Contribution of Antibody-based Protein Profiling to the Human Chromosome-centric Proteome Project (C-HPP)

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · DECEMBER 2012

Impact Factor: 4.25 · DOI: 10.1021/pr300924j · Source: PubMed

32 AUTHORS, INCLUDING:

Cristina Al-Khalili Szigyarto
KTH Royal Institute of Technology
25 PUBLICATIONS **1,736** CITATIONS
SEE PROFILE

Peter Nilsson
KTH Royal Institute of Technology
143 PUBLICATIONS **6,083** CITATIONS
SEE PROFILE

Jochen M Schwenk
KTH Royal Institute of Technology
67 PUBLICATIONS **1,358** CITATIONS
SEE PROFILE

Mathias Uhlen
KTH Royal Institute of Technology
616 PUBLICATIONS **28,509** CITATIONS
SEE PROFILE

Journal of proteome research

Article

pubs.acs.org/jpr

Contribution of Antibody-based Protein Profiling to the Human Chromosome-centric Proteome Project (C-HPP)

Linn Fagerberg,[†,‡] Per Oksvold,[†,‡] Marie Skogs,[†] Cajsa Älgenäs,[§] Emma Lundberg,[†] Fredrik Pontén,[∥] Åsa Sivertsson,[†] Jacob Odeberg,[†] Daniel Klevebring,[†] Caroline Kampf,[∥] Anna Asplund,[∥] Evelina Sjöstedt,[∥] Cristina Al-Khalili Szigyarto,[§] Per-Henrik Edqvist,[∥] IngMarie Olsson,[∥] Urban Rydberg,[∥] Paul Hudson,[§] Jenny Ottosson Takanen,[§] Holger Berling,[§] Lisa Björling,[†] Hanna Tegel,[§] Johan Rockberg,[§] Peter Nilsson,[†] Sanjay Navani,[⊥] Karin Jirström,[#] Jan Mulder,[¶] Jochen M. Schwenk,[†] Martin Zwahlen,[†] Sophia Hober,[†] Mattias Forsberg,[†] Kalle von Feilitzen,[†] and Mathias Uhlén*[,†,§]

[†]Science for Life Laboratory, KTH-Royal Institute of Technology, Stockholm, Sweden

[§]School of Biotechnology, AlbaNova University Center, KTH-Royal Institute of Technology, Stockholm, Sweden

[∥]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

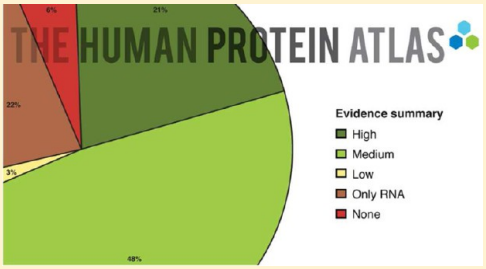
[⊥]Lab Surgpath, Mumbai, India

[#]Department of Clinical Sciences, Pathology, Lund University, Sweden

[¶]Science for Life Laboratory, Karolinska Institute, Stockholm, Sweden

Supporting Information

ABSTRACT: A gene-centric Human Proteome Project has been proposed to characterize the human protein-coding genes in a chromosome-centered manner to understand human biology and disease. Here, we report on the protein evidence for all genes predicted from the genome sequence based on manual annotation from literature (UniProt), antibody-based profiling in cells, tissues and organs and analysis of the transcript profiles using next generation sequencing in human cell lines of different origins. We estimate that there is good evidence for protein existence for 69% (n = 13985) of the human protein-coding genes, while 23% have only evidence on the RNA level and 7% still lack experimental evidence. Analysis of the expression patterns shows few tissue-specific proteins and approximately half of the genes expressed in all the analyzed cells. The status for each gene with regards to protein evidence is visualized in a chromosome-centric manner as part of a new version of the Human Protein Atlas (www.proteinatlas.org).



KEYWORDS: *Antibody-based protein profiling, C-HPP*

INTRODUCTION

The success of the Human Genome project has demonstrated the power of systematic efforts to map the building blocks of man and shown a path of pursuit for "omics" approaches to study biology and medicine using high-throughput methods. A significant challenge for the future is to use this genomic knowledge base to gain a proteome-wide understanding of the protein products encoded by the genome. Recently, a Human Proteome Project has been proposed to systematically map the human proteins in a gene-centric manner,[1] suggesting that an international effort should be initiated to characterize at least one representative isoform from every protein-coding gene using a multitude of technology platforms, including mass spectrometry and antibody-based characterization.

One of the issues for such an effort is the complexity of the human proteome, including temporal and spatial differences, transient and stable interactions and the vast amounts of isoforms and protein variants. In addition, it is difficult to define clear end-points that can be accepted throughout the scientific community.[2,3] We have earlier described the Human Protein Atlas effort to, in a gene-centric manner, map the human proteome using an antibody-based approach with the objective to generate publicly available subcellular localization data and expression data for most major human tissues and organs.[4−6] Here, we describe how this effort can contribute to a Human Proteome Project. The human protein-coding genes predicted by the Ensembl team[7] and the manual annotation performed by the UniProt Consortium[8] have been used as a starting point for an analysis based on antibody-based profiling, including Western blot analysis, immunohistochemistry and immunofluorescence-based confocal microscopy. We also include transcriptomics data from a few selected human cell lines of different origin as a pilot

Special Issue: Chromosome-centric Human Proteome Project

Received: September 30, 2012
Published: December 31, 2012

to demonstrate the values of combining data from RNA and proteins. The integration of data from various sources has allowed us to generate a score for each of the 20251 putative human genes on all the 24 chromosomes describing the level of evidence for the existence of the corresponding protein(s).

On the basis of these new features, a new version of the Human Protein Atlas (www.proteinatlas.org) is launched with protein evidence score for all putative genes with links to the experimental evidence. We have added a new chromosome-centered visualization of all protein-coding genes based on UniProt, Protein Atlas and transcriptomics data. All data is publicly available to facilitate complementary efforts to characterize the proteome components, including protein isoforms, subcellular localization, and distribution profiles in cells, tissues and organs, for all genes covering all human chromosomes. In this manner, the antibody-based profiling can be complemented with other experimental results from various sources, including mass spectrometry methods and gene tagging technologies, to deliver integrated knowledge-based resources covering all human protein-coding genes.

RESULTS

Transcriptomics in Human Cell Lines

An important aspect when evaluating the evidence for the existence of gene-products on a protein-level is to study the corresponding transcripts using quantitative RNA profiling. For this reason a transcriptomics study was included, based on next generation sequencing of eight human cell lines, selected on the bases of originating from diverse cellular phenotypes and being frequently used as model systems for human biology (Table 1).

Table 1. Cell Lines Used for the Transcriptomics Study[a]

cell line	description	no. of detected genes	no. of cell-specific genes	reference
MCF-7	Metastatic breast adenocarcinoma cell line	13575	61	This work
A-549	Lung carcinoma cell line	13817	46	This work
CACO-2	Colon adenocarcinoma cell line	13692	100	This work
Hep-G2	Hepatocellular carcinoma cell line	13653	102	This work
HeLa	Cervical epithelial adenocarcinoma cell line	13839	77	This work
RT-4	Urinary bladder transitional cell carcinoma cell line	14087	146	This work
PC-3	Metastatic poorly differentiated prostate adenocarcinoma cell line	13866	74	This work
HEK 293	Embryonal kidney cell line, transformed by adenovirus type 5	14308	113	This work
U-2 OS	Osteosarcoma cell line	13869	144	Lundberg et al.[10]
U-251 MG	Glioblastoma cell line	13438	11	Lundberg et al.[10]
A-431	Epidermoid carcinoma cell line	13090	54	Lundberg et al.[10]

[a]Total number of detected genes as well as the number of cell-specific genes are shown. The data for the first eight cell lines are new for this study, while the data for the last three cell lines have been described in Lundberg et al.[10].

The amount of transcript from each putative protein-coding gene in all of the eight human cell lines was calculated by an RNA-Seq approach.[9] This data was integrated with transcriptomics data published previously[10] for three additional human cell lines. The results, summarized in Table 1, show that each of the studied cell lines had detectable transcripts corresponding to approximately 14000 genes. This is in line with previous studies showing that human cell lines have detectable levels of transcripts to between 13000 and 15000 of the protein-coding genes.[10,11]

Analysis of biological replicates showed a high correlation for all eight cell lines (Supplementary Figure 1, Supporting Information) with a Spearman correlation between the biological replicates from 0.96 to 0.98. The high correlations suggest good technical and biological reproducibility between samples. Detecting transcripts from the protein-coding part of a putative gene is obviously supportive evidence for the existence of the corresponding protein, but cannot be used alone to claim protein evidence with high reliability. Hebenstreit et al.[12] recently suggested that many low abundant transcripts in mammalian cells, in particular genes with an expression value of RPKM value less than 1, are not translated into stable proteins. The results summarized in Figure 1a show that even if genes with an expression value of RPKM lower than 1 are excluded, transcripts for approximately three-quarters (76%) of the genes can be found in at least one of the 11 cell lines (Figure 1a). This supports earlier suggestions[13] that a large fraction of genes are expressed in human cells.

Cell Line Specificity Based on Transcript Profiling

Transcript profiling is an attractive tool to study cell-specific expression in the different cell lines. In Figure 1b, the number of the putative genes expressed in 0, 1, 2 etc up to all 11 cell lines are shown. Approximately half of the genes (n = 10078) are "house-keeping", defined as genes with detectable transcripts in all analyzed cells, and only 5% of the genes are cell-type specific with detectable expression in only a single cell line. Few genes (n = 2798) were not detected in any of the analyzed cell lines, a surprisingly low number considering that these eleven adherent cell lines only covers a small fraction of the functionality of the human body, but is consistent with earlier studies suggesting that human cells express very few cell- or tissue-specific proteins.[13] In Table 1, the number of cell-type specific genes in each of the human cell lines is presented. Approximately 100 genes are specific for each of the cell lines corresponding in each case to less than 1% of the total human genes.

We have earlier made the striking observation that many of the house-keeping genes not only display a broad expression pattern, but also normally have relatively high expression levels in the different cell types, while genes displaying a tissue-specific pattern are often expressed at lower relative level.[10] An analysis of the eight cell lines supports this hypothesis showing significant higher average RNA levels for the genes detected in all eight cell lines as compared to the genes detected in less cells (Figure 1c). A more in-depth analysis showing the differential level of expression for all the 16901 genes detected in at least one the eight cell lines (Figure 1d) confirms the conclusion that cell-type specific genes show lower abundance then the house-keeping genes.

Protein Evidence Summary

The UniProt team with nodes in Switzerland, UK and USA[8] has during many years manually curated literature for protein evidence for many species, including humans. This effort has resulted in one of the worlds most visited portals for biological data showing protein evidence coupled with references to facilitate further in-depth studies of each protein. For humans, the latest version of UniProt (2011_08) contains 20244 genes

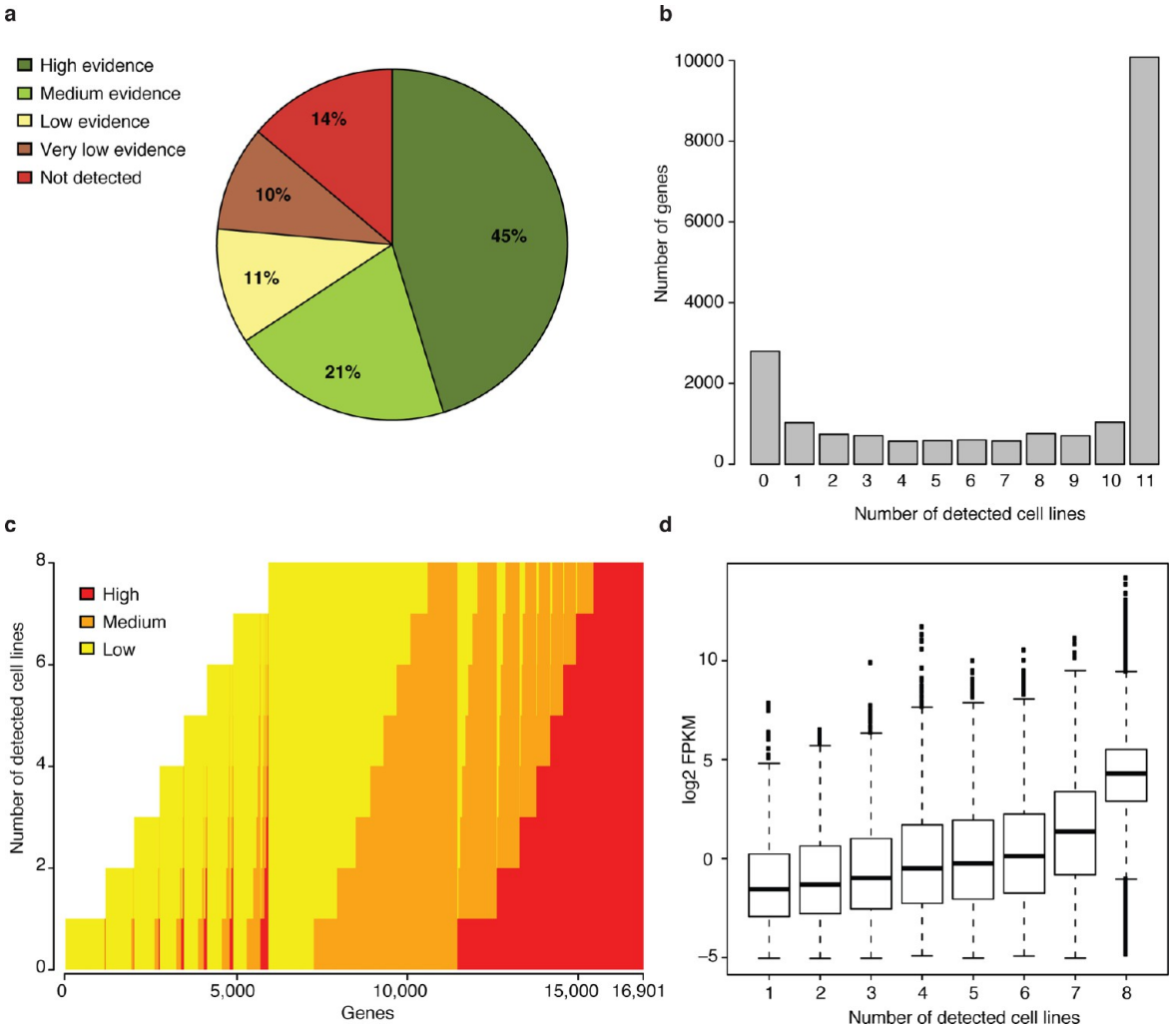


Figure 1. Transcriptomic profiling of human cell lines. (a) Distribution of genes in categories based on the combined set of RNA-seq data from 11 human cell lines (Table 1). For each gene, the category is defined by the most highly abundant cell line. For the low category, a subgroup "very low" was used for the genes with all FPKM/RPKM values <1. (b) Distribution of the number of genes detected across different numbers of cell lines, ranging from 0 to 11. 50% (n = 10078) of all genes are detected in all 11 cell lines, whereas 14% (n = 2798) are not detected at all and 5% (n = 1028) are only detected in a single cell line. (c) Barplot of the number of detected cell lines for each of the 16901 genes detected in at least one of the eight cell lines analyzed here. Each bar also shows the corresponding abundance class for each gene and each of the detected cell lines calculated as the bottom/middle/top third of all positive FPKM values, respectively, for a given cell line. Genes are arranged according to the total number of detected cell lines with genes only detected in a single cell line to the far left and genes detected in all eight cell lines to the right. (d) Boxplots showing the distribution of log2 FPKM values across genes detected in the different number of cell lines ranging between 1 and 8, using all positive FPKM values from the eight cell lines analyzed using Illumina sequencing.

with protein evidence for 13524 genes and RNA-based evidence for an additional 5789 genes.

The Human Protein Atlas team with nodes in Sweden, India, Korea and China has generated more than 11 million high-resolution images covering a large portion of the organs and tissues in the human body and each of the immunohistochemical images has been manually curated using a certified pathologist. Each antibody has been assigned a validation score based on the staining pattern along with support from bioinformatics data and/or previous literature. For all proteins mapped and characterized using two or more antibodies, a knowledge-based annotated protein expression score have been implemented. This annotation score describes the extent to which the staining of one antibody is validated by the staining of the other antibody binding a separate region of the target protein.[6] Similarly, each antibody has been analyzed using Western blot assays using a standardized format containing two human cell lines, a pool of plasma samples and extracts from liver and tonsils. Based on the results of 11232 genes, a protein evidence score was calculated based on the manual curation of each gene using Western blot, immunohistochemistry and/or immunofluorescence data, as described in the Methods section. The analysis yields 7429 genes with good (high or medium) evidence for protein existence based on the Protein Atlas alone.

A new protein evidence summary has been determined for all genes predicted from the genome sequence based on manual annotation by (i) UniProt, (ii) antibody-based profiling in cells, tissues and organs, and (iii) analysis of the transcript profiles using next generation sequencing in the 11 human cell lines of different origins described here. A classification for each gene is calculated as described in detail in the Material and Methods section. To reach the highest classification, the gene must have evidence on the protein level as determined by UniProt as well as having a high reliability score by the antibody-based Protein Atlas

dx.doi.org/10.1021/pr300924j | J. Proteome Res. 2013, 12, 2439−2448

a

b

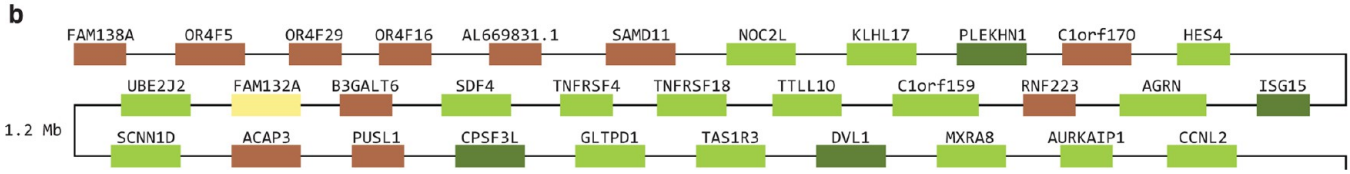


Figure 2. Protein evidence based on UniProt, the Human Protein Atlas and transcriptomics analysis in cell lines. Visualization of a part of the human chromosome 19 is shown with the status for protein evidence summary shown in color code.

data. The classification for evidence at the RNA-level takes into account both the results from the cell line analysis performed here (Table 1 and Figure 1), but also the classification on evidence on the transcript level reported in literature and manually annotated by UniProt. The classification on an individual gene level can be obtained at a new version of the Human Protein Atlas (www.proteinatlas.org).

Chromosome-specific Protein Evidence

The Human Protein Project has been launched to characterize the human proteins in a chromosome-centered manner.[14] As a contribution to this project, four different views of the evidence status for every gene of each chromosome has been visualized as part of the Protein Atlas, including (i) the UniProt classification, (ii) the antibody-based Human Protein Atlas experimental data, (iii) the cell line transcript profiling data and (iv) the aggregated evidence summary. The protein evidence score for each of the putative genes for all chromosomes is shown, ordered by their relative position on the chromosome. An example of the protein evidence summary view for part of chromosome 19 is shown in Figure 2. Note that for several of the genes on the far end on the chromosomes, no or little evidence for existence can yet be found, suggesting that more in-depth analysis is needed to determine if these genes are pseudogenes or genes coding for functional proteins.

A summary of the classification of all predicted genes of the human genome is presented in Figure 3a, demonstrating that there is good protein evidence (high or medium) for 69% (n = 13971) of the protein-coding genes, while 22% (n = 4421) of the genes have only evidence on the RNA level and 6% (n = 1311) still lack experimental evidence. In addition, 3% (n = 548) of the genes have a low evidence score, meaning no evidence on the protein level by UniProt, but evidence by the antibody-based profiling with medium reliability. In summary, the classification shows that there is evidence on the protein or RNA level for more than 90% of the putative protein-coding genes. However, more studies are needed to confirm the existence of these putative protein products with no or only indirect evidence (transcripts).

The overall status for all genes on the various human chromosomes has been summarized in Figure 3b. There is good protein evidence for a majority of the genes on all human chromosomes and the genes with no protein evidence or only evidence on the RNA level are distributed relatively evenly over the 24 chromosomes. The corresponding evidence summary for the UniProt classification, the antibody-based Human Protein Atlas experimental data and the cell line transcript profiling data have also been summarized as part of the "Chromosome progress" page at the Human Protein Atlas portal (www.proteinatlas.org).

Subcellular Specificity

An important mission of the Human Proteome Project is to determine the subcellular distribution of all the human proteins. An effort has therefore been undertaken to map all the proteins in selected human cell lines using the antibodies generated within the framework of the Human Protein Atlas effort. We have earlier described an analysis of 4,005 gene products with regards to subcellular localization in 16 annotated subcellular localizations, including mitochondria, nucleus, ER, Golgi and plasma membrane[15] and this has here been extended to 6515 manually curated genes. In Figure 4a, a Cytoscape[16] network plot is presented with the number of proteins found in various subcellular locations. The analysis shows that the locations with most proteins presented are the cytoplasm and nucleus followed by mitochondria and vesicles, but interestingly a large fraction of the proteins are found in more than one location.

The profiling has provided evidence for subcellular localization for a large number of proteins with no previous evidence at the protein level or with a strong indication for another localization that has been previously reported in literature. As an example, Figure 4b and c shows that the protein products of the genes C14orf119 and C6orf64 is localized to mitochondria and vesicles, respectively. Similarly, three antibodies targeting different epitopes of the gene product of the gene C9orf78 all suggest a nuclear localization of this unknown protein (Figure 4d). Two antibodies toward the AHNAK nucleoprotein (Figure 4e) and AHNAK nucleoprotein 2 (Figure 4f), respectively, suggest that these

dx.doi.org/10.1021/pr300924j | J. Proteome Res. 2013, 12, 2439−2448

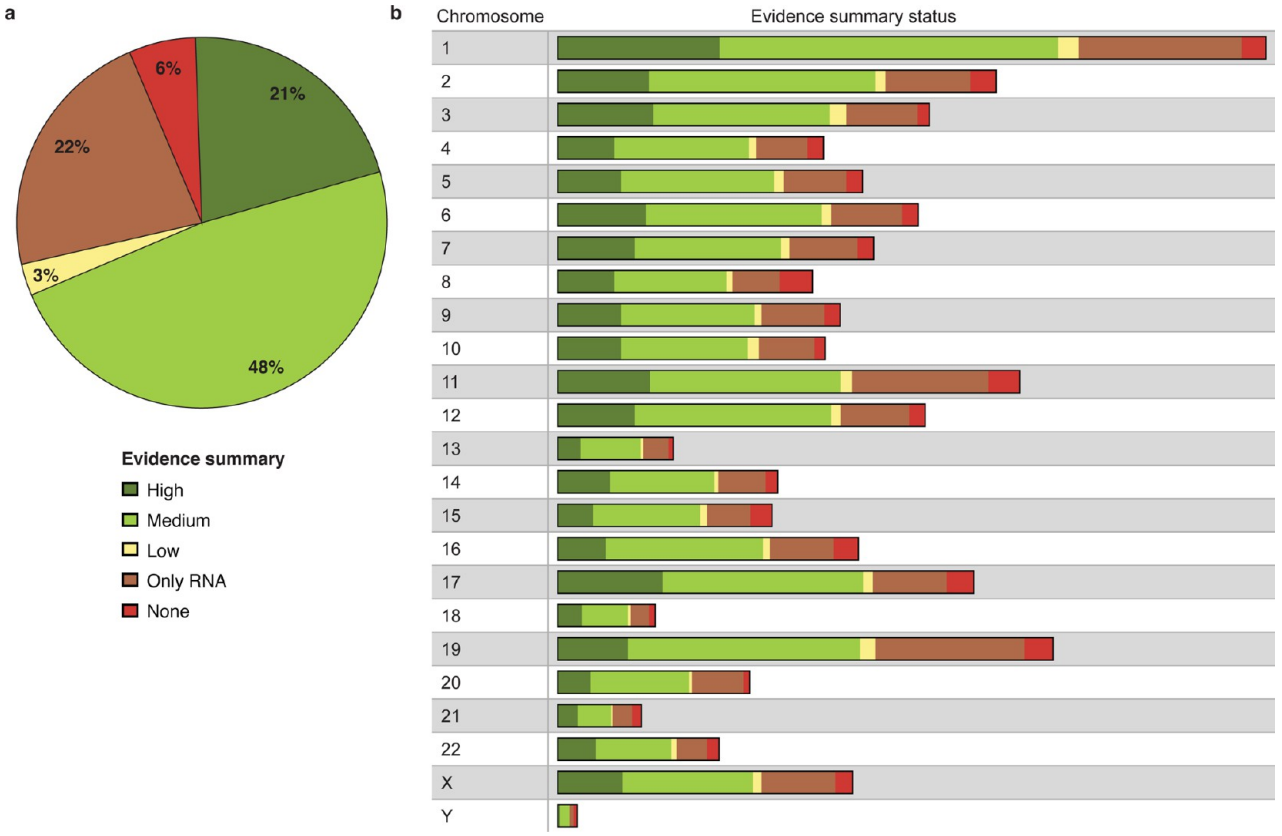


Figure 3. Protein evidence status for all human chromosomes. (a) Fraction of human genes with various protein evidence scores. (b) Protein evidence status based on three underlying data sources are shown for each of the chromosomes.

so-called nucleoproteins are localized to the cytoplasm and plasma membrane and that the expression of AHNAK2 varies in a cell cycle dependent manner.

Tissue Specificity

Another important issue for the characterization of the human proteins is to determine the cell-specificity in various tissues and organs. In Figure 5a, the tissue-specificity of the Human Protein Atlas (version 8) is shown, based on 11260 genes in 66 cell types corresponding to most major tissues and organs in the human body, including heart, lung, GI-tract, hematopoetic cells, liver, kidney and brain. The results suggest that very few proteins are tissue-specific supporting earlier studies[13] and this is further supported by the analysis of data from 2854 genes with annotated protein expression based on two or more paired antibodies for each protein target (Figure 5b).

Despite the low number of cell type specific proteins, we have found many proteins with interesting tissue-specific patterns, for which no previous evidence at the protein level exist. A few examples of these are shown in Figure 5c. In each case, the tissue profiles have been analyzed with at least two antibodies to generate an annotated expression profile[6] and most of the examples have no previous evidence on the protein level. TEX1010 protein shows restricted expression to maturing spermatocytes in testis. Rho GTPase activating protein 28 (ARHGAP28) shows expression restricted to few distinct cell compartments including spermatocytes and spermatids in the inner layers of seminiferous ducts of testis and a sub set of cells in kidney glomeruli. Chromosome 1 open reading frame 114 (C1orf114) shows restricted expression to cilia in ciliated cells from various locations including respiratory mucosa in nasopharynx and bronchi, ciliated cells lining the epididymis in males and the fallopian tube in females. Chromosome 9 open reading frame 11 (C9orf11) shows cell type specific expression pattern with positivity restricted maturing spermatids, with a "cap-like" subcellular distribution pattern resembling that of the acrosomes. Alpha-kinase 1 (ALPK1) shows ubiquitous cytoplasmic expression in various cell types including respiratory and glandular epithelia, mesenchymal cells, glomeruli and tubules of kidney and neurons in the CNS. Chromosome 12 open reading frame 63 (C12orf63) showed ubiquitous expression in all cell types, with variable intensities of cytoplasmic positivity. Pronounced granular cytoplasmic expression was found in neurons of the cerebral cortex. Enoyl CoA hydratase domain containing 2 (ECHDC2) shows similar pattern of selective expression in liver and kidney with some additional expression in glandular cells of the GI-tract. Expression appears cytoplasmic with a granular structure resembling a mitochondrial expression pattern. Ermin, ERM-like protein (ERMN) shows similar expression in selected cell types, with highest expression in glial cells and white matter of the CNS, demonstrated in myelinated nerve bundles present in the lateral ventricle wall. AHNAK nucleoprotein 2 (AHNAK2) has been reported to show evidence at protein level as a nuclear protein,[17] but we instead find a selective cytoplasmic expression pattern with highest expression in squamous epithelia, such as skin and respiratory epithelia, including a clear cytoplasmic expression pattern of this protein by immunofluorescence (data not shown).

The examples shown here and other proteins with similar narrow expression profile are obviously interesting starting points for further studies to study human biology and diseases. To facilitate such studies, the tissue-specific pattern for each gene analyzed using the antibody-based effort is summarized in the

dx.doi.org/10.1021/pr300924j | J. Proteome Res. 2013, 12, 2439−2448

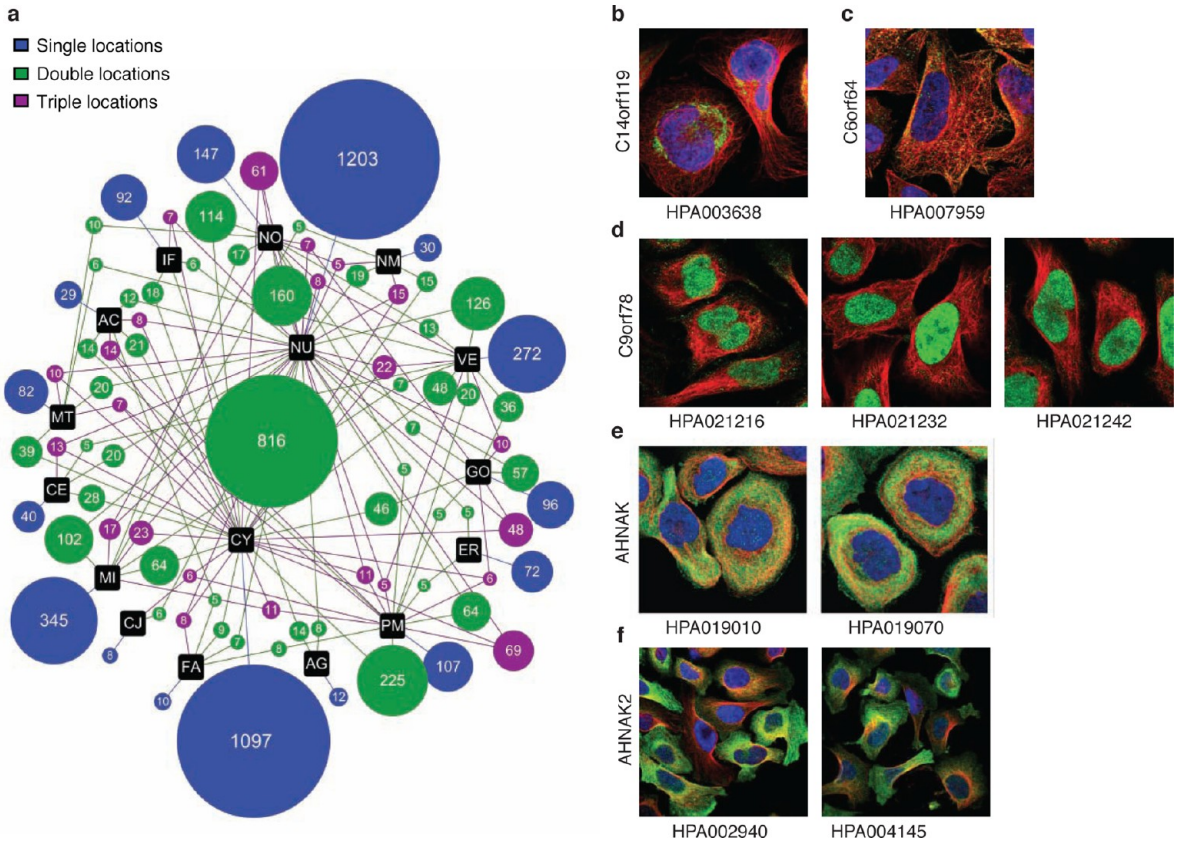


Figure 4. Distribution of subcellular locations and subcellular specificity of proteins with no previous localization evidence. (a) Cytoscape plot showing the distribution of 6515 proteins localized in the human cell line U-2 OS. Each circle represents a combination of subcellular compartments and the size is related to the number of proteins detected in a particular combination, indicated by the number in the middle. Node and edge colors represent single (blue), double (green) and triple (purple) locations. Subcellular compartments are represented by rounded rectangles in black and a two letter code: AC, actin filaments; AG, aggresome; CE, centrosome; CJ, cell junctions; CY, cytoplasm; ER, endoplasmic reticulum; FA, focal adhesions; IF, intermediate filaments; GO, Golgi apparatus; MI, mitochondria; MT, microtubules; NM, nuclear membrane; NO, nucleoli; NU, nucleus; PM, plasma membrane; VE, vesicles. (b–f) Some examples of proteins with no previous localization evidence. The nucleus is shown in blue, microtubules in red and the protein of interest in green. (b) Confocal image of C14orf119 in U-2 OS cells. (c) Confocal image of C6orf64 in U-2 OS cells. (d) Confocal images of C9orf78 targeted by three independent antibodies in U-2 OS cells. (e) Confocal images of AHNAK in A-431 cells targeted by two independent antibodies. (f) Confocal images of AHNAK2 in U-2 OS cells targeted by two independent antibodies.

column "Tissue-specificity (IH)" (Figure 2a). The protein expression across 66 cell types in 48 tissues and organs in the human body is summarized for 11260 genes covering 72% of the human protein-coding genes.

DISCUSSION

Here we have described a new protein evidence classification of all the putative protein-coding genes predicted from the human genome sequence. The classification is based on manual annotation of the literature (UniProt), antibody-based profiling as part of the Human Protein Atlas effort and transcriptomics data generated as part of this work of selected human cell lines. The classification suggest that there is solid evidence for approximately two-thirds (69%) of the gene products predicted by the Ensembl effort and evidence on a transcript level for an additional 22% of the genes. There are thus only 1859 genes (9%) with no evidence yet for their existence as proteins. In the new version of the Human Protein Atlas (www.proteinatlas.org) portal, it is possible to visualize the protein evidence classification in a chromosome-centered manner to allow a rapid inspection of the status for all genes on a chromosome or around a specific gene or region. This visualization tool will be updated with every new version of the Protein Atlas to provide a more and more accurate description as more data are collected.

The classification and conclusion from this study depends on the integration of data both from the study of RNA and the corresponding proteins. The quantitative transcriptomics approaches using next generation sequencing technology (RNA-seq) provide an attractive tool for expression studies, including analysis of cell and tissue specificity. It is important to point out that cell lines are releatively poor models for tissues, but the advantage with cell lines is that the analysis can be performed on relatively homogeneous cells, which is important for this work. RNA analysis of tissues are not easy due to the heterogeneity of the cell types in a particular tissue or organ, which makes it difficult to identify cell specific proteins, but we are exploring many different ways to solve the heterogeneity problems. In the meanwhile, we feel most comfortable with only using the homogeneous cell lines as a basis to classify the protein-coding genes.

Here, we have only used data from a limited number of human cell lines, but obviously it is important in the future to extend this analysis to many more cell lines and defined tissues and cells from normal and disease individuals, as well as developmental stages. The existence of transcript profiling resources, such as ArrayExpress[18] and NCBI Gene Expression Omnibus,[19] integrating

dx.doi.org/10.1021/pr300924j | J. Proteome Res. 2013, 12, 2439–2448

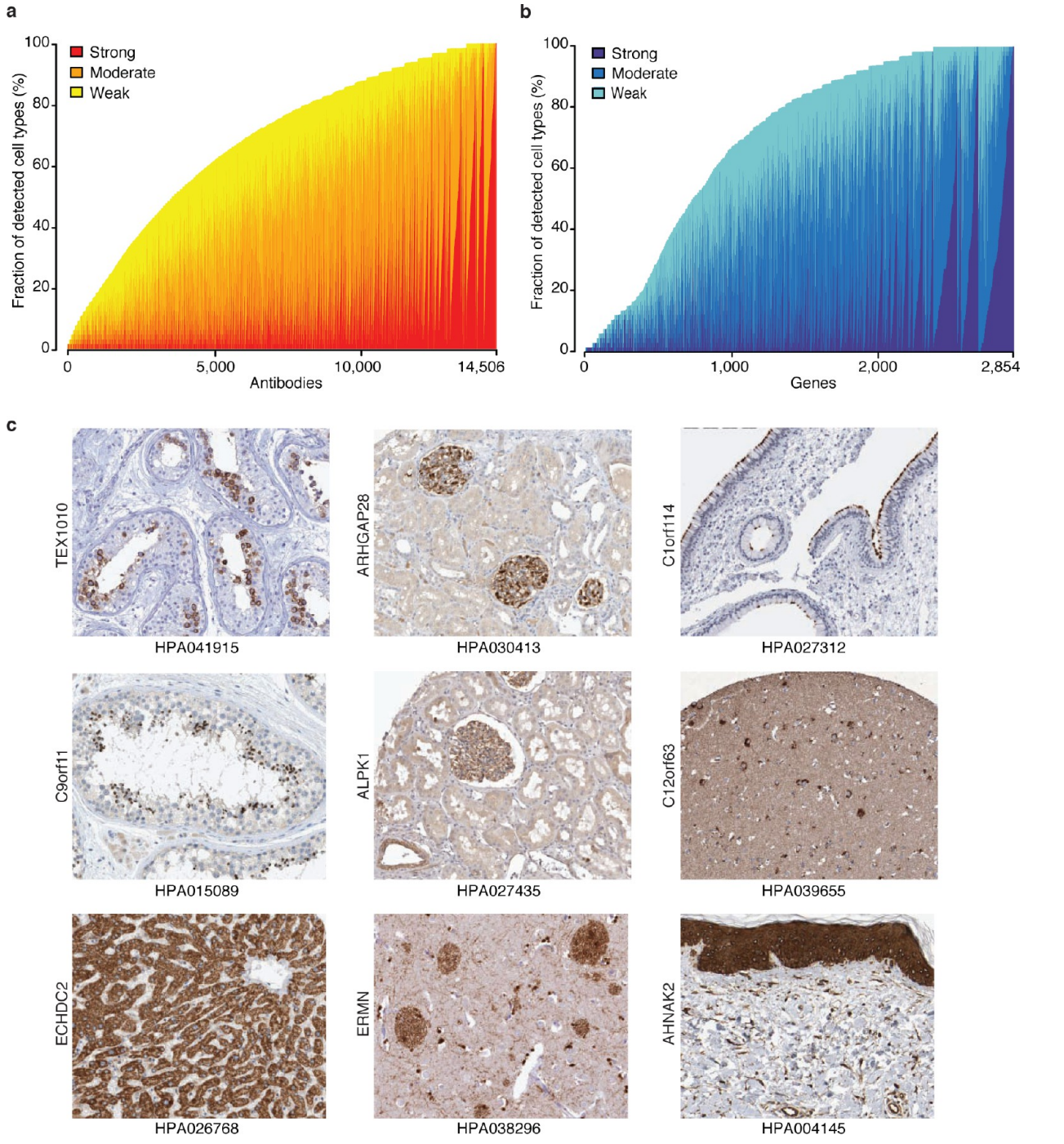


Figure 5. Tissue specificity of proteins in the Human Protein Atlas. (a) The fraction (%) of cell types with the antibody staining levels weak, moderate and strong for 11260 genes detected in at least one of the total 66 analyzed cell types. (b) The fraction of cell types with annotated expression levels weak, moderate and strong for 2854 genes with paired antibodies. Genes are arranged according to the total number of detected cell types with genes detected in all analyzed cell types to the right. (c) Tissue-specific expression of human proteins with no previous evidence on the protein level. In each case, the tissue profiles have been analyzed with at least two antibodies to generate an annotated expression profile. Only one of the antibodies are shown in the figure. Protein profiling in tissues shows expression of TEX101 protein in maturing germinal cells of the testis, Rho GTPase activating protein 28 (ARHGAP28) protein in glomeruli of the kidney, Chromosome 1 open reading frame 114 (C1orf114) protein in cilia of respiratory mucosa, Chromosome 9 open reading frame 11 (C9orf11) protein in maturing spermatids, Alpha-kinase 1 (ALPK1) protein in glomeruli of the kidney, Chromosome 12 open reading frame 63 (C12orf63) protein in neurons of the cerebral cortex, Enoyl CoA hydratase domain containing 2 (ECHDC2) protein in hepatocytes of the liver, Ermin, ERM-like protein (ERMN), in glial cells and white matter of the CNS, and AHNAK nucleoprotein 2 (AHNAK2), in squamous epithelia.

publicly available data from the research community are important assets in this context. However, despite the usefulness for researchers trying to unravel the existence and functionality of the protein-coding genes, it is important to point out that the ultimate study of the consequence of the gene expression should preferably be done on the protein level, including the analysis of localization, modifications, local concentration and interactions in space and time for each protein and its isoforms.

dx.doi.org/10.1021/pr300924j | J. Proteome Res. 2013, 12, 2439−2448

The study of the complete set of protein-coding genes in the human genome is complicated by the fact that some proteins are highly homologous, either involving individual domains or across the whole protein, which generates proteins with highly homologous structures and sequences. Proteins with only regional homology do normally not cause major problems, since the analysis using either antibodies or mass spectrometry can be directed to regions unique for each protein. Proteins with homology across the whole protein are obviously more difficult to study, including large protein families with related function, such as the olfactory receptors ($n = 374$), responsible for sensing smell and taste, and the keratin-associated proteins ($n = 92$), involved in the assembly of hair follicles, but also other highly similar genes evolved through gene duplications. Excluding the olfactory receptors and the keratin-associated proteins, there exists an additional 1170 genes in the current prediction of human protein-coding genes with 80% or higher sequence similarity to at least one other putative human gene across the entire sequence (Sivertsson and Uhlen, unpublished). Dedicated studies are probably needed to resolve the individual functionality of many of these highly homologous proteins in human biology.

The results suggest that approximately one-third of the putative protein-coding genes lack evidence on the protein level. This emphasizes the need for a systematic effort to characterize this part of the human protein-coding genes. An important mission, in this context, is to determine which genes have been wrongly predicted from the genome sequence to provide a more accurate list of genes coding for proteins. Thus, experimental data from various experimental platforms can be generated supporting the existence or nonexistence of proteins predicted from the genome sequence and annotated as putative protein-coding genes by efforts such as Ensembl.[7] In this way, it is possible to support the ongoing efforts to manually annotate the human proteins, such as UniProt[8,20,21] and to complement related resources such as neXtProt,[22] Array Express[23,24] and Peptide Atlas.[25] The strength of combining data from several separate experimental pillars is that experimental evidence from one platform can be validated with the results from an alternative, independent platform. An example of this is the study of the subcellular proteomes with the aim to determine the localization of all the proteins in a given organelle or other cellular compartment, which can be tackled using antibody-based methods,[26] GFP-fusions[27] or organelle fractionation followed by comprehensive proteomics analysis.[28,29]

An important part of the functional annotation of the human proteome is to define the expression pattern of the human proteins across cells and tissues of the human body. The analysis presented here supports earlier suggestions[6,13] that there are few cell and tissue-specific proteins. Despite this, many examples of tissue-specific proteins, not previously described, have been identified with a few examples shown in Figure 5. A list of all the cell-type specific genes are shown in Supplementary Table 2 (Supporting Information).

In conclusion, we describe a new knowledge-based resource for visualization of protein evidence for the putative human protein-coding genes. The analysis also provides evidence on which proteins are expressed in a cell- or tissue-specific manner and who are expressed in a more ubiquitous manner across all cells and tissues studied. The results demonstrate that we lack evidence for existence for approximately one-third of the putative protein-coding genes and an important mission for systematic efforts to characterize the human proteins in the near future is thus to gather information about these gene products and to adopt an integrated approach based on many experimental pillars to characterize the human proteins in cells, tissues and organs.

METHODS

Transcript Profiling (RNA-seq)

All cell lines were cultivated at 37 °C in a 5% CO_2 environment and a media recommended by the provider. Two separate cultivations were performed for each cell line to generate two independent RNA samples (biological replicates). A-431, A-549, CACO-2, HEK 293T, HeLa, Hep-G2, MCF-7 and PC-3 were obtained from DSMZ. U-2 OS were obtained from LGC/ATCC, RT-4 from ECACC and U-251 MG from Professor Bengt Westerberg at Uppsala University. We used the standard Illumina RNA-seq protocol with a read length of 2 × 100 bases. Mycoplasma tested (negative) cell lines were used for extraction of RNA. Early split samples were used as duplicates and the cells were harvested at 50% confluency, log-phase growth, and immediately frozen at −80 °C. Total RNA was extracted using the RNeasy mini kit according to the manufacturer's instructions (Qiagen, Hilden, Germany). RNA integrity was analyzed using an Experion automated electrophoresis system with the Experion RNA StdSens kit (Bio-Rad Laboratories, Hercules, CA, USA). For this study, the RPKM (reads per kilobase of exon model per million mapped reads) and FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated by dividing the number of reads mapping to the protein coding part of each gene by the length of the protein coding part of the gene and the total number of reads from the library to compensate for slightly different read depths for different samples. The softwares used to perform the mapping and FPKM calculations where Tophat v1.0.14 and Cufflinks v1.0.3.[30] We followed the approach used by Ramsköld et al.,[31] where intergenic background regions are used to calculate statistical cutoff FPKM values which can are used to estimate the gene detection limit in each sample. All genes with FPKM values below the cutoff for a sample are considered not detected. For each cell line, the total set of all FPKM values have been ordered into three classes: low (the bottom third of the set), medium (middle third of the set) and high (top third of the set), used as an estimate for the abundance level for each gene. These three classes are used to determine the abundance level for each gene in the cell line(s) where it was detected and to classify each gene into the categories "High", "Medium", "Low", and "Not detected".

Immunofluorescence Microscopy

Immunofluorescence microscopy was systematically used to determine the protein subcellular location in three human cell lines; the osteosarcoma U-2 OS, the epithelial carcinoma A-431 and the malignant gliomaU-251 MG. Cells were fixed, permeabilized and immunostained as previously described.[32,33]

Tissue Profiling

Tissue microarrays (TMA) containing triplicate 1-mm cores of 46 different types of normal tissue and duplicate 1-mm cores of 216 different cancer tissues representing the 20 most common forms of human cancer were generated as previously described.[34] TMA sections were immunostained as previously described.[35] Briefly, slides were deparaffinized in xylene, hydrated in graded alcohols and blocked for endogenous peroxidase in 0.3% hydrogen peroxide diluted in 95% ethanol. For antigen retrieval, a Decloaking chamber (Biocare Medical, Walnut Creek, CA) was used. Slides were immersed and boiled in Citrate buffer, pH 6 (Lab Vision, Freemont, CA) for 4 min at 125 °C and then

allowed to cool to 90 °C. Automated IHC was performed essentially as previously described,[36] in brief, using an Autostainer 480 instrument (Lab Vision). Primary antibodies and a dextran polymer visualization system (UltraVision LP HRP polymer, Lab Vision) were incubated for 30 min each at room temperature and slides were developed for 10 min using Diaminobenzidine (Lab Vision) as chromogen. All incubations were followed by rinse in wash buffer (Lab Vision). Slides were counterstained in Mayers hematoxylin (Histolab) and coverslipped using Pertex (Histolab) as mounting medium. Incubation with PBS instead of primary antibody served as negative control. The Aperio ScanScope XT Slide Scanner (Aperio Technologies, Vista, CA) system was used to capture digital whole slide images with a 20X objective. Slides were dearrayed to obtain individual cores. The outcome of IHC stainings in the screening phase, that included various normal and cancer tissues, was manually evaluated and scored by certified pathologists using a web-based annotation system (unpublished). In brief, the manual score of immunohistochemistry-based protein expression was determined as the fraction of positive cells defined in different tissues: 0 = 0−1%, 1 = 2−25%, 2 = 26−75%, 3 > 75% and intensity of immunoreactivity: 0 = negative, 1 = weak, 2 = moderate and 3 = strong staining.

Evidence Scores

For each gene, a protein evidence summary score was calculated based on three parameters: UniProt protein existence (Uniprot evidence), transcript profiling categories (RNA evidence) and a Protein Atlas antibody based score (HPA evidence). The UniProt protein existence data was assigned to classes; evidence at protein level (class 1), evidence at transcript level (class 2), inferred from homology (class 3), predicted (class 4) and uncertain (class 5). The UniProt data was downloaded from www.uniprot.org on 21st of July, 2011, and filtered to include only reviewed data from taxonomy 9606 (*Homo sapiens*). The UniProt protein ids were mapped to genes from Ensembl release version 63.37 (www.ensembl.org). The HPA evidence was calculated based on the manual curation of Western blot, tissue profiling and subcellular location as described in Supplementary Table 1 (Supporting Information). The protein evidence summary score for each gene was assigned "High" if a gene was found having both Uniprot evidence class 1 and "High" in the HPA evidence; "Medium" if the gene had Uniprot evidence class 1 or was scored "High" in the HPA evidence; "Low" if the HPA evidence was "Medium" and the UniProt evidence class was 2, 3, 4 or 5; "Only RNA" if UniProt evidence class 2 or RNA evidence was "High"; and "None" if RNA evidence was "Medium" or lower and the gene was scored as UniProt evidence class 3, 4 or 5.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary tables and figure. This material is available free of charge via the Internet at http://pubs.acs.org.This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mathias.uhlen@scilifelab.se.

Author Contributions

‡These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge the entire staff of the Human Protein Atlas project. This work was supported by grants from the Knut and Alice Wallenberg Foundation and the EU 7th framework programs Affinomics and PROSPECTS.

■ REFERENCES

(1) A gene-centric human proteome project: HUPO--the Human Proteome organization. *Mol. Cell. Proteomics*, **2010**, *9*, 427−429.
(2) The call of the human proteome. *Nat. Methods*, **2010**, *7*, 661.
(3) Hochstrasser, D. Should the Human Proteome Project be gene- or protein-centric? *J. Proteome Res.* **2008**, *7*, 5071.
(4) Uhlen, M.; et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **2005**, *4*, 1920−1932.
(5) Nilsson, P.; et al. Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **2005**, *5*, 4327−4337.
(6) Uhlen, M.; et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28*, 1248−1250.
(7) Flicek, P.; et al. Ensembl 2011. *Nucleic Acids Res.* **2011**, *39*, D800−806.
(8) UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **2011**, *39*, D214−219.
(9) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5*, 621−628.
(10) Lundberg, E.; et al. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **2010**, *6*, 450.
(11) Ramskold, D.; Wang, E. T.; Burge, C. B.; Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **2009**, *5*, e1000598.
(12) Hebenstreit, D.; et al. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **2011**, *7*, 497.
(13) Ponten, F.; et al. A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* **2009**, *5*, 337.
(14) Legrain, P.; et al. The human proteome project: Current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009993.
(15) Fagerberg, L.; et al. Mapping the Subcellular Protein Distribution in Three Human Cell Lines. *J. Proteome Res.* **2011**, *10*, 3766−3777.
(16) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2011**, *27*, 431−432.
(17) Komuro, A.; et al. The AHNAKs are a class of giant propeller-like proteins that associate with calcium channel proteins of cardiomyocytes and other cells. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4053−4058.
(18) Parkinson, H.; et al. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* **2009**, *37*, D868−872.
(19) Barrett, T.; et al. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.* **2011**, *39*, D1005−1010.
(20) UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142−148.
(21) Jain, E.; et al. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **2009**, *10*, 136.
(22) Lane, L. neXtProt, a new knowledgebase on human proteins. Available from *Nature Precedings* <http://dx.doi.org/10.1038/npre.2010.5104.1>, 2010.
(23) Brazma, A.; et al. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **2003**, *31*, 68−71.
(24) Parkinson, H.; et al. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* **2011**, *39*, D1002−1004.
(25) Desiere, F.; et al. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34*, D655−658.
(26) Lundberg, E.; Uhlen, M. Creation of an antibody-based subcellular protein atlas. *Proteomics* **2010**, *10*, 3984−3996.

(27) Vermeulen, M.; et al. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **2010**, *142*, 967−980.
(28) Andersen, J. S.; Mann, M. Organellar proteomics: turning inventories into insights. *EMBO Rep.* **2006**, *7*, 874−879.
(29) Foster, L. J.; et al. A mammalian organelle map by protein correlation profiling. *Cell* **2006**, *125*, 187−199.
(30) Trapnell, C.; et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511−515.
(31) Ramskold, D.; Wang, E. T.; Burge, C. B.; Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **2009**, *5*, e1000598.
(32) Barbe, L.; et al. Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* **2008**, *7*, 499−508.
(33) Stadler, C.; Skogs, M.; Brismar, H.; Uhlen, M.; Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **2010**, *73*, 1067−1078.
(34) Ponten, F.; Jirstrom, K.; Uhlen, M. The Human Protein Atlas–a tool for pathology. *J. Pathol.* **2008**, *216*, 387−393.
(35) Kampf, C.; Andersson, A. C.; Wester, K.; Björling, E.; Uhlén, M.; Pontén, F. Antibody-based tissue profiling as a tool in clinical proteomics. *Clin. Proteomics* **2004**, *1*, 285−300.
(36) Paavilainen, L.; et al. The impact of tissue fixatives on morphology and antibody-based protein profiling in tissues and cells. *J. Histochem. Cytochem.* **2010**, *58*, 237−246.