted models, 15% (116) are pseudogenes, 11% (89) are miscRNA, 11% (89) are uncertain/unknown models and 2% (13) are others. These results reveal that a majority of the unidentified genes potentially could be invalid protein-coding genes. It is possible that some of the valid protein-coding genes could be hidden in the scientific literature either in the main text or in the supplementary tables, which can't be queried. Perhaps some of them could be expressed only at certain stages in development or in specific tissues at certain time points. Further analysis need to be performed in order to add protein evidence to these molecules or retract them from the curated protein-coding list.

Beyond the target human proteome set

While a majority of the protein identifications from various datasets that were collated could be mapped to the target proteome, a minor set could not be mapped. Such unmapped protein entries can be categorized into two groups, -1, proteins that are deemed invalid and withdrawn from databases due to genome annotation pipelines, and -2, proteins that exist in non-curated databases but are not part of the curated versions. We analysed such unmapped proteins that do exist in non-curated databases but not in curated versions. 1,308 unique genes could be mapped from such entries (accessible through HPB). Among these, 1,208 were identified by MS and 671 (51%) are characterized as pseudogenes. Presumably, some of the peptides identified by MS could

be identical to other valid protein-coding gene products. In order to check for bona fide protein identifications, we examined whether any of these proteins have been identified by proteotypic peptides (unique peptides that spans only one specific gene products). Interestingly, 297 proteins could be identified by proteotypic peptides. This clearly suggests that some of these protein identifications can be valid and the curation projects fail to catalogue them as bona fide protein-coding genes. For example, T cell antigen receptor alpha (TRA@) is detected by IHC, MS (5 independent studies), Western blotting, X-ray crystallography, in vivo and in vitro methods. Additionally, the membrane protein is detected by a proteotypic peptide (SDSYGYLLLQELQMK). Though some of these protein identifications can be valid, we emphasize caution in using these protein identifications as further experimental validations/manual curation need to be performed to confirm their protein-coding potentials.

Discussion

Pitfalls associated with large scale integration

Whilst the strength of the study is the integration of protein data obtained from multiple resources and heterogeneous experiment platforms, the weakness of the study is also the same where the assembled datasets can have false positives. Though the protein identification can be strengthened by multiple evidences, it cannot be ignored that individual false positive identifications still exist similar to other databases. In this study, we have mapped 96% of the human proteome and collated experimental evidence for the existence of 19,716

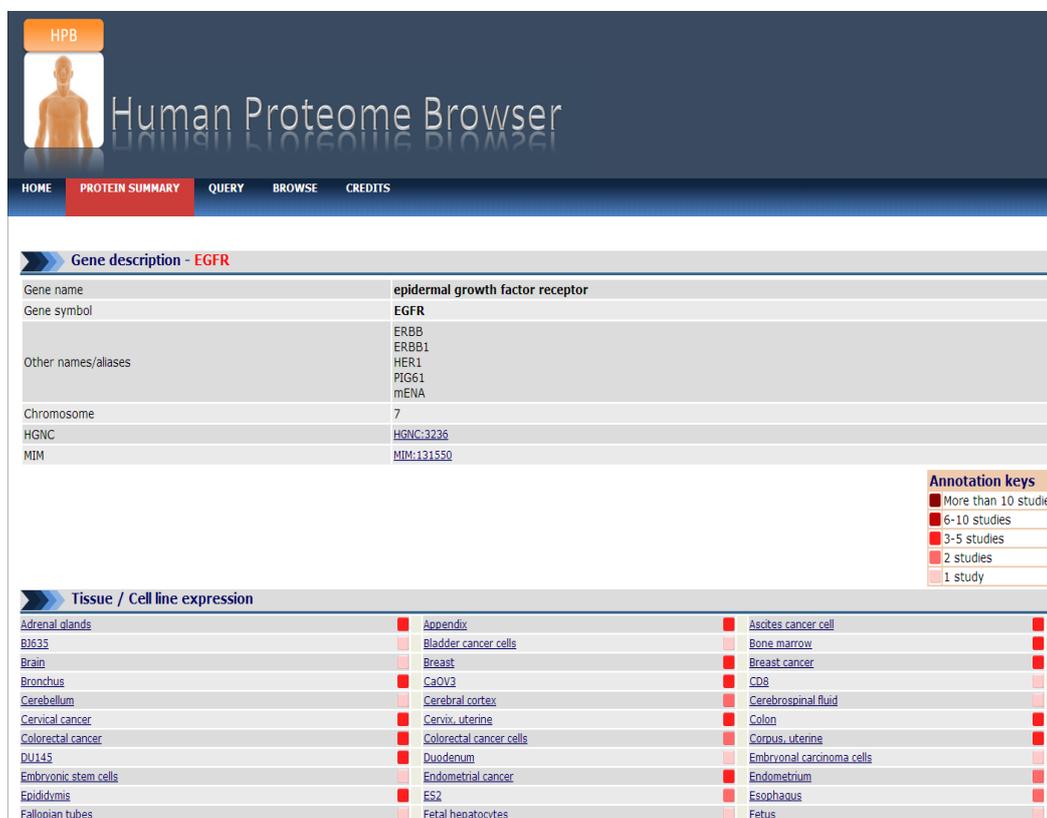


Figure 4: A snapshot of Human Proteome Browser protein page. EGFR protein page with tissue/cell line expression displayed from Human Proteome Browser. Biomedical users can browse heterogeneous protein annotations in one website, the Human Proteome Browser.

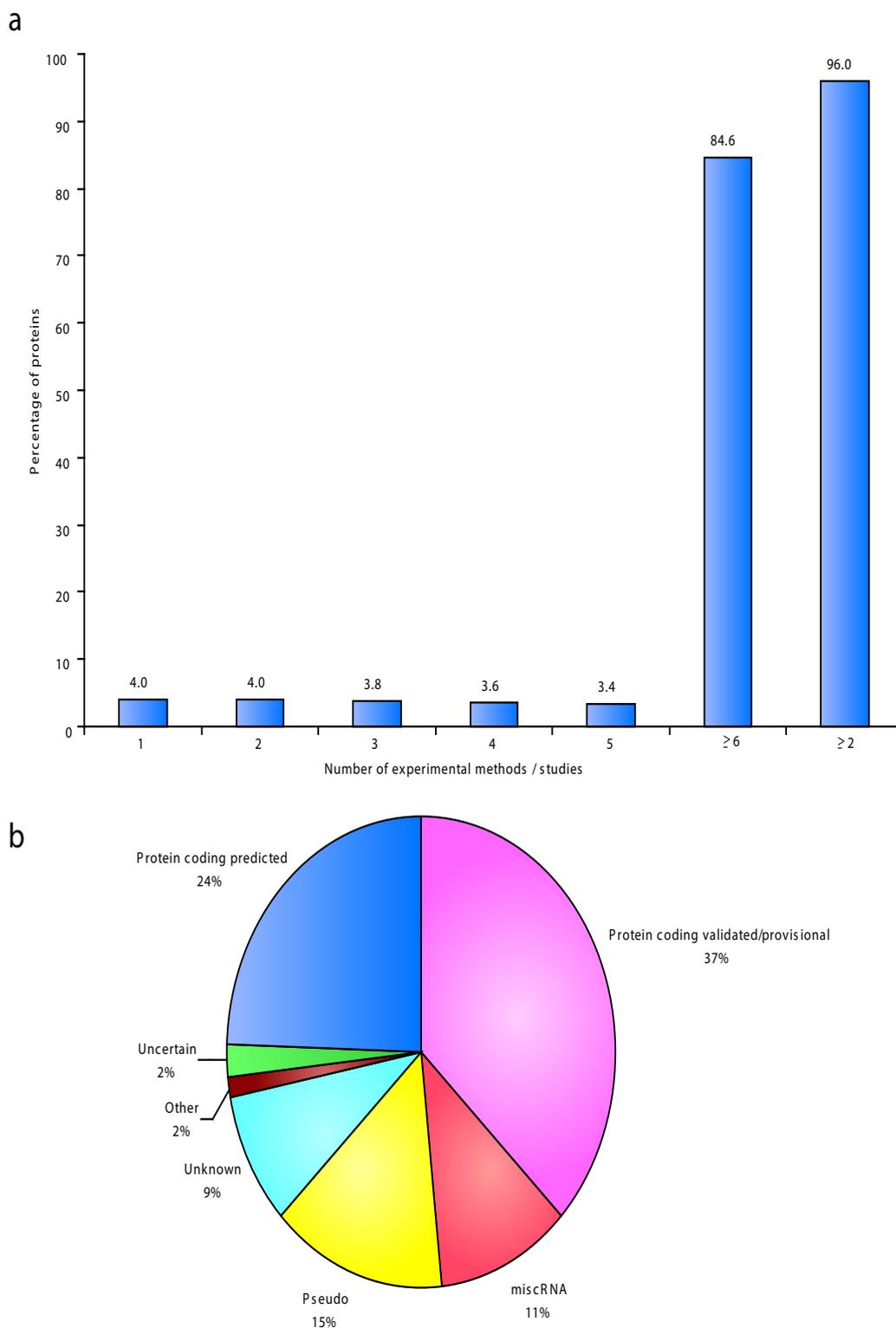


Figure 5: Distribution of proteins by number of experimental methods/studies and gene status.

- a) Statistics of protein identifications in Human Proteome Browser by the total number of experimental methods or studies is displayed. Only, 4% of the proteins were identified by one experiment or study. A staggering 96% of the proteins were identified by 2 or more independent studies or methods while 84.6% of the proteins were identified by 6 or more.
- b) Genes status distribution for 799 genes that could not be identified at the protein level is shown. 37% of the genes were denoted as protein-coding models with either provisional or validated RefSeq entries. 24% were predicted protein-coding models. The remaining 39% of the genes were denoted as pseudo, miscRNA, unknown, uncertain and others by the genome annotation pipelines.

genes at the protein level. We have indexed human proteins based on their subcellular localization, tissue/cell line expression, PTMs, protein interactions, 3D structures and enzyme-substrate relations. We believe that this study will be the one of the many that is aimed at deciphering the human proteome at various functional levels. HPP aims to characterize and quantitate every single protein in the human proteome. There are still a multitude of features yet to be unravelled in order to functionally characterize the entire human proteome. Protein isoforms, interaction partners for 8,000 proteins, protein abundances in various tissue and cell lines and finite number of protein-coding genes are yet to be documented.

Acknowledgement

This work was supported by the Australian National Health and Medical Research Council fellowship (1016599) and Australian Research Council Discovery Project Grant (DP130100535) to SM. The funders had no role in study design, data collection and analysis, decision to publish and preparation of the manuscript. We thank Prof. Richard J Simpson for minor editing of the introduction.

References

- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
- International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.
- Gregory SG, Barlow KF, McLay KE, Kaul R, Swarbreck D, et al. (2006) The DNA sequence and biological annotation of human chromosome 1. *Nature* 441: 315-321.
- Stein L (2001) Genome annotation: from sequence to biology. *Nat Rev Genet* 2: 493-503.
- UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 38: D142-148.
- Venter JC (2010) Multiple personal genomes await. *Nature* 464: 676-677.
- Orchard S, Hermjakob H, Apweiler R (2005) Annotating the human proteome. *Mol Cell Proteomics* 4: 435-440.
- Rabilloud T, Hochstrasser D, Simpson RJ (2010) Is a gene-centric human proteome project the best way for proteomics to serve biology? *Proteomics* 10: 3067-3072.
- Kuster B, Schirle M, Mallick P, Aebersold R (2005) Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* 6: 577-583.
- Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457-463.
- Pearson H (2008) Biologists initiate plan to map human proteome. *Nature* 452: 920-921.
- (2010) The call of the human proteome. *Nat Methods* 7: 661.
- Cyranoski D (2010) China pushes for the proteome. *Nature* 467: 380.
- Liu S, Im H, Bairoch A, Cristofanilli M, Chen R, et al. (2013) A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J Proteome Res* 12: 45-57.
- Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, et al. (2013) The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J Proteome Res* 12: 23-27.
- Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, et al. (2012) The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol* 30: 221-223.
- Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39: D698-704.
- Aranda B, Achuthan P, Alam-Farouque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 38: D525-531.
- Mathivanan S, Ahmed M, Ahn NG, Alexandre H, Amanchy R, et al. (2008) Human Proteinpedia enables sharing of human protein data. *Nat Biotechnol* 26: 164-167.
- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res* 37: D767-D772.
- Vizcaino JA, Côté R, Reisinger F, Foster JM, Mueller M, et al. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 9: 4276-4283.
- Deutsch EW (2010) The PeptideAtlas Project. *Methods Mol Biol* 604: 285-296.
- Slotta DJ, Barrett T, Edgar R (2009) NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nat Biotechnol* 27: 600-601.
- Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, et al. (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28: 1248-1250.
- Nanjappa V, Thomas JK, Marimuthu A, Muthusamy B, Radhakrishnan A, et al. (2014) Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res* 42: D959-D965.
- Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, et al. (2013) neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res* 12: 293-298.
- Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 3: 1234-1242.
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61-D65.
- Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, et al. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res* 34: D302-305.
- Simpson RJ, Kalra H, Mathivanan S (2012) ExoCarta as a resource for exosomal research. *J Extracell Vesicles* 1.
- Kalra H, Simpson RJ, Ji H, Aikawa E, Altevogt P, et al. (2012) Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol* 10: e1001450.
- Muthusamy B, Hanumanth G, Suresh S, Rekha B, Srinivas D, et al. (2005) Plasma Proteome Database as a resource for proteomics research. *Proteomics* 5: 3531-3536.
- Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4: 1551-1561.
- Wu C, Orozco C, Boyer J, Leglise M, Goodale J, et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 10: R130.
- Pandey A, Mann M (2000) Proteomics to study genes and genomes. *Nature* 405: 837-846.
- Brenner SE (1999) Errors in genome annotation. *Trends Genet* 15: 132-133.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31: 365-370.
- Salzberg SL (2007) Genome re-annotation: a wiki solution? *Genome Biol* 8: 102.
- Reeves GA, Talavera D, Thornton JM (2009) Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface* 6: 129-147.
- O'Donovan C, Martin MJ, Glemet E, Codani JJ, Apweiler R (1999) Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics* 15: 258-259.
- Descorps-Declère S, Barba M, Labeledan B (2008) Matching curated genome databases: a non trivial task. *BMC Genomics* 9: 501.
- Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 35: D26-31.

44. Mathivanan S, Simpson RJ (2009) ExoCarta: A compendium of exosomal proteins and RNA. *Proteomics* 9: 4997-5000.
45. Mathivanan S, Pandey A (2008) Human Proteinpedia as a resource for clinical proteomics. *Mol Cell Proteomics* 7: 2038-2047.
46. Mishra GR, Mathivanan S, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database--2006 update. *Nucleic Acids Res* 34: D411-D414.
47. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 13: 1222-1230.
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
49. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol* 22: 177-183.
50. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101: 6062-6067.
51. Wu CC, Yates JR 3rd (2003) The application of mass spectrometry to membrane proteomics. *Nat Biotechnol* 21: 262-267.
52. Nilsson T, Mann M, Aebersold R, Yates JR 3rd, Bairoch A, et al. (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* 7: 681-685.
53. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437: 1173-1178.
54. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, et al. (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125: 801-814.
55. Lee YH, Tan HT, Chung MC (2010) Subcellular fractionation methods and strategies for proteomics. *Proteomics* 10: 3935-3956.