Published online 23 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D773–D781***
doi:10.1093/nar/gkn701

Human Proteinpedia: a unified discovery resource for proteomics research

Kumaran Kandasamy[1,2,3], Shivakumar Keerthikumar[1,2], Renu Goel[1], Suresh Mathivanan[1,2], Nandini Patankar[1], Beema Shafreen[1], Santosh Renuse[1], Harsh Pawar[1], Y. L. Ramachandra[2], Pradip Kumar Acharya[1], Prathibha Ranganathan[1], Raghothama Chaerkady[1,3], T. S. Keshava Prasad[1] and Akhilesh Pandey[3,*]

[1]Institute of Bioinformatics, International Tech Park, Bangalore 560 066, [2]Department of Biotechnology, Kuvempu University, Shankaraghatta, Karnataka, India and [3]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University, Baltimore, MD 21205, USA

Received August 16, 2008; Revised September 24, 2008; Accepted September 26, 2008

ABSTRACT

Sharing proteomic data with the biomedical community through a unified proteomic resource, especially in the context of individual proteins, is a challenging prospect. We have developed a community portal, designated as *Human Proteinpedia* (http://www.humanproteinpedia.org/), for sharing both unpublished and published human proteomic data through the use of a distributed annotation system designed specifically for this purpose. This system allows laboratories to contribute and maintain protein annotations, which are also mapped to the corresponding proteins through the Human Protein Reference Database (HPRD; http://www.hprd.org/). Thus, it is possible to visualize data pertaining to experimentally validated posttranslational modifications (PTMs), protein isoforms, protein–protein interactions (PPIs), tissue expression, expression in cell lines, subcellular localization and enzyme substrates in the context of individual proteins. With enthusiastic participation of the proteomics community, the past 15 months have witnessed data contributions from more than 75 labs around the world including 2710 distinct experiments, >1.9 million peptides, >4.8 million MS/MS spectra, 150 368 protein expression annotations, 17 410 PTMs, 34 624 PPIs and 2906 subcellular localization annotations. *Human Proteinpedia* should serve as an integrated platform to store, integrate and disseminate such proteomic data and is inching towards evolving into a unified human proteomics resource.

INTRODUCTION

Proteomics is the large-scale analysis of proteins whose major goals are to characterize features of gene products such as posttranslational modifications (PTMs), protein isoforms, subcellular localization, protein–protein interactions (PPIs) and tissue expression (1,2). Many proteomic experiments make use of high-throughput technologies such as yeast two-hybrid, mass spectrometry (MS), protein/peptide arrays or fluorescence microscopy techniques to yield multi-dimensional datasets. These datasets, which are often quite large, are not usually published in their entirety or are published as supplementary information that is not easily searchable. Given the heterogeneous nature of proteomic data, integration of such diverse data from multiple sources is quite difficult. Without a system for standardizing and sharing such data in place, it is less fruitful for the biomedical community to contribute these types of data to centralized repositories. Even more difficult is the annotation and display of pertinent information in the context of the corresponding protein. In light of these issues, we have developed *Human Proteinpedia* (http://www.humanproteinpedia.org/) as a portal that overcomes many of these obstacles to provide a unified view of the human proteome.

Researchers contributing data to *Human Proteinpedia* belong to various arena of biomedical research including biochemistry, molecular biology, bioinformatics, genetics, cell biology and pathology. Different types of MS datasets are included from mass spectrometers ranging from ion traps to quadrupole time-of-flight to Fourier transform instruments. To aid comparison and interpretation, all of the datasets are annotated with the sample, method of isolation and experimental platform-specific information (labeling method, protease used, fractionation

*To whom correspondence should be addressed. Tel: +410 502 6662; Fax: +410 502 7544; Email: pandey@jhmi.edu

mode, ionization mode as well as database search annotations such as search algorithm, peptide score for MS, primary antibody source organism, primary antibody company or catalog information and primary antibody titration for antibody-based experiments). In addition to the types of data, more than 80 human tissues and cell types, including liver, serum, brain, plasma, B lymphocytes, saliva, platelets and pancreas are represented. With the recent advancements in proteomics technologies and large-scale proteomics initiatives like Human Proteome Organization (HUPO) and Cancer Genome Anatomy Project (CGAP), proteomic data are sure to explode. We aim to make *Human Proteinpedia* as an exhaustive platform to store and disseminate diverse proteomic data.

DEVELOPMENT OF A DISTRIBUTED ANNOTATION SYSTEM FOR PROTEINS

Distributed annotation system (DAS) is a set of defined protocols that is used to share biological data. This system was originally developed to share annotations for genomic data (i.e. DNA information) of various organisms and comprised a genome server for hosting genome maps, sequences and sequencing information, an annotation viewer, which is basically a computer used to access a genome browser and one or more annotation servers. Some of the resources that already have DAS implementations include WormBase (3), FlyBase (4) and Ensembl (5) allowing integration and visualization of data in a browser. A major limitation of DAS is that the specifics of the protocol mainly relate to DNA (6) or, more recently, mRNA data (7). Furthermore, significant technical expertise is required for any laboratory to participate in the data sharing process. Thus, given the complexity of proteomic data, it is a prerequisite to alter the specifics of DAS before it can be used to share data pertaining to proteins.

Even though initial attempts to extend the specifics of the DAS protocol to suit protein data have been made by SPICE (7), Dasty (8), UniProt (9) and ProteinDAS (5), all of the inherent properties of proteins such as PTMs, protein isoforms, PPIs, tissue expression, subcellular localization, functional annotations and enzyme substrates have not yet been consolidated into a single annotation system. Our annotation system follows DAS standards but is extended to fit protein data that characterize protein functions. *Human Proteinpedia* simplifies data sharing by allowing contributors to provide data in four different ways: (i) annotation of data on individual proteins along with experimental evidence through the use of web forms; (ii) upload of data via the web in a batch mode; (iii) sending data through FTP/e-mail to the *Human Proteinpedia* team that carries out data processing, formatting, mapping and upload of data; and (iv) DAS servers set up by the contributing labs for upload of data. Supplementary Figure 1 shows the web interface for annotation of PTMs and tissue expression using simple pull-down menus and pop-up windows for specifying and mapping sites of modification or tissue expression using standardized vocabulary terms. We have included these simpler options for contributing data because although the laboratories can also set up their own annotation servers, many investigators are precluded from volunteering to provide their data because of the technical expertise required. In addition, we have also developed a semi-automatic author tracking system for capturing proteomic data. Here, we are scanning through the published issues till date in all of the major proteomics journals for possible inclusion in *Human Proteinpedia*. Validity checks are carried out to minimize logical errors (e.g. typographic errors). For instance, if a user submits a PTM of serine located at position 162 for a given protein accession number, NP_0123456.1, the automated validation system will check whether the accession provided belongs to humans and whether there is indeed a serine residue at the given position. Once verified, the data is uploaded into *Human Proteinpedia*.

All of the data submitted to *Human Proteinpedia* can be viewed through Human Protein Reference Database (HPRD) (10,11) in the context of an individual protein molecule, which provides an integrated view of the existing literature curated data along with the deposited annotations. The peptide data in *Human Proteinpedia* has been mapped onto the Ensembl human genome browser and can be viewed as separate tracks.

DATA STORAGE

Human Proteinpedia is the portal for querying, browsing and downloading the contributed datasets (Supplementary Figure 2). All meta-annotations pertaining to experiments (e.g. description of experiment, experimental platform, publication details, etc.) along with the protein annotations (e.g. identified peptides, autoradiographs, microscopy images, etc.) are stored in the server hosting *Human Proteinpedia*. On the other hand, storage and dissemination of the MS data is more challenging because it poses unique issues pertaining to hardware, file formats and file transfer. For this purpose, we have taken advantage of ProteomeCommons.org, which is a public repository for digital content relating to proteomics and incorporates most of the tools for interchanging file formats. The Tranche file-sharing network supporting ProteomeCommons.org (http://www.proteomecommons.org/dev/dfs) is a secure transactional system for proteomics data that incorporates public key/private key encryption techniques and a peer-to-peer distributed file system, which provides considerable flexibility for file transfer issues. The system is designed to support both raw proteomics data and metadata independent of file format. Currently, over 16 servers ($\sim$50 TB of aggregate capacity) host the MS datasets in triplicate including two servers (in Japan and USA) set up specifically for this initiative.

All of the raw and processed MS datasets referenced by *Human Proteinpedia* are deposited in Tranche, which delivers the datasets as required. The data can be accessed either by clicking a link to download a file via web browser or by a link that will launch the Tranche downloader. Tranche can also act as a repository for reference datasets for other metadatabases and includes the contents of

Peptide Atlas (12), TheGPMdb (13), Open Proteomics Database (OPD) (14) and PRoteomics IDEntifications database (PRIDE) (15,16). Tranche is file format independent so there are no barriers to importing data in standardized file formats (e.g. XML-based) or in proprietary file formats. Support for file conversion within Tranche is provided by the IO Framework (17), which supports all major standardized and proprietary MS file formats.

PROTEIN FEATURES ANNOTATED IN *HUMAN PROTEINPEDIA*

Protein features that can be annotated are briefly described below along with relevant examples that illustrate the functionality and utility of *Human Proteinpedia*.

Posttranslational modifications

Human Proteinpedia facilitates annotation and visualization of PTMs by mapping them onto the modified amino acid residue of the corresponding modified protein. This information is not readily available even in the published literature as either amino acid residue numbers or a short peptide sequence are usually reported as sites of modifications, both of which are disconnected from the protein level information. An example of a novel phosphorylation site on vimentin protein, annotated through the use of *Human Proteinpedia*, is shown in Figure 1. Vimentin is a cytoskeletal protein that is important for maintaining the integrity of cytoplasm and is involved in processes such as wound healing. It has recently also been shown to be secreted into the extracellular space by activated macrophages, an event that is dependent on phosphorylation of vimentin on serine and threonine residues (18). Thirteen novel phosphorylation sites on vimentin shared using *Human Proteinpedia* should provide more insight into the role of phosphorylation events that regulate its function and subcellular localization. As shown in the figure, additional details regarding the experiment pertaining to the phosphorylation data are shown under the heading '*Human Proteinpedia*'. The tandem MS (MS/MS) spectrum for one of the phosphopeptides is displayed through a spectrum viewer, obtained from PRIDE.

Tissue expression

Cataloging sites of expression for human proteins is an important task and one that cannot easily be tackled by a single laboratory or research group. The data pertaining to organ-based localization has grown considerably given several HUPO (e.g. plasma, liver, brain) and other initiatives [e.g. Human Protein Atlas project (19)]. Figure 2 shows the tissue expression data obtained by mass spectrometric analysis and immunohistochemical labeling from other laboratories for vimentin. This includes documentation of expression in human plasma from PeptideAtlas, in B cell, liver, platelets and serum from four different mass spectrometry laboratories and in lung, ovary, spleen, thyroid and tonsil from Human Protein Atlas.

Expression in cell lines

Given the fact that a large majority of biochemical and cell biology experiments are carried out in cell lines, investigators can also deposit data pertaining to proteins in cell lines such as proteins identified in a mass spectrometric analysis. Given the knowledge of expression of proteins in cell lines, investigators will be able plan their experiments in a wider variety of cell lines than would otherwise be possible. It could also cut costs, as investigators would not need to randomly test cell lines for expression themselves in a large battery of cell lines.

Protein–protein interactions

PPIs are crucial to most cellular events and represent an important component of current high-throughput screening modalities for therapeutic development. High-throughput proteomic techniques, such as identification of immunoprecipitated complexes, have recently become popular for systematically cataloging physical interactions between proteins on a large scale. *Human Proteinpedia* allows annotation of PPIs obtained from a number of platforms including yeast two-hybrid experiments, pull-down assays and protein complex identification by mass spectrometry. In all of these cases, standardized vocabulary is used to describe the experiments and the detection methods and has direct links to the published citation, if any.

Subcellular localization

Human Proteinpedia permits annotation of information pertaining to subcellular localization of proteins. The vocabulary used for this purpose is in compliance with cellular component of Gene Ontology terms. Also annotated are details about the experiment, fluorescence microscopy images (where available) and links to contributing laboratories. Figure 3 shows how a protein of unknown function that is present in the databases simply as a hypothetical protein (accession numbers MGC33867 or KIAA2013), was localized to the endoplasmic reticulum and the Golgi apparatus by two independent groups. While the report describing localization to the endoplasmic reticulum is only preliminary, the localization to the Golgi complex is supported by three lines of evidence: first, the investigators identified the proteins from a Golgi membrane preparation of rat liver tissue; second, they showed localization to the Golgi using a GFP fusion protein; and last, they raised an antibody against the human protein and demonstrated localization of the endogenous protein as well to the Golgi. Integration of all this information in one place is likely to spur further research on such proteins and exemplifies how new discoveries could be enabled by *Human Proteinpedia*.

Enzyme substrates

The transient nature of an enzyme/substrate reaction makes it difficult to be captured by many of the standard high-throughput experiments. The instances where such information is actually available are quite valuable. *Human Proteinpedia* allows enzymatic reactions to be

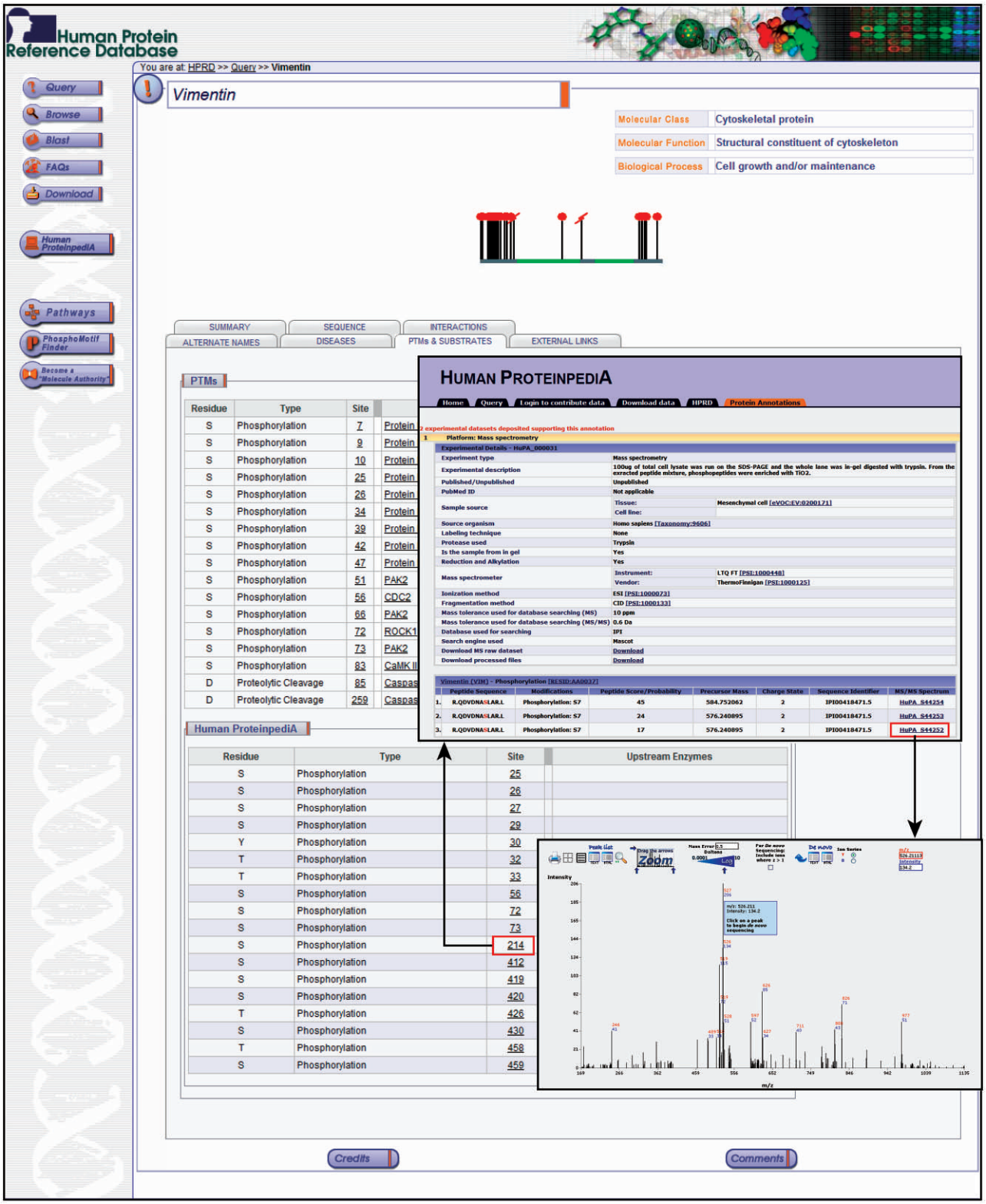

Figure 1. HPRD molecule page with 13 novel phosphorylation sites deposited in *Human Proteinpedia*. HPRD molecule page for vimentin is shown with a novel serine phosphorylation site that was contributed through *Human Proteinpedia*. Thirteen novel phosphorylation events were identified and shared using *Human Proteinpedia*. The panel on the middle displays description of the experiment, phosphopeptide identifications and the peptide score calculated by the respective search engine that was used. The tandem mass spectrometry (MS/MS) spectrum is linked to all of the peptides and the lower right panel shows the MS/MS spectrum of a phosphopeptide in a spectrum viewer.

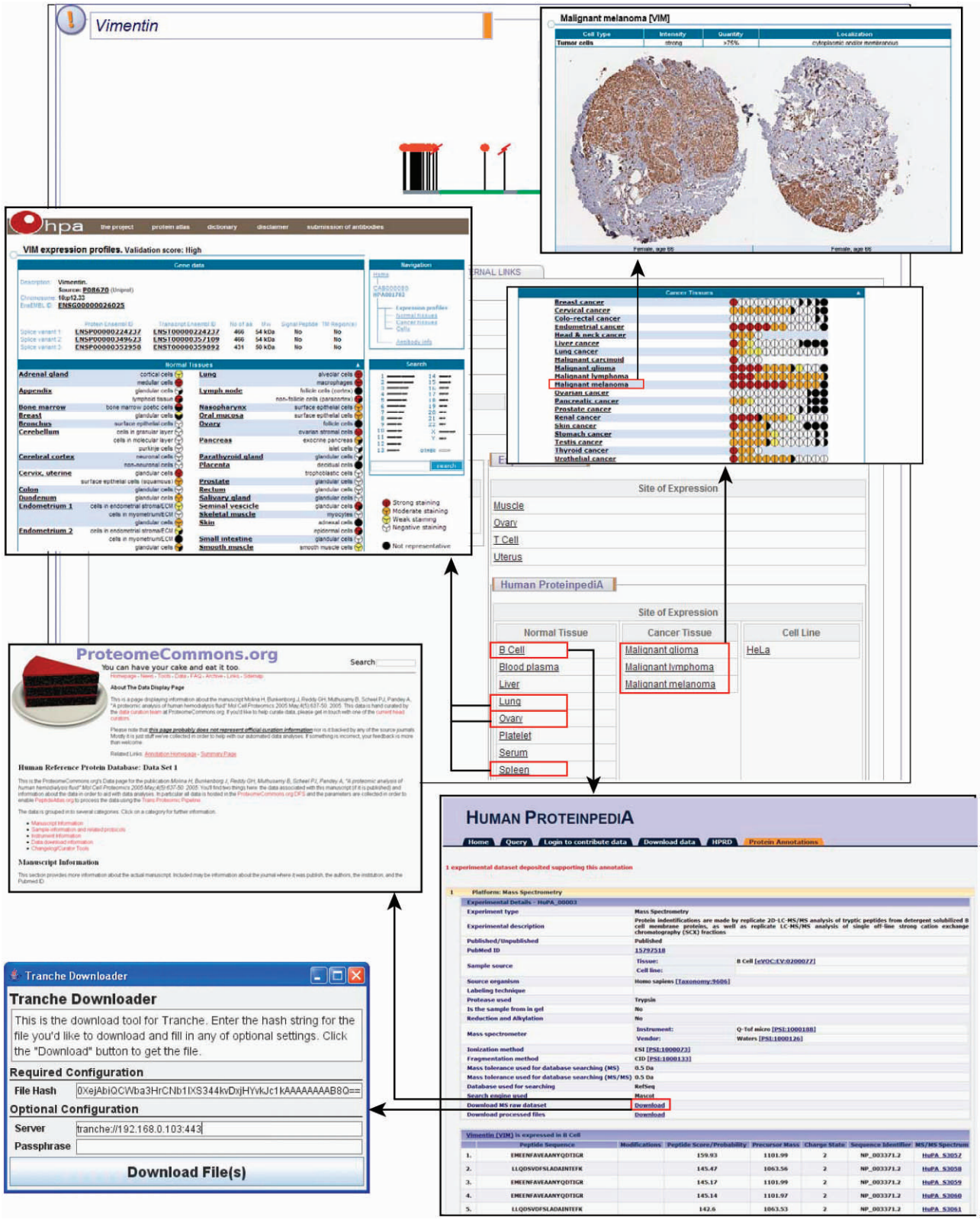

Figure 2. Display of tissue expression data submitted to *Human Proteinpedia*. Molecule page of vimentin in HPRD is displayed with novel tissue expression sites. It is shown as expressed in B cell and is hyperlinked to the experimental description, peptide identification data, peptide score, precursor mass, charge state, sequence identifiers and their corresponding MS/MS spectrum. Processed and raw MS/MS files are linked to ProteomeCommons as indicated in the lower left panel either using the web interface or by launching the Tranche downloader. Clicking on either one of lung, ovary, spleen, tonsil or thyroid gland will hyperlink to HPA page for vimentin that is shown in the upper right panel.

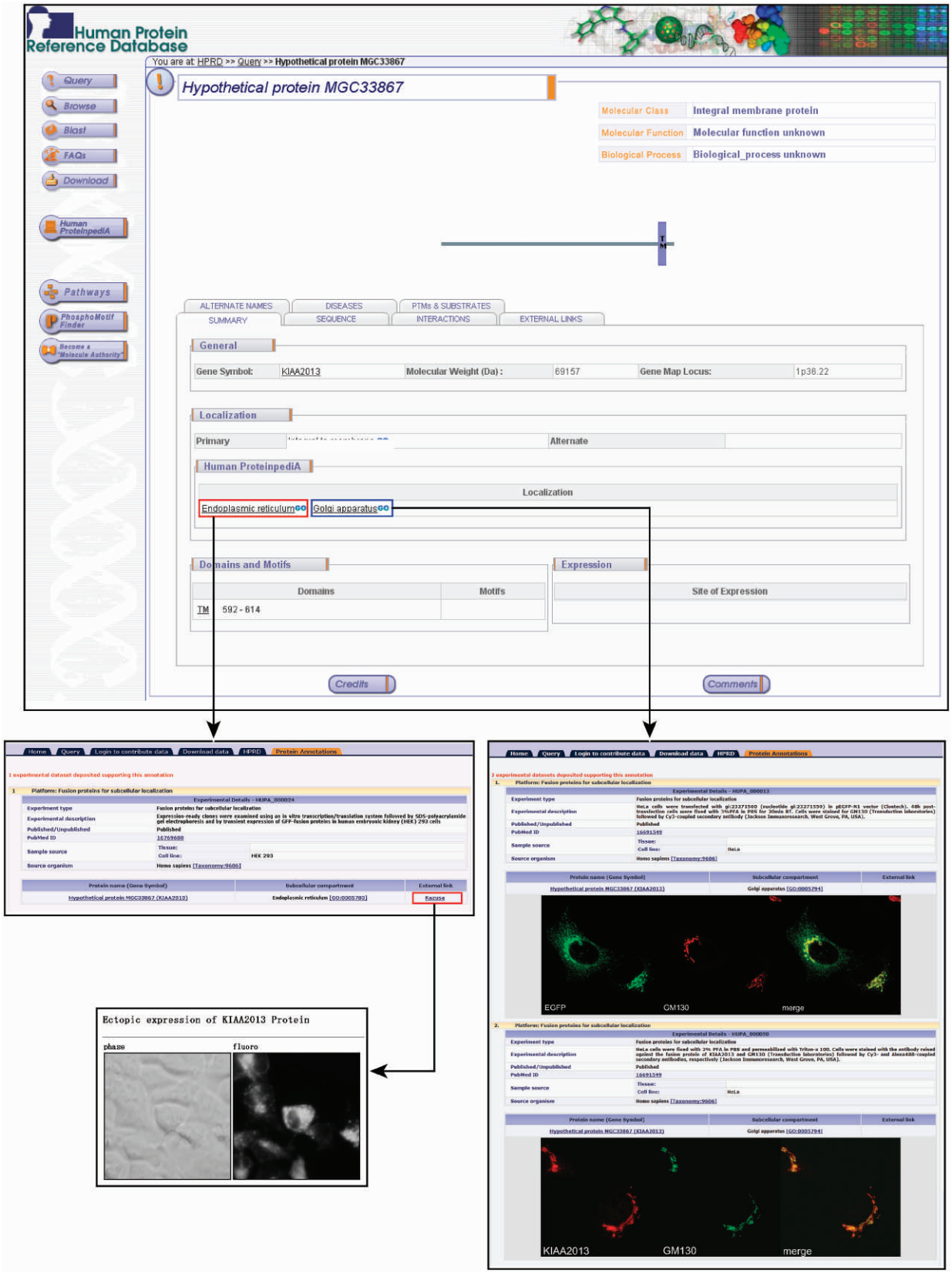

Figure 3. Display of subcellular localization data for a hypothetical protein: MGC33867/KIAA2013. Molecule page of a hypothetical protein MGC33867/KIAA2013 is displayed. Lower left panel shows subcellular localization of the protein to the endoplasmic reticulum, while the right panel shows identification of the same protein to be localized in the Golgi apparatus. Fluorescence microscopy images are provided along with the experimental annotations.

annotated along with details of the residue that is modified. The experimental type (e.g. *in vitro* experiment) and the detection method (e.g. phosphopeptide mapping or mass spectrometry) are also annotated.

COMPARISON OF *HUMAN PROTEINPEDIA* WITH EXISTING PROTEOMIC RESOURCES

Most of the proteomic repositories in the public domain are restricted to one or two experimental platforms. PeptideAtlas and PRIDE are specialized public repositories of peptides/proteins identified by tandem mass spectrometric experiments. PRIDE, in addition, also accepts two-dimensional gel data. *Human Proteinpedia* differs significantly from these repositories in its scope, data types accepted, annotation features and various options provided for submission of data.

(1) *Human Proteinpedia* does not exclusively contain mass spectrometry-derived data. The data that can be submitted to *Human Proteinpedia* ranges from yeast two-hybrid screens and peptide arrays to immunohistochemistry and western blots. The relevant details of the experiment performed are provided along with the data.
(2) *Human Proteinpedia*, with the effective use of DAS, allows participating laboratories to contribute both published and unpublished data. Some of the unpublished data belong to a category of data that are identified to be of high quality by the contributing laboratories, but will never get published as they were of secondary importance (e.g. two phosphorylation sites identified by a laboratory that was looking primarily for protein identifications in liver using MS).
(3) Data from all of the diverse proteomic experiments are viewed at the context of an individual protein through HPRD. *Human Proteinpedia* allows for the integration of data from multiple platforms in the context of an individual protein and links it with the detailed information on the annotation.
(4) Data submission is simplified to a greater extent with users having an option of submitting data one by one or batch wise via the web forms or transferring data in native format to the data processing team or setting up their own annotation server. A biologist with no technical expertise can login and contribute data using the easy to use web forms, which is one of the major goals of the *Human Proteinpedia*.
(5) *Human Proteinpedia* also incorporates data from articles that are published already. Our data processing team have retrieved data from various journals supplementary information and processed the large-scale proteomic identifications.
(6) PRIDE stores processed, but not raw, mass spectrometry-derived data. *Human Proteinpedia*, PeptideAtlas, ProteomeCommons and OPD store all types of mass spectrometry data.
(7) *Human Proteinpedia* restricts the data to that derived from human tissues or cell lines. All other repositories do not have this restriction.
(8) *Human Proteinpedia* allows contributing laboratories to annotate data pertaining to various features of proteins including PTMs, tissue expression, cell line expression, subcellular localization, enzyme—substrate interaction and PPIs.

STANDARDIZATION OF VOCABULARY

Standardization of data formats alone is not sufficient for facilitating data exchange. It is equally important to use standardized vocabulary before different types of data can be freely exchanged, queried or retrieved. eVOC (20), a set of controlled terms in a set of hierarchical vocabularies, are used to standardize human tissue nomenclature. Gene Ontology (21) terms for cellular component are used to designate subcellular compartments while RESID (22) and Proteomics Standard Initiative-Molecular Interaction (PSI-MI) (23) vocabularies are used to standardize PTMs and experiment types, respectively. Proteomics Standard Initiative-Mass Spectrometry (PSI-MS) vocabularies are used to standardize MS-based experimental annotations.

POTENTIAL USES OF DATA IN *HUMAN PROTEINPEDIA*

A large number of potential uses of the data deposited in *Human Proteinpedia* can be envisaged. The following is a brief list of some of the important uses that can be envisaged in the near future:

(1) Integration of data obtained from most of the proteomic assays, which were performed to capture the functional properties of proteins, in a single repository by itself promises great potential in elucidating functions performed by a protein.
(2) Data mining for additional information from the deposited MS datasets. For example, if the original MS/MS dataset obtained from characterization of serum or a particular cell line was simply used to annotate protein identifications, it could be used to specifically look for peptides with PTMs such as phosphorylation, acetylation, etc.
(3) A list of proteotypic peptides can be assembled from the high confidence peptides in the deposited data.
(4) Antibody that is specific against a protein is used in various functional assays including immunohistochemical staining, western blotting, fusion proteins for subcellular localization, ELISA and co-immunoprecipitation. Information on such primary antibodies, that have subsequently worked well, will be of great importance to a biomedical scientist.
(5) Meta analysis of data from different ionization methods, mass spectrometers or separation methods can be carried out.
(6) The repository of spectra can be used to generate a spectral 'library' that can be used for protein identification.
(7) PPI data can be used to study protein networks and pathways in cells.

Table 1. Statistics of submitted data

Number of individual laboratories submitting data	75
Number of experiments	2710
Number of protein annotations	221 578
MS/MS spectra	4 855 122
Number of peptide sequences deposited	1 960 352
Tissue expression	150 368
PPIs	34 624
Subcellular localization	2906
PTMs	17 410

(8) Information on the subcellular localization of proteins can be used to emulate a microenvironment of one single compartment, wherein detailed PPI analysis can be carried out.

COMMUNITY PARTICIPATION

In this effort, the data contributed by the proteomics community include over 34 000 PPIs obtained from yeast two-hybrid and co-immunoprecipitation assays, more than 17 000 PTMs obtained from mass spectrometric analyses, over 150 000 sites of protein tissue expression from immunohistochemistry and mass spectrometric analyses, 2906 protein subcellular localization obtained from fluorescence microscopy and mass spectrometric analyses, over 1.9 million peptides and more than 4.8 million MS/MS-spectra (Table 1). We have also imported mass spectrometry data from several HUPO initiatives including human plasma proteome project (HPPP), brain proteome project (HBPP) and liver proteome project (HLPP) that were deposited in two public repositories of MS data: PRIDE and PeptideAtlas. All of the data present in *Human Proteinpedia* is freely available to the community for downloading at http://www.humanproteinpedia.org/. We anticipate that ready availability of such diverse datasets will spur research in many new areas of human biology including signaling networks, biomarkers and cellular and developmental studies.

CONCLUSIONS AND OUTLOOK

Active participation by members of the proteomics community in sharing data through *Human Proteinpedia* has resulted in a prolific increase in the amount of protein annotations. Manual curation of scientific literature over the span of four years resulted in greater than 228 800 protein annotations in HPRD. The enthusiastic participation of the proteomics community over the last 15 months has almost doubled the number of protein annotations over short duration. This initial effort with 75 contributing laboratories enabled sharing of this large amount of data. We hope to include more human proteomic data in the near future as more investigators generating large proteomics datasets participate in this initiative. This would not be feasible without the continued participation of the community and we encourage all investigators to share their published and unpublished proteomic datasets with *Human Proteinpedia*. We anticipate that contribution of experimental data to a public repository will be made an essential criterion for publication as is already the case for nucleotide sequences, gene expression profiles and protein structures. With the participation of scientific community, we foresee the *Human Proteinpedia* to serve as a unified resource for proteomics research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health (U54 RR020839); Roadmap Initiative for Technology Centers for Networks and Pathways, partially. Funding for open access charge: National Institutes of Health (U54 RR020839).

Conflict of interest statement. None declared.

REFERENCES

1. Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.
2. Phizicky,E., Bastiaens,P.I., Zhu,H., Snyder,M. and Fields,S. (2003) Protein analysis on a proteomic scale. *Nature*, **422**, 208–215.
3. Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.*, **29**, 82–86.
4. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
5. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
6. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
7. Prlic,A., Down,T.A. and Hubbard,T.J. (2005) Adding some SPICE to DAS. *Bioinformatics*, **21 (Suppl 2)**, ii40–ii41.
8. Jones,P., Vinod,N., Down,T., Hackmann,A., Kahari,A., Kretschmann,E., Quinn,A., Wieser,D., Hermjakob,H. and Apweiler,R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198–3199.
9. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
10. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
11. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **32**, D497–D501.
12. Desiere,F., Deutsch,E.W., Nesvizhskii,A.I., Mallick,P., King,N.L., Eng,J.K., Aderem,A., Boyle,R., Brunner,E., Donohoe,S. *et al.* (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.*, **6**, R9.
13. Craig,R., Cortens,J.C., Fenyo,D. and Beavis,R.C. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.*, **5**, 1843–1849.
14. Prince,J.T., Carlson,M.W., Wang,R., Lu,P. and Marcotte,E.M. (2004) The need for a public proteomics repository. *Nat. Biotechnol.*, **22**, 471–472.
15. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.

16. Jones,P., Cote,R.G., Martens,L., Quinn,A.F., Taylor,C.F., Derache,W., Hermjakob,H. and Apweiler,R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **34**, D659–D663.
17. Falkner,J.A., Falkner,J.W. and Andrews,P.C. (2007) ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats. *Bioinformatics*, **23**, 262–263.
18. Mor-Vaknin,N., Punturieri,A., Sitwala,K. and Markovitz,D.M. (2003) Vimentin is secreted by activated macrophages. *Nat. Cell Biol.*, **5**, 59–63.
19. Uhlen,M., Bjorling,E., Agaton,C., Szigyarto,C.A., Amini,B., Andersen,E., Andersson,A.C., Angelidou,P., Asplund,A., Asplund,C. *et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics*, **4**, 1920–1932.
20. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V., McCarthy,M.I. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.
21. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
22. Garavelli,J.S. (2003) The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res.*, **31**, 499–501.
23. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *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.