A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies*

Lisa Berglund‡§, Erik Björling‡§, Per Oksvold‡, Linn Fagerberg‡, Anna Asplund¶, Cristina Al-Khalili Szigyarto‡, Anja Persson‡, Jenny Ottersson‡, Henrik Wernérus‡, Peter Nilsson‡, Emma Lundberg‡, Åsa Sivertsson‡, Sanjay Navani¶, Kenneth Wester¶, Caroline Kampf¶, Sophia Hober‡, Fredrik Pontén¶, and Mathias Uhlen‡**

An attractive path forward in proteomics is to experimentally annotate the human protein complement of the genome in a genecentric manner. Using antibodies, it might be possible to design protein-specific probes for a representative protein from every protein-coding gene and to subsequently use the antibodies for systematical analysis of cellular distribution and subcellular localization of proteins in normal and disease tissues. A new version (4.0) of the Human Protein Atlas has been developed in a genecentric manner with the inclusion of all human genes and splice variants predicted from genome efforts together with a visualization of each protein with characteristics such as predicted membrane regions, signal peptide, and protein domains and new plots showing the uniqueness (sequence similarity) of every fraction of each protein toward all other human proteins. The new version is based on tissue profiles generated from 6120 antibodies with more than five million immunohistochemistry-based images covering 5067 human genes, corresponding to ~25% of the human genome. Version 4.0 includes a tentative list of members in various protein classes, both functional classes, such as kinases, transcription factors, G-protein-coupled receptors, etc., and project-related classes, such as candidate genes for cancer or cardiovascular diseases. The exact antigen sequence for the internally generated antibodies has also been released together with a visualization of the application-specific validation performed for each antibody, including a protein array assay, Western blot analysis, immunohistochemistry, and, for a large fraction, immunofluorescence-based confocal microscopy. New search functionalities have been added to allow complex queries regarding protein expression profiles, protein classes, and chromosome location. The new version of the protein atlas thus is a resource for many areas of biomedical research, including protein science and biomarker discovery. Molecular & Cellular Proteomics 7:2019–2027, 2008.

One of the largest challenges in the postgenome era is to explore and understand the instructions embedded in the human genome. The sequencing of the human genome has revealed ~20,500 protein-encoding genes (1), and it is now possible to envision a genecentric proteomics approach to map the protein-based molecular architecture of the human body (2). The generation of “probes” in the form of protein-specific antibodies (3) allows for studies of the corresponding proteins and protein isoforms using a wide range of assays, including analysis of tissue and cell protein profiles.

The major repository for protein information is the Universal Protein Resource (UniProt) (4), a collaborative effort between the Swiss Institute of Bioinformatics, the European Bioinformatics Institute, and the Protein Information Resource. An alternative portal was recently described, the Human Proteinpedia (5), in which researchers themselves can submit annotated protein data from a large number of technology platforms, such as mass spectrometry, Western blots, and immunohistochemistry, with accompanying experimental evidence. In the current version of UniProtKB (release 14), there are 20,069 reviewed human protein entries. Only approximately half (11,371) of these entries have evidence at the protein level. This emphasizes the need for protein-specific antibodies to extend the number of verified proteins with the aim to create a complete list of all human proteins forming a basis for protein-related research in human biology and medicine.

Here we report on a new version (4.0) of the publicly available Human Protein Atlas with new features and a new structure based on a genecentric and genome-complete view of the protein complement of the human genome. The protein atlas portal is the result of a large scale effort to create a resource for human protein expression profiles in normal tissues, cancer cells, and cell lines with validated antibodies as reagents. The results show a path forward to create a comprehensive resource covering all human proteins.

THE PROGRESS OF THE HUMAN PROTEIN ATLAS

The first version (1.0) of the Human Protein Atlas portal was released in August 2005 and contained 275 internally generated antibodies and 443 external antibodies obtained from various commercial antibody vendors (Table I). The antibodies
A Genecentric Human Protein Atlas

TABLE I
The progress of the Human Protein Atlas

<table>
<thead>
<tr>
<th>Version</th>
<th>Year</th>
<th>No. of genes</th>
<th>No. of external Abs</th>
<th>No. of internal Abs</th>
<th>Total no. of Abs</th>
<th>New features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2005</td>
<td>650</td>
<td>443</td>
<td>275</td>
<td>718</td>
<td>Protein profiles for 48 normal tissues (IHC); protein profiles for 20 cancer types (IHC)</td>
</tr>
<tr>
<td>2.0</td>
<td>2006</td>
<td>1344</td>
<td>759</td>
<td>755</td>
<td>1514</td>
<td>Protein profiles for 59 cell lines and cells (IHC)</td>
</tr>
<tr>
<td>3.0</td>
<td>2007</td>
<td>2600</td>
<td>1241</td>
<td>1774</td>
<td>3015</td>
<td>Subcellular localization in three cell lines (IF); validation results (Western blot/protein array); search function (tissues and cancers)</td>
</tr>
<tr>
<td>4.0</td>
<td>2008</td>
<td>5067</td>
<td>2251</td>
<td>3869</td>
<td>6120</td>
<td>Genecentric structure with all predicted genes; advanced searches (protein classes and cells); visualization of antigen</td>
</tr>
</tbody>
</table>

*Abs, antibodies.  
<sup>a</sup> Immunofluorescence.

A GENECENTRIC PROTEIN ATLAS

The first three versions of the protein atlas were “antibody-centric” with a database structure based on the results from each antibody. The new version 4.0 has been changed to a “genecentric” structure in which all genes predicted by Ensembl (12) are included as part of the database. Altogether the 21,528 predicted protein-coding genes of Ensembl release 50.36 are contained in the database at present; this is slightly higher than the 20,500 genes predicted by Clamp et al. (1). A search in the portal will result in a list of protein-coding genes, and this is illustrated in Fig. 1 in which a partial list of all the genes on chromosome X is displayed. The Antibody ID column shows whether an antibody has been used to generate tissue profiles as is the case for most of the genes shown here. However, the protein product of the first gene (ATP1B4) and two other genes further down (ATXN3L and AVPR2) do not yet have approved antibodies, and consequently there are no tissue profiles in the database at present as shown by the lack of an Antibody ID and no data in the Validation column. Other genes have several antibodies to the same protein target as exemplified by the second gene (ATP2B3), which has both an internally generated antibody (HPA) and an external commercial antibody (CAB), and the last gene (BCAP51), which has three antibodies directed toward its protein product, one internally generated (HPA) and two external antibodies (CAB), in this case from two different antibody vendors. Consequently three different antibodies from three different origins are available to this gene product, and this allows for comparative studies and better validation of the staining results.

VALIDATION OF INTERNALLY GENERATED ANTIBODIES

The specificity of the internally generated antibodies was validated by protein microarrays (8, 13) in which 384 human recombinant protein fragments were spotted on a microarray, and the antibody specificity was determined using a fluorescence-based analysis. The antibodies were further validated by Western blotting using tissue extracts from liver and tonsil, pooled human plasma depleted of IgG and albumin, and cell extracts from two human cell lines (8). In addition, an IHC

<sup>1</sup> The abbreviations used are: IHC, immunohistochemistry; HPA, Human Protein Atlas; CAB, collaborator antibody; GI, gastrointestinal; ID, identifier.
validation was performed using a tissue microarray containing 13 normal tissue types, four tumor types, and eight cell lines in duplicate or triplicate with 42 microarray cores altogether. The results of the Western blots and the immunohistochemistry were annotated based on known literature and bioinformatics data (presence of signal peptide, transmembrane regions, or other localization signals).

The results of the validation of more than 9000 internally generated antibodies are shown in Table II. Approximately half (49%) of the antibodies fail because of a staining pattern that is not consistent with literature or bioinformatics data. Antibodies with supportive validation constituted 22% of the reagents, and of these 7% were paired antibodies with similar staining pattern, consistent with experimental and/or bioinformatics data (score 1). 29% of the antibodies were categorized as uncertain with no literature data available or with staining pattern only partly consistent with earlier experimental data and/or bioinformatics. The results show that approximately half (51%) of the internally generated antibodies are scored 1–4, and these are subsequently made public through the Human Protein Atlas.

It is noteworthy that the validation of the antibodies (immunohistochemistry and Western blotting) depends on subjective decisions based on a comparison of experimental results with information obtained via bioinformatics prediction methods and literature, which in many cases are inconclusive. As an example, the lack of developmental tissues in the microarray setup excludes the validation of antibodies to protein targets that are only expressed during developmental stages. Observed discordant IHC patterns can be explained by the presence of protein isoforms, such as splice variants or post-translationally modified proteins, but cross-reactivity to other proteins cannot be excluded. The availability of validation data on a public database thus facilitates input from the scientific community regarding the possible cross-reactivity.
and specificity of each antibody (14). In this context, we have decided to make the supporting validation results available with the supporting images available for review (Fig. 1). Our aim is to encourage open access to the validation results to facilitate international efforts to compare antibodies generated by different programs.

VALIDATION OF EXTERNAL ANTIBODIES

A call for antibodies toward human proteins was sent to a large number of commercial vendors. 5436 external antibodies from 51 different antibody providers were obtained and subsequently validated. Of these, 1410 monoclonal antibodies and 1316 polyclonal antibodies were approved by a standardized validation using Western blotting and IHC on tissue microarrays as described above. The success rates stratified by the different providers (Fig. 2) showed large differences, ranging from 0 to 100% of the antibodies with an average success rate of 49%. It is important to point out that many of these antibodies have not been approved by the antibody providers for immunohistochemistry; this might explain the low success rate in our hands. In addition, some of the antibodies might give supportive data if a more specialized assay was designed for each antibody making sure that the target protein is present in the assay. Anyhow the results enforce the need for publicly available validation results for antibodies because half of the commercially available antibodies did not pass our quality assurance.

INFORMATION ABOUT THE ANTIGEN

We have earlier argued (13) that a long term objective for a human antibody initiative could be to generate two independent antibodies toward all non-redundant proteins, allowing the results and validation of one antibody to be used to validate the other. The objective to generate paired antibodies with non-overlapping epitopes opens up the need to

**TABLE II**

Validation of 9358 internally generated antibodies

<table>
<thead>
<tr>
<th>Validation score</th>
<th>Validation category</th>
<th>Criteria</th>
<th>Number of antibodies</th>
<th>Fraction %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. High</td>
<td>Supportive</td>
<td>Two independent antibodies with similar staining pattern consistent with experimental and/or bioinformatics data</td>
<td>687</td>
<td>7</td>
</tr>
<tr>
<td>2. Medium</td>
<td>Supportive</td>
<td>Staining pattern consistent with experimental and/or bioinformatics data</td>
<td>1414</td>
<td>15</td>
</tr>
<tr>
<td>3. Low</td>
<td>Uncertain</td>
<td>Staining pattern partly consistent with experimental and/or bioinformatics data</td>
<td>1853</td>
<td>20</td>
</tr>
<tr>
<td>4. Very low</td>
<td>Uncertain</td>
<td>No literature or bioinformatics data available</td>
<td>861</td>
<td>9</td>
</tr>
<tr>
<td>5. Failed</td>
<td>Failed</td>
<td>Staining pattern not consistent with experimental and/or bioinformatics data</td>
<td>4543</td>
<td>49</td>
</tr>
</tbody>
</table>

**Fig. 2. Success rates in validation of antibodies from external providers.** 5399 antibodies from external providers were validated by the standardized approach using Western blotting and tissue microarrays. The figure displays the fraction of approved antibodies (grouped by monoclonal or polyclonal antibodies) per external provider (2374 monoclonal antibodies and 3025 polyclonal antibodies from 29 antibody providers). Only external providers that supplied at least five monoclonal/polyclonal antibodies are represented in the figure. ab, antibody.
gain precise knowledge about the binding region of each antibody either by epitope mapping of each antibody, such as the use of overlapping synthetic peptides, or if a synthetic peptide or a recombinant fragment has been used to gain access to the sequence used to generate the antibody. All the internally generated antibodies have been produced via a protocol for polyclonal (monospecific) antibodies immunized and affinity-purified using a recombinant protein epitope signature tag (PrEST) as antigen (15). With the new version of the Human Protein Atlas (4.0), all antigen sequences for the internally generated antibodies (63% of all the antibodies) are disclosed as part of the antibody information page to facilitate such development and to aid in international efforts to generate paired antibodies. The position of the antigen HPA003906 used for generation of the antibody is shown in the upper right corner of the protein view.

**Protein Classes**

The new version of the atlas portal also includes the possibility to stratify the proteome according to various protein classes. This can be done in three different ways: (i) functional classes, such as kinases, transcription factors, or G-protein-coupled receptors, (ii) a list of interesting proteins, such as plasma proteins, candidate markers of cardiovascular disease, or candidate cancer markers, and/or (iii) chromosomal location of the corresponding gene. Some examples of protein classes included in the protein atlas are shown in Table III with information about the number of genes in each class. The table also shows the fraction of proteins with tissue profiles (validated antibodies) for each protein class ranging from 26% for the transcription factors to 65% for the candidate markers of cardiovascular disease (Table III).

**Advanced Searches**

The search function described by Björling et al. (10) has been extended to allow complex queries, including combined searches based on protein classes, chromosomal location,
The number of genes in each protein class is given by “Predicted genes” (as determined by Ensembl version 50.36 (12)). “Tissue profiles” gives the number of genes for each class analyzed by antibodies; “Fraction” shows the percentage of analyzed proteins. “Source” is the source from which the original gene or protein identifiers for each class have been retrieved. TCDB, Transporter Classification Database; DBD, DNA-binding domain.

<table>
<thead>
<tr>
<th>Protein class</th>
<th>Predicted genes</th>
<th>Tissue profiles</th>
<th>Fraction</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinases</td>
<td>503</td>
<td>273</td>
<td>54</td>
<td>UniProt (4)</td>
</tr>
<tr>
<td>Transports</td>
<td>472</td>
<td>162</td>
<td>34</td>
<td>TCDB (30)</td>
</tr>
<tr>
<td>Transcription factors</td>
<td>1492</td>
<td>391</td>
<td>26</td>
<td>DBD (31)</td>
</tr>
<tr>
<td>Enzymes</td>
<td>2492</td>
<td>980</td>
<td>39</td>
<td>ENZYME (32)</td>
</tr>
<tr>
<td>G-protein-coupled receptorsa</td>
<td>392</td>
<td>119</td>
<td>30</td>
<td>UniProt (4)</td>
</tr>
<tr>
<td>Plasma proteins</td>
<td>687</td>
<td>384</td>
<td>56</td>
<td>Plasma Proteome Inst. (33)</td>
</tr>
<tr>
<td>Candidate markers of cardiovascular disease</td>
<td>166</td>
<td>108</td>
<td>65</td>
<td>Plasma Proteome Inst. (34)</td>
</tr>
<tr>
<td>Candidate cancer biomarkers</td>
<td>1067</td>
<td>668</td>
<td>63</td>
<td>Plasma Proteome Inst. (22)</td>
</tr>
<tr>
<td>Chromosome 22</td>
<td>504</td>
<td>176</td>
<td>35</td>
<td>Ensembl (12)</td>
</tr>
<tr>
<td>Chromosome X</td>
<td>884</td>
<td>321</td>
<td>36</td>
<td>Ensembl (12)</td>
</tr>
</tbody>
</table>

a Excluding olfactory and gustatory receptors.

and/or tissue profiles. It is also possible to include searches based on the expression in the cell lines in addition to tissue profiles in the normal and/or cancer tissues. In Fig. 4, two examples of combined searches are presented to illustrate the new search function. The first example (Fig. 4A) is aimed to identify proteins from a certain protein class encoded from a particular chromosome and expressed in a defined cell type. A search was made for transcription factors located on the human chromosome X with high expression in ovary stromal cells, and four proteins fit these criteria. The first protein is the erythroblast transformation specific domain-containing protein Elk-1, a transcription factor part of the ternary complex factor subfamily. This protein is the nuclear target for the Ras-Raf-MAPK (mitogen-activated protein kinase) signaling cascade, and the atlas summary shows strong expression in ovary and placenta but also in parts of the brain (cerebellum). The second protein is the methyl-CpG-binding protein 2 (MeCP2), which is a nuclear protein that binds to methylated DNA. This protein can suppress transcription by binding to CpG-containing promoters, and the protein is essential for embryonic development. The gene is subject to X chromosome inactivation and is the major cause of female mental retardation. The tissue profiling summary shows high expression in almost all tissues with an expected nuclear localization confirmed by the confocal images. The third protein is the multifunctional transcription factor YY2 (yin and yang 2), suggested to antagonize YY1 (also known as δ transcription factor NF-E1), which shows high sequence similarity to YY2. The protein YY1 is a ubiquitously distributed transcription factor belonging to the GLI-Kruppel class of zinc finger proteins, and it is involved in repressing and activating a diverse number of promoters. The tissue profiles of the atlas show an expected nuclear localization and expression in most cells except the brain. Finally the query found zinc finger protein 75, which is also known as zinc finger protein 82. The function of this protein is unknown. The tissue profile summary shows expression in all cells. Again the immunohistochemistry shows a nuclear staining, and this is supported by the confocal microscopy images showing staining of nuclei in all three cell lines.

The second query is aimed to identify proteins overexpressed in cancer as compared with its normal cellular counterpart. Based on the list of candidate cancer proteins identified by Polanski and Anderson (22), a search was made to investigate whether any proteins on this list were expressed at high levels in at least six of 12 colorectal cancer patients, whereas the expression in normal glandular cells in the colon is low (Fig. 4B). Again four proteins fit these criteria. The first protein is the serine/threonine-protein kinase 12 (Aurora-B), also known as Aurora-related kinase 2. The Aurora kinases associate with microtubules during chromosome movement and segregation, and this protein localizes to microtubules near kinetochores, specifically to the specialized microtubules called K-fibers. The tissue profiles of the Human Protein Atlas show strong expression in six of 12 colon cancers, and the expression is, as expected, in the nuclei. The protein is also expressed highly in a subset of lung cancer and cervical cancer patients. The second protein is the G2/mitotic-specific cyclin-B1. This protein is a regulatory protein involved in mitosis, and the gene product complexes with p34 (Cdc2) to form the maturation-promoting factor. The tissue profiles show low expression in the normal colon (and the GI tract), whereas six of 12 colon cancer patients express this protein abundantly. The protein is expressed highly in a few breast cancers and malignant lymphomas. The third protein is the macrophage migration-inhibitory factor, also known as the glycosylation-inhibiting factor, which is a lymphokine involved in cell-mediated immunity, immunoregulation, and inflammation. It plays a role in the regulation of macrophage function in host defense through the suppression of anti-
inflammatory effects of glucocorticoids. This protein is expressed in low amounts in normal colon and the GI tract, whereas six colon cancer patients show high expression. Finally the search identified the human transcription factor SOX-9. This protein recognizes the sequence CCTTGAG along with other members of the high mobility group box class DNA-binding proteins. It acts during chondrocyte differentiation and regulates transcription of the anti-Müllerian hormone gene. Deficiencies result in the skeletal malformation syndrome camptomelic dysplasia. The tissue profiles show a distinct nuclear staining, which is confirmed by the analysis using confocal microscopy. Eight of 12 colon cancers express high amounts of this protein in contrast to the normal GI tract. SOX-9 is also highly expressed in other cancers, in particular skin cancer.

**CONCLUDING REMARKS**

Here we describe a first attempt to experimentally annotate the human protein complement of the genome in a genecentric manner. A new version (4.0) of the Human Protein Atlas...

---

**Fig. 4. Lists of genes resulting from advanced searches.** The new advanced search allows for complex queries involving expression profiles, protein classes, and chromosomal localization of genes (Chr). A, a search for genes coding for transcription factors on chromosome X with strong staining in ovarian stromal cells results in a list of four candidate genes. B, a search for high levels of expression of candidate cancer biomarkers in colorectal cancer with low levels in normal colon results in a list of four candidate genes. PA, protein array; WB, Western blotting; IH, immunohistochemistry; IF, immunofluorescence.
A Genecentric Human Protein Atlas

has been developed with the inclusion of all human genes and splice variants from the Ensembl (12) database. The new version is based on tissue profiles generated from 6120 antibodies with more than five million immunohistochemistry-based images covering 5067 human genes.

An important objective of a genecentric antibody-based proteomics project is to generate paired antibodies with non-overlapping epitopes (8). In this context, an attractive path forward might be to use the results from the antibody pairs with separate and distinct epitopes to subtract staining that is only observed using one antibody. In this manner, the staining observed for both antibodies can be scored as specific, although caution has to be taken not to exclude true differences between the antibodies due to the presence of splice variants or protein modifications. A non-redundant set of the human proteome corresponds to ~21,000 proteins (1), and the objective of an initial human antibody-based proteome project would then, according to this strategy, be to generate 42,000 validated affinity reagents directed to two non-overlapping epitopes of each human protein. Ideally the binder pairs should consist of different types of affinity reagents, such as monoclonal antibodies (23), in vitro selected antibody fragments (24, 25), affibodies (26), or aptamers (27). If one of the affinity reagents could be generated by the strategy used here involving the generation of affinity-purified monospecific antibodies, the other affinity reagent should thus be generated with an alternative technology platform, preferably generating a renewable resource of reagents available to the scientific community. This emphasizes the need for international efforts to coordinate such antibody-based efforts as previously discussed by Haab et al. (28) and Taussig et al. (29).

The antibody-based proteomics strategy described opens up the possibility to perform antibody-based functional studies of the complete proteome with an almost endless set of analyses to investigate the corresponding proteins. Such studies should be complemented with alternative proteomics techniques, such as green fluorescent protein tagging, interaction analysis, and high resolution mass spectrometry analysis, to further validate the results from protein probe-based assays. Such studies could include detailed analysis of protein isoforms and global spatial and temporal protein profiles in both normal and disease tissues with the objective to create a genecentric human proteome map as a basis for biomedical and biological research.

Acknowledgments—We are grateful to Claes Wilhelmsson and Kenneth Nilsson for useful comments and advice. The entire staffs of the Human Proteome Resource centers in Stockholm and Uppsala, Sweden and in Mumbai, India are acknowledged for tremendous efforts.

* This work was supported by grants from the Knut and Alice Wallenberg Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Joint first authors. ** To whom correspondence should be addressed. Tel.: 46-8-5537-8325; Fax: 46-8-5537-8482; E-mail: mathias@biotech.kth.se.

REFERENCES


strategy for protein profiling in cell microarrays using automated image analysis. Proteomics 7, 2142–2150


