Toward a Confocal Subcellular Atlas of the Human Proteome*§

Laurent Barbe‡, Emma Lundberg‡, Per Oksvold‡, Anna Stenius‡, Erland Lewin§, Erik Björling‡, Anna Asplund¶, Fredrik Pontén¶, Hjalmar Brismar§, Mathias Uhlen‡, and Helene Andersson-Svahn‡

Information on protein localization on the subcellular level is important to map and characterize the proteome and to better understand cellular functions of proteins. Here we report on a pilot study of 466 proteins in three human cell lines aimed to allow large scale confocal microscopy analysis using protein-specific antibodies. Approximately 3000 high resolution images were generated, and more than 80% of the analyzed proteins could be classified in one or multiple subcellular compartments(s). The localizations of the proteins showed, in many cases, good agreement with the Gene Ontology localization prediction model. This is the first large scale antibody-based study to localize proteins into subcellular compartments using antibodies and confocal microscopy. The results suggest that this approach might be a valuable tool in conjunction with predictive models for protein localization. Molecular & Cellular Proteomics 7:499–508, 2008.

The identification of the subcellular localization of all proteins can provide key insights into the cellular function of the individual proteins as well as their probable interacting partners (1). So far systematic studies of subcellular localization of proteins have been performed either with cellular fractionation (2) or fluorescence microscopy (3, 4). The most commonly used approach in the latter case is based on transfection of cells with cDNA clones fused to a fluorescent reporter protein, like green fluorescent protein. The fluorescent reporter protein is cloned to either the N or C terminus of the target protein. After coating each well with human 12.5 g/ml fibronectin, (Sigma-Aldrich) for 1 h followed by washes with PBS, 15,000 cells (10,000 for A-431 cells) were seeded per well and grown for 3 h. All the subsequent steps were performed in 96-well format using a pipetting work

EXPERIMENTAL PROCEDURES

Cell Culture—Three different human cell lines were used in this study: U-251MG, a glioblastoma cell line, provided by Prof. Bengt Westermarck, Uppsala University (Uppsala, Sweden); A-431, an epidermoid carcinoma cell line, obtained from the Deutsche Sammlung von Mikroorganismen und Zellkultwien (DSMZ) cell line bank (Braunschweig, Germany); U-2OS, an osteosarcoma cell line, obtained from ATCC-LGC Promochem (Boras, Sweden). All cell lines were grown at 37 °C in a 5% CO2 environment in media suggested by the provider. All growth media were supplemented with 10% fetal bovine serum and an antibiotic/antimycotic solution (all from Invitrogen).

Immunofluorescent Sample Preparation—Immunofluorescent stainings were prepared in 96-well glass bottom plates (Whatman). After coating each well with human 12.5 μg/ml fibronectin, (Sigma-Aldrich) for 1 h followed by washes with PBS, 15,000 cells (10,000 for A-431 cells) were seeded per well and grown for 3 h. All the subsequent steps were performed in 96-well format using a pipetting work

1 The abbreviations used are: HPA, Human Protein Atlas; GO, Gene Ontology; PFA, paraformaldehyde; DAPI, 4′,6-diamidino-2-phenylindole.
Toward a Confocal Subcellular Atlas of the Human Proteome

RESULTS

Work Flow—Fig. 1 shows the work flow used in this study from high throughput production of monospecific polyclonal antibodies to protein profiling and subcellular localization (7–9). The first step (step 1) is the design of protein fragments (100–150 amino acids) using bioinformatics tools and the Ensembl database. The next step (step 2) yields monospecific antibodies by means of cloning, protein expression, immunization, and affinity purification (8). A key step is the standardized quality controls of the antibodies (step 3 in Fig. 1), and here we analyzed the antibody specificity and selectivity by protein arrays and Western blots to validate protein size and expression pattern (9). All antibodies are further used for immunohistochemical staining of microarrays comprising a variety of normal and cancer tissue as well as cell samples.
and cell lines (7, 10) (step 4). Each image from the tissue microarrays is annotated by certified pathologists to score protein expression levels in the different organs. For cell microarrays, automated annotation is achieved with a tailored image analysis application (10). The confocal atlas (step 5), described in this study, adds subcellular information to the previously described protein profiling tools. The data presented here are embedded in a publicly available Website (step 6) with ~2.8 million immunohistochemistry images corresponding to 3015 antibodies (version 3.0).

Sample Preparation and Image Acquisition—Three cell lines, U-2OS, A-431, and U-251MG, originating from different human tissues were chosen to be included in the atlas. To enhance the probability for a large number of proteins to be expressed, we selected cell lines from different lineages, i.e. tumor cell lines from mesenchymal, epithelial, and glial tumors. The selection was furthermore based on morphological characteristics, widespread use, and multitude of publications using these cell lines. The U-2OS cell line was derived from an osteosarcoma from a 150-year-old patient and exhibits an epithelial adherent morphology despite that it originates from a malignant cell type displaying a mesenchymal phenotype. The A-431 cell line, derived from a squamous cell carcinoma from an 85-year-old female patient, originates from an epidermoid carcinoma and represents an epithelial cell lineage. This cell line has been extensively used as a model for epithelial cancer. The third cell line, U-251MG, has also been extensively used in glioma research. This cell line is derived from a high grade malignant glioma (glioblastoma multiforme) and is expected to express several brain-specific proteins, which may not be present in cells derived from epithelial or sarcomatous tumors.

Several different cellular probes were tested as reference for identification of organelles and other cellular compartments. Due to limitation of the excitation/emission characteristics of the fluorescent probes used (i.e. to avoid problems associated with spectrally overlapping fluorophores) and the need to maintain a high throughput operation, we decided to limit the pilot study to three cellular probes in addition to the probe for the analyte. The following three probes were selected: (i) DNA stain DAPI for the nucleus, (ii) tubulin as internal control for fixation quality and homogeneity, and (iii) calreticulin for the endoplasmic reticulum.

When preparing samples for immunofluorescence labeling, an important factor is the choice of fixative and detergent to permeabilize the cells. There are many different options available, all with different advantages and disadvantages (11, 12). Several different alternatives to suit a high throughput approach were investigated, and examples of PFA/Triton X-100 and PFA/saponin are shown in Supplemental Fig. 1. Although saponin yielded excellent images for proteins localized in the cytoplasm, our results suggest that stronger detergents, such as Triton X-100, are needed to stain proteins localized to mitochondria or nuclei (data not shown) (11). The Triton permeabilization was therefore chosen as the standard procedure for the high throughput effort. In Supplemental Fig. 2, an example of the split view of the four channels is presented showing staining of a putative mitochondrial protein (A) with probes for nucleus (B), tubulin (C), and endoplasmic reticulum (D).

Annotation of Subcellular Localization—The analysis of three cell lines using four probes was performed for 466 antibodies selected at random from the validated list of antibodies generated within the framework of the protein atlas program. The set of antibodies covers proteins originating from all chromosomes with slightly higher representation on chromosomes 1, 3, 11, 14, 19, 22, and X (see Supplemental Fig. 3). Fig. 2 shows an example of the nine different subcellular localizations annotated in this work, namely cytoplasm, microtubules, nucleus, mitochondria, vesicles, microfilaments, endoplasmic reticulum, Golgi apparatus, and extracellular matrix. The microscopic parameters were set to reach a lateral resolution close to the diffraction limit, i.e. about 200 nm. All images in the atlas were manually annotated using a Web-based annotation tool (Supplemental Fig. 4). The subcellular localization, staining pattern characteristics (e.g. smooth, speckled, etc.), and intensity were manually annotated.

The diversity of staining patterns is illustrated by examples of nuclear staining in Fig. 3. Eight different proteins are shown, all having different distributions within the nucleus (nucleus, nucleoli, and nuclear membrane) as well as varying characteristics of the staining (smooth, granular, speckled, dotty, and clusters of spots). More detailed annotation could provide more information about the function of individual proteins, but the large spectrum of staining patterns makes a more detailed annotation risky and could lead to false positives (13). Therefore the annotations are focused on the main organelles, and to ensure that no information is lost, both the images and annotation comments are accessible for the viewers.

Quality of Antibodies—All HPA antibodies were diluted to a final concentration of 2 μg/ml. To analyze whether the specificity of the staining could be related to the initial antibody concentration, a diagram showing the relation between the antibody stock concentrations and staining specificity is displayed in Supplemental Fig. 6. It is noteworthy that even though the antibody concentration was normalized there were a higher number of unspecific and negative stainings for low stock concentration compared with higher concentration suggesting that high concentrations of specific antibodies after immunization is an indicator of good antibody quality. However, it is worth pointing out that unspecificity of the staining can also be associated with the fixation/permeabilization protocol used as shown previously (11, 12).

Evaluation of Localization—In Fig. 4, the distribution of the different subcellular localization is shown for the three cell lines. A, B, and C show protein localization distributed in three major compartments: nucleus, cytoplasm, and nucleus with

---

2 L. Barbe and H. Andersson-Svahn, unpublished data.
A fourth category groups unspecific and negative staining. Antibodies were annotated as unspecific when there was a homogeneous staining all over the cell, as shown in Supplemental Fig. 5A, in contrast to the compartment “nucleus + cytoplasm” (Supplemental Fig. 5B) where the annotator was able to distinguish different staining intensities within the two localizations. Negative staining was identified when laser gain was above a defined threshold about which the noise level is too high compared with signal intensity. Data are similar between U-2OS (Fig. 4A) and A-431 (Fig. 4C) with about 30% of all antibodies located in the cytoplasmic compartment, 17% located in the nuclear compartment, and 30% located in both. The U-251MG cell line (Fig. 4B) had more proteins with an apparent nuclear localization (43%), and this finding needs to be further investigated. For the three cell lines, an average of 18% of antibodies was considered as unspecific or negative.

A comparison of the concordance of subcellular localization in the three cell lines is shown in Fig. 4D. Because it was expected that many of the protein targets would only be expressed in one or two of the cell lines, a complete match between the cell lines was not expected for many of the antibodies. The comparison of proteins with an exact match of all subcellular localizations (stringent) demonstrated that 68% of the antibodies show identical localization of the target protein in three (23%) or two (45%) cell lines. A comparison in which at least one subcellular localization was overlapping (non-stringent) showed that 82% of the protein targets have identical subcellular localization in at least two cell lines and that almost half of the antibodies (49%) have the same localization in all three cell lines.

Fig. 4E gives a more detailed overview of subcellular localizations. For this figure, only antibodies specific to one organelle were taken into account. Differences among the three
cell lines can be noticed. For example, more antibodies are annotated to the mitochondria compartment in U-2OS cell line (26 antibodies) compared with the other cell lines (21 and four antibodies for U-251MG and A-431, respectively). About 200 antibodies show a strict localization, 200 show localizations in multiple cellular compartments, and about 70 are considered unspecific (or negative). The latter figure can be explained through multiple parameters, independent or interrelated, like sample preparation, antibody specificity in a particular cell line, cell morphology, or annotation.

**Comparisons with Gene Ontology Annotations**—In Fig. 5, localization annotations from this study are compared with the Gene Ontology (GO) annotations in a “score map.” To be able to link the two annotation ensembles, the first database was populated to connect each HPA antibody with its corresponding cellular component GO terms (data are mined from the antibody information pages available on our servers). A second database links our annotation terms (i.e., nuclear, cytoplasm, mitochondrion, etc.) with GO terms from the cellular component category (available at the Gene Ontology Website) based on term similarity. Finally a color score was generated for our 10 subcellular compartments to evaluate a match or mismatch between our annotations and Gene Ontology (Fig. 5).

Fig. 5 enables a direct comparison between our subcellular annotations of protein localization and GO terminology of the gene products. For instance, the antibody cluster denoted by a in Fig. 5A highlights a good agreement for localization into nuclear and cytoplasmic compartments. b illustrates a group of about 20 proteins annotated to the nucleus and cytoplasmic compartments but without any GO annotations for cytoplasmic localization. However, this localization is supported by other GO information based on molecular functions or biological processes. c indicates a group of proteins for which annotations are in good agreement between our study and GO, and c’ shows a mismatch between the two databases highlighting antibodies directed against cytoskeleton proteins as predicted by GO. For some of these antibodies, the sample preparation (especially the fixation step of the immunofluorescence protocol) might have masked antigens or modified protein conformations and ultimately hindered annotations. d represents a group of proteins predicted by the Gene Ontology to belong to the vesicle subcellular compartment. Vesi
cles (e.g., endosomes or clathrins) are by definition small subcellular organelles and thus are more difficult to annotate as described by Glory and Murphy (14). e shows a set of 20 proteins for which there is no existing GO terms and where our results add new localization information. Overall the agreement between GO and our annotations is larger than 80%, calculated as the percentage of proteins in which the GO and the experimental data predict the presence (shown in green in Fig. 5) or absence (light blue in Fig. 5) of a protein in a defined subcellular compartment. It is obvious that comparisons of the nuclear and cytoplasmic compartments are more complex than comparisons for the other compartments (Fig. 5). These are the most often annotated compartments, and here the agreement with GO data was lower (45 and 55%, respectively). One explanation for this discrepancy is the limitations due to using only one method for sample preparation. For example the clusters indicated by c’ and d in Fig. 5A (cytoskeletal and vesicular proteins, respectively) had in the experimental evaluation been annotated as nuclear or cytoplasmic. This might be expected because some types of proteins will not easily be resolved in their intact subcellular compartment and hence fall into a less resolved compartment, i.e., cytoplasm or nucleus. Another reason for this discrepancy is that annotations often take multiple localizations into account. The nuclear and cytoplas-

![Fig. 3. Examples of different nuclear and subnuclear patterns in U-2OS cell lines for the proteins indicated. A, methyl-CpG-binding protein 2; B, RNA-binding motif protein 14; C, zinc finger and SCAN (SRE-ZBP, CTlin51, AW-1[ZNF174], and 18 cDNA or ZnF20) domain-containing protein 1; D, emerin; E, Bcl-2-associated transcription factor 1; F, probable RNA-binding protein 25; G, bromodomain-containing protein 1; H, mucin-4 precursor.](image-url)
mic compartments could in this context be considered as metacompartments including cytoplasmic and nuclear or- ganelles. This will sometimes lead to false mismatches when comparing with predictive data. Interestingly when compiling data from all three cell lines (Fig. 5, A, B, and C), there is a 60% agreement between the GO prediction and the experimental localization determined here (data not shown). The presented data therefore add valuable experimental informa-
tion to the Gene Ontology terminology by evaluating subcel-
lar localization in cells with different phenotypes.

**Integration into the Human Protein Atlas**—The images and subcellular localization data resulting from this project has been released as a new feature of the currently available Human Protein Atlas. In Fig. 6, a screen shot of the layout of the

---

**Fig. 4.** Distribution of localization into broad compartments (nucleus, cytoplasm, nucleus + cytoplasm, and unspecific + negative) for U-2OS cell lines (A), U-251MG (B), and A-431 (C). D, localization agreement for the three cell lines. Overall statistics are separated in two categories. Non-stringent condition represents a group of antibodies with at least one match in a subcellular compartment. Stringent condition only includes antibodies with the exact same localization pattern. The two categories have three different criteria to calculate the concordance among the three cell lines: agreement in all three cell lines, agreement in two cell lines; agreement in any cell line. *, this calculation includes the case where three cell lines are in concord.) E, subcellular localization distribution for the three cell lines. ER, endoplasmic reticulum; ECM, extracellular matrix.
subcellular atlas is shown. The objective is to include this analysis for every antibody/protein included in the atlas in the future.

**DISCUSSION**

Here we describe a pilot study for the systematic localization of human proteins with confocal microscopy. A single fixation/permeabilization protocol was used, and a manual annotation was conducted with the aid of three organelle probes, representing nucleus, cytoskeleton, and endoplasmic reticulum. Over the past decade, automated image systems for cell-based assays have been developed, often referred to as high content screening or high throughput microscopy systems, and these are increasingly being used for basic research on biological pathways (15–17). However, systems with ultrahigh throughput capacities are often lacking the single cell or subcellular resolution. Therefore, we used a conventional confocal microscope and manual image capturing to generate a subcellular atlas with high spatial resolution.

In the past decade, fluorescence microscopy imaging has gone through a dramatic change due to development of new types of fluorescent probes, new advancement in imaging instrumentation (18, 19), and new applications (3, 20). All proteomics studies using cells and imaging rely on three major steps: sample preparation, image acquisition, and data analysis. For the first step, two different techniques are available: live cell studies (21) or immunostaining of fixed cells (16). An advantage of live cell studies compared with fixed cells is the possibility to study temporal patterns of protein expression, which is a parameter of importance in cell behavior. The cell lines used in this study do not have their cell cycles synchronized, meaning that at a fixed time point the cell population exhibits cells randomly distributed in G1, S, G2, and M phases. Manual image acquisition alleviates this problem by choosing a representative region of interest. An attractive expansion of the work described here could therefore be to include specific cell cycle markers (e.g., DNA probe, bromo-2-deoxyuridine incorporation, and cyclin expression) to bring a pseudodynamic input for protein localization (22). New imaging techniques, such as imaging-bleaching cycles could also be use on selected sets of proteins to study protein networks and interactions (23).

Here we performed manual annotation on each image to classify antibody staining patterns into different subcellular compartments. As shown in a recent review, human and computer accuracy scores are comparable for large or-

![Score map comparing localization information from our study and Gene Ontology annotations.](image-url)

**Fig. 5.** Score map comparing localization information from our study and Gene Ontology annotations. The number of antibodies, on the y axis, represents the number of antibodies with a defined localization. The cell lines represented are U-2OS in A, U-251MG in B, and A-431 in C. The scores for the different subcellular compartments are shown in the following order on the x axis: cytoplasm (cyto), nucleus, nucleolus (nucleo), nuclear membrane (n.m), mitochondrion (mito), endoplasmic reticulum (ER), Golgi apparatus, cytoskeleton (cytosk), vesicles, and extracellular matrix (ecm). Dark blue indicates no available GO cellular component information but experimental data from our study. Light blue indicates no protein localization for a specific compartment in both databases, and green indicates when both databases agree on the localization. Dark red indicates conflicting subcellular localization, and orange indicates a localization in a particular compartment not predicted by GO. Thus, supporting data are shown in light blue and green, whereas conflicting data are shown in orange and dark red. Dark blue indicates that there is no available information from Gene Ontology, but experimental data exist from our study. The clustering on the y axis is merely performed to enable the identification of groups of proteins localized in the same compartment and may thus be different for the three cell lines.
ganelles like mitochondria, endoplasmic reticulum, or cytoskeleton (14). However, human annotation is labor-intensive and is susceptible to annotator subjectivity, and automated image analysis is therefore attractive. Such analysis can be performed using two different principles: (i) co-localization or (ii) pattern recognition. For co-localization analysis, organelle-specific markers are needed for every subcellular compartment to be analyzed, and three-dimensional image acquisitions are desired, making such studies difficult and demanding. A pilot study has been conducted on a reduced set of organelle markers demonstrating the feasibility of such an approach, although more technical development is needed to implement such a strategy in a streamlined, high throughput scheme. The pattern recognition strategy has shown feasibility and accuracy in large scale studies (24–26). It can be divided into two groups, supervised and unsupervised classifications. Supervised learning requires an initial training phase before the subcellular location class of the protein can be determined. Unsupervised learning methods are able to define classes based on the distance of objects in a feature space (27). These different approaches are currently being investigated to enable automated image annotation.

Here our manual annotations of 466 antibodies led to more than 80% classified in one or multiple subcellular compart-

Fig. 6. Preview of the confocal Web-based atlas to be added to the existing protein atlas. The first level shows all information available for an antibody. Clicking in the last section “cell lines – immunofluorescence” a new page (second level) appears and gives detailed localization information for the chosen cell line in addition to the fluorescence images. By selecting one image, it opens (third level) another window displaying a high resolution picture where the different channels can be switched on and off. IHC, immunohistochemistry; IF, immunofluorescence.
ment(s) and about 20% left with undefined localization. Several factors can be responsible for these non-localized antibodies such as (i) the lack of the protein target in the analyzed cell lines, (ii) the failure to permeabilize the compartment in which the protein target is present, or (iii) the failure to recognize the epitope of the protein target due to denaturation of the corresponding protein by the fixation (detergent). More work is needed to distinguish between these alternatives including the use of an alternative antibody recognizing the same target but recognizing another epitope.

As reviewed recently (14), there is a need for additional experimental determination of protein localization, and this project attempts to contribute to this effort. The annotated images in this atlas classify proteins into subcellular compartments and can be compared with existing databases, such as the Gene Ontology consortium, Proteome Analyst WoLF PSORT, or hum-PLOC (see Nucleic Acids Research for an extensive list). The data in this pilot study were evaluated on the degree of agreement in protein localization using the GO database. The results show the usefulness of complementing theoretical prediction methods with experimental procedures, such as the antibody-based method presented or with a fluorescent tagging method as demonstrated earlier (15).

The scope of proteomics is broad from identification and quantification of proteins in cells and tissues (in normal and cancer states) to deeper understanding of molecular processes within cells. The recent addition of cytomics to the “omics” data sets brings knowledge of cell systems and molecular architecture thanks to imaging techniques and molecular biology. The presented subcellular atlas aims to provide a publicly available atlas for subcellular localization of all human proteins and to facilitate studies comparing normal and disease protein profiles with the ultimate objective to discover new biomarkers and targets for the design of diagnostic tools and drugs.

* This work was supported by The Royal Academy of Sciences. The Human Proteome Resource program is funded by 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 be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

** The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

To whom correspondence should be addressed. E-mail: mathias@biotech.kth.se.

REFERENCES


19. Oheimg, M. (2007) High-throughput microscopy must re-invent the micro-
Toward a Confocal Subcellular Atlas of the Human Proteome

scope rather than speed up its functions. Br. J. Pharmacol. 152, 1–4