Human Proteinpedia enables sharing of human protein data

Proteomic technologies, such as yeast twohybrid, mass spectrometry (MS), protein/peptide arrays and fluorescence microscopy, yield multi-dimensional data sets, which are often quite large and either not published or published as supplementary information that is not easily searchable.
Without a system in place for standardizing and sharing data, it is not 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 proteins. Wikipedia, an online encyclopedia that anyone can edit, has already proven quite successful1 and can be used as a model for sharing biological data. However, the need for experimental evidence, data standardization and ownership of data creates scientific obstacles.
Here, we describe Human Proteinpedia (http://www.humanproteinpedia.org/) as a portal that overcomes many of these obstacles to provide an integrated view of the human proteome. Human Proteinpedia also allows users to contribute and edit proteomic data with two significant differences from Wikipedia: first, the contributor is expected to provide experimental evidence for the data annotated; and second, only the original contributor can edit their data.
Human Proteinpedia's annotation system provides investigators with multiple options for contributing data including web forms and annotation servers. Although registration is required to contribute data, anyone can freely access the data in the repository. The web forms simplify submission through the use of pull-down menus for certain data fields and pop-up menus for standardized vocabulary terms. Distributed annotation servers using modified protein DAS (distributed annotation system) protocols developed by us (DAS protocols were originally developed for sharing mRNA and DNA data) permit contributing laboratories to maintain protein annotations locally. All protein annotations are visualized in the context of corresponding proteins in the Human Protein Reference Database (HPRD)3. **Figure 1** shows tissue expression data for alpha-2-HS glycoprotein derived from three different types of experiments.
Our unique effort differs significantly from existing repositories, such as PeptideAtlas and PRIDE5 in several respects. First, most proteomic repositories are restricted to one or two experimental platforms, whereas Human Proteinpedia can accommodate data from diverse platforms, including yeast two-hybrid screens, MS, peptide/protein arrays, immunohistochemistry, western blots, co-immunoprecipitation and fluorescence microscopy–type experiments.
Second, Human Proteinpedia allows contributing laboratories to annotate data pertaining to six features of proteins (posttranslational modifications, tissue expression, cell line expression, subcellular localization, enzyme substrates and protein-protein interactions;).
No existing repository currently permits annotation of all these features in proteins. Third, all data submitted to Human Proteinpedia are viewable through HPRD in the context of other features of the corresponding proteins. To aid comparison and interpretation, meta-annotations pertaining to samples, method of isolation and experimental platform-specific information are provided (e.g., labeling method, protease used, ionization method, details of primary antibody used).
And fourth, in spite of accommodating multiple data types, the data submission is simplified. This means that a biologist with no technical expertise can login and contribute data.
Thus far, a considerable body of data has been contributed to Human Proteinpedia by the community, with a total of >1.8 million peptides and >4 million MS/MS-spectra deposited.

The abovementioned data were derived from 2,695 individual experiments (single experiments are defined as immunohistochemistry performed with a specific antibody, a single MS run or a yeast two-hybrid screen). We have imported MS data from two Human Proteome Organization initiatives, including the human plasma proteome project (HPPP)6 and the human liver proteome project (HLPP). Data from other initiatives like the Human Protein Atlas8, Human Unidentified Gene-Encoded (HUGE) project9 and LIFEdb are also included in this effort. The vocabulary follows accepted community standards, such as eVOC, Gene Ontology, RESID, PSI-MI14 and PSI-MS.

All data in Human Proteinpedia are freely available to the community for downloading. Storage and dissemination of the raw and processed MS data sets is through the Tranche file-sharing network supporting ProteomeCommons.org15 (http://www.proteomecommons.org/dev/dfs/) whereas the remainder of the data is available directly from Proteinpedia website. The MS data sets are currently hosted by over 16 servers (~50 TB of aggregate capacity) in triplicate including three servers (in India, Japan and the United States) set up especially for this initiative.

Four years of continued manual curation of the scientific literature has yielded >228,800 protein annotations in HPRD. The information content in Proteinpedia collected over the course of the past year alone has resulted in the addition of roughly an equal number of entries. Although large collaborations for genome projects are now commonplace, this collaborative initiative of 71 participating laboratories is unprecedented for proteomic data. We are now beginning to contact authors of published proteomic studies in a systematic fashion requesting their participation. We eventually hope to capture most of the published data sets, which could be facilitated if depositing experimental data to a public repository is made mandatory for publication, as is already the case for nucleotide sequences, gene expression profiles and protein structures. In this regard, we have instituted a mechanism that allows referees to access data sets submitted by authors before publication in an authorized and anonymous fashion for evaluation purposes.

Table 1 Statistics of submitted data

Data type	Number
Individual laboratories submitting data	71
Experiments submitted	2,695
Protein annotations	203,293
MS/MS spectra	4,567,235
Peptide sequences deposited	1,851,124
Protein expression	138,487
Post-translational modifications	17,108
Protein-protein interactions	31,476
Subcellular localizations	2,906
Phosphorylation	16,062

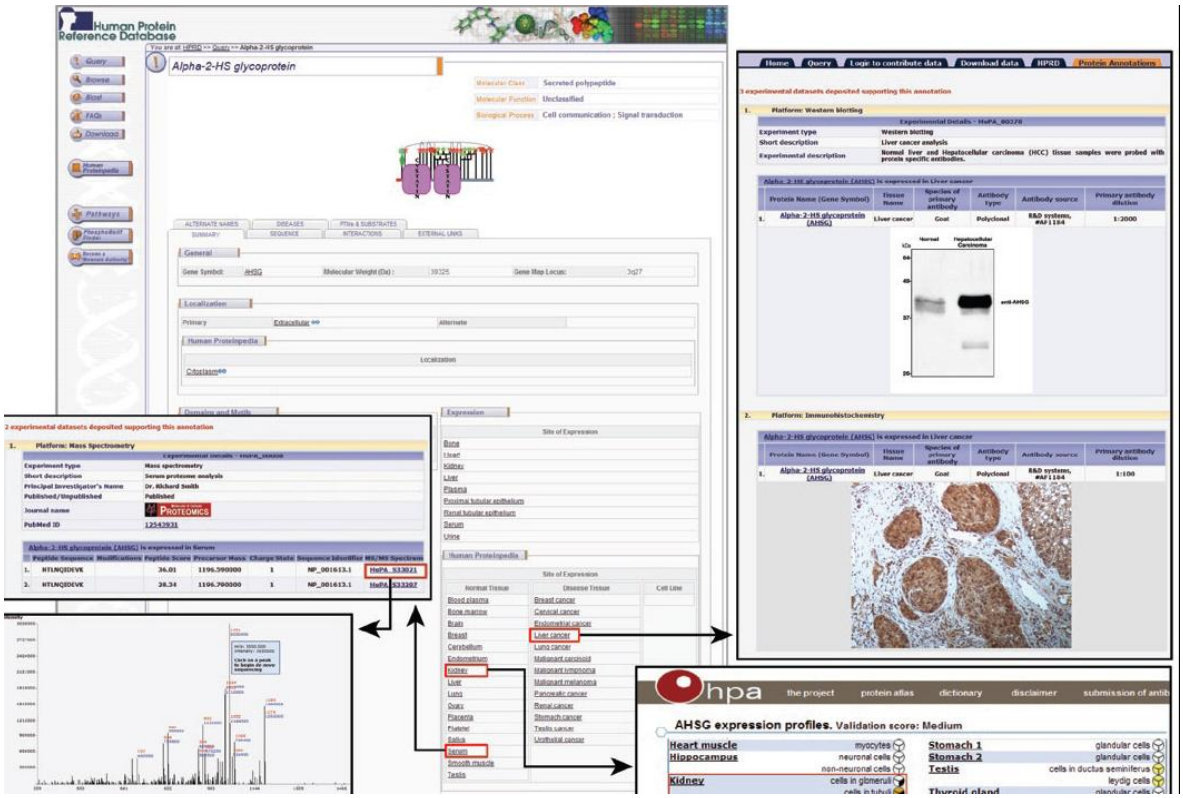

Figure 1 Display of tissue expression data obtained from immunohistochemistry, mass spectrometry and western blot analysis that were submitted to Human Proteinpedia. The molecule page for α-2-HS glycoprotein in Human Protein Reference Database is shown. Annotations pertaining to sites of tissue expression from three types of experimental platforms contributed to Human Proteinpedia are shown. The meta-annotation of a mass spectrometry experiment confirming expression in serum is shown (lower left) along with the corresponding MS/MS spectrum for a peptide displayed using a spectrum viewer obtained from PRIDE. Expression of α-2-HS glycoprotein in liver cancer is provided by two entries: one from western blotting (top right) and the other from immunohistochemical labeling, each detailing the antibody used. Expression in kidney is hyperlinked to the corresponding page in the Human Protein Atlas, which contains data from immunohistochemical labeling experiments (lower right).