⌘ Author's Choice

Antibody-based Protein Profiling of the Human Chromosome 21*[S]

Mathias Uhlén‡§¶, Per Oksvold§, Cajsa Älgenäs§, Carl Hamsten§, Linn Fagerberg§, Daniel Klevebring‖, Emma Lundberg‡, Jacob Odeberg‡, Fredrik Pontén, Tadashi Kondo‡‡, and Åsa Sivertsson§**

The Human Proteome Project has been proposed to create a knowledge-based resource based on a systematical mapping of all human proteins, chromosome by chromosome, in a gene-centric manner. With this background, we here describe the systematic analysis of chromosome 21 using an antibody-based approach for protein profiling using both confocal microscopy and immunohistochemistry, complemented with transcript profiling using next generation sequencing data. We also describe a new approach for protein isoform analysis using a combination of antibody-based probing and isoelectric focusing. The analysis has identified several genes on chromosome 21 with no previous evidence on the protein level, and the isoform analysis indicates that a large fraction of human proteins have multiple isoforms. A chromosome-wide matrix is presented with status for all chromosome 21 genes regarding subcellular localization, tissue distribution, and molecular characterization of the corresponding proteins. The path to generate a chromosome-specific resource, including integrated data from complementary assay platforms, such as mass spectrometry and gene tagging analysis, is discussed. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.013458, 1–10, 2012.***

The Human Proteome Project has been proposed (1) to systematically map the human proteins in a chromosome-specific manner using mass spectrometry-based methods combined with antibody-based characterization. One of the major challenges to such a project is the dynamics of the human proteome, including temporal and spatial parameters, transient and stable interactions, and the vast amount of isoforms and protein variants (2). There have also been proposals for alternative strategies, such as a more disease-driven proteome project with the objective to explore various human diseases using mass spectrometry-based methods (3). These two approaches have now been combined into the Human Proteome Project launched by the Human Proteome Organization (HUPO) (4). The questioning of a gene-centric approach as the most suitable strategy for a systematic exploration of human proteins calls for pilot projects to demonstrate feasibility and to facilitate the definition of suitable milestones and deliverables for a complete genome-wide proteome project.

Here, we describe a pilot study to investigate the genes encoded on human chromosome 21 using antibody-based profiling with the aim of characterizing the proteome components, including protein isoforms, subcellular localization, and distribution profiles in cells, tissues, and organs. Chromosome 21 is the smallest autosomal chromosome, regarding both size and gene numbers, in humans, and three copies of the chromosome (trisomy 21) is the underlying cause for Down syndrome. With regards to chromosome 21, a first attempt to generate antibodies to the gene products from this chromosome was published already in 2003 (5), as a prelude to the Human Protein Atlas effort, aimed to generate publicly available subcellular localization data and expression data for most major human tissues and organs (6, 7). Recently, version 7 of the Human Protein Atlas portal was launched (8) with expression data for more than 50% ($n = 10{,}170$) of the human protein-coding genes.

We report on a first attempt on a chromosome-wide analysis using antibody-based methods, including tissue profiles to cover 131 of the 240 protein-coding genes defined by the Ensembl database, and extended the analysis by molecular characterization of the proteins, including an isoform analysis of selected proteins. In addition, we have included RNA data to provide evidence for existence of the protein-coding genes on the transcriptional level. The results demonstrate the power of an integrated approach to characterize the protein-coding genes using a gene-centric approach.

EXPERIMENTAL PROCEDURES

Western Blot—A panel comprising two cell lines (RT-4 and U-251 MG), two human tissues (liver and tonsil), and HSA/IgG depleted

From the ‡Science for Life Laboratory, Royal Institute of Technology, SE-17121 Solna, Sweden, the §School of Biotechnology, AlbaNova University Center,Royal Institute of Technology, SE-106 91 Stockholm, Sweden, the ‖Department of Medical Epidemiology, Karolinska Institute, SE-171 77 Stockholm, Sweden, the **Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden, and the ‡‡Division of Pharmacoproteomics, National Cancer Center Research Institute, Tokyo 104-0045, Japan

⌘ *Author's Choice*—Final version full access.
Received August 14, 2011, and in revised form, October 20, 2011
Published, MCP Papers in Press, October 31, 2011, DOI 10.1074/mcp.M111.013458

human plasma was selected for protein characterization using Western blot analysis. 15 μg of total protein lysate and 25 μg of depleted plasma were subjected to a precast 10–20% Criterion™ SDS-PAGE gradient gel (Bio-Rad Laboratories, CA) under reducing conditions followed by transfer to a PVDF membrane using Criterion™ gel blotting sandwiches (Bio-Rad Laboratories, CA) according to the manufacturer's recommendations. PVDF membranes were presoaked in methanol and blocked (5% dry milk, 0.5% Tween 20, 1*TBS (150 mM NaCl, 10 mM Tris HCL)) for 45 min at room temperature followed by 1 h of incubation with primary antibody, diluted 1:250 in blocking buffer. After four 5-min washes in TBST (0.1 M Tris-HCl, 0.5 M NaCl, 0.05% Tween 20), the membranes were incubated for 1 h with an horseradish peroxidase-conjugated polyclonal swine anti-rabbit antibody (Dako, Glostrup, Denmark) diluted 1:3000 in blocking buffer. A final round of four 5-min TBST washes was performed before chemiluminescence detection, using a CCD camera (Bio-Rad Laboratories, CA) and Immobilon Western chemiluminescent horseradish peroxidase substrate (Millipore Corporation, Billerica, MA).

Isoelectric Focusing—Fourteen genes on chromosome 21 were transfected to HEK 293 cells, and proteins were extracted. The resulting protein lysates were purchased from OriGene Technologies (Rockville, MD). Protein concentration was measured by a Bio-Rad protein assay kit. Five micrograms of protein were diluted with 320 μl of rehydration buffer containing 6 M urea, 2 M thiourea, 3% CHAPS,[1] 1% Triton X-100, 13 mM DTT, 1% Pharmalyte pH 3–10 (GE Healthcare, Japan). The samples were loaded on a 18-cm IPG DryStrip gel (pH 3–10; GE Healthcare, Japan) by sample rehydration overnight. Subsequently, the strips were subjected to isoelectric focusing on a Multiphor II (GE Healthcare, Japan) at 20 °C under the following conditions: 500 V (gradient over 0.5 h), 3500 V (gradient, 1.5 h), or 3500 V (hold, 6.5 h), resulting in 16250 Vh. Subsequently, the strips were stored at -80 °C. The gel-separated protein samples were blotted onto PVDF membrane by passive diffusion for 3 h with conventional transfer buffer containing 50 mM Tris, pH 7.4, 200 mM NaCl, 0.05% Tween 20. The membranes were blocked with blocking buffer and 5% skimmed milk in TBS-T buffer at room temperature for 1 h and washed four times in TBS-T for 5 min. The primary antibodies were diluted with the blocking buffer (1:250) and incubated with the membranes at 4 °C overnight. The membranes were washed four times in TBS-T for 5 min. Subsequently, the primary antibodies were reacted with horseradish peroxidase-conjugated polyclonal sheep anti-rabbit IgG antibody in the blocking buffer (1:3000 dilution). After incubating for 1 h, the membranes were washed four times in TBS-T buffer for 5 min. The bound peroxidase-conjugated anti-rabbit antibody was detected using the ECL-plus kit (GE Healthcare, Japan) and LAS-3000 (Fuji-Film, Tokyo, Japan). The observed isoelectric point was calculated by measuring the electrophoretic migration in a linear pH gradient. The theoretical isoelectric point was obtained by the on-line software Compute pI/Mw tool at the ExPASy website.

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 glioma U-251 MG. The cells were fixed, permeabilized, and immunostained as previously described (9, 10).

Tissue Profiling—Tissue microarrays 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 (11). Tissue microarray sections were immunostained as previously described (12). Briefly, the 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. The 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 immunohistochemistry was performed essentially as previously described (13), in brief, using an Autostainer 480 instrument (Lab Vision, Freemont, CA). Primary antibodies and a dextran polymer visualization system (UltraVision LP horseradish peroxidase polymer; Lab Vision, Freemont, CA) were incubated for 30 min each at room temperature, and the slides were developed for 10 min using diaminobenzidine (Lab Vision, Freemont, CA) as chromogen. All of the incubations were followed by rinse in wash buffer (Lab Vision, Freemont, CA). The slides were counterstained in Mayers hematoxylin (Histolab, Sweden) and coverslipped using Pertex® (Histolab, Sweden) as mounting medium. Incubation with PBS instead of primary antibody served as negative control. The Aperio Scan Scope CS slide scanner (Aperio Technologies, Vista, CA) system was used to capture digital whole slide images with a 20× objective. The slides were de-arrayed to obtain individual cores. The outcome of immunohistochemistry stainings in the screening phase, which included various normal and cancer tissues, was manually evaluated and scored by certified pathologists using a web-based annotation system as previously described (14). In brief, the manual score of immunohistochemistry-based protein expression was determined as the fraction of positive cells defined in different tissues: $0 = 0–1\%$, $1 = 2–25\%$, $2 = 26–75\%$, and $3 = >75\%$ and intensity of immunoreactivity: $0 =$ negative, $1 =$ weak, $2 =$ moderate, and $3 =$ strong staining. All of the tissues used as donor blocks were acquired from the archives at the Department of Pathology of Uppsala University Hospital in agreement with approval from the Research Ethics Committee at Uppsala University (Uppsala, Sweden) (Ups 02-577).

Transcript Profiling (RNA-seq)—The RNA-seq method using the SOLiD3 platform has been previously described (15). For this study, the RPKM (reads per kilobase of exon model per million mapped reads) value was 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 total set of all RPKM values from all genes and all three cell lines 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). 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: "supportive" (medium to high levels), "uncertain" (low levels), and "not supportive" for genes not detected in any of the cell lines.

RESULTS

Overall Annotation of Human Chromosome 21—In Fig. 1, the 240 putative protein-coding genes on chromosome 21 as defined by the Ensembl effort (16) (release 59) are outlined with a color code to show the current knowledge base according to UniProt (17–19). These putative genes are ranging from very well known genes, such as the amyloid β precursor protein responsible for Alzheimer disease, to many genes of unknown function or even questionable existence. Chromosome 21 has 48 keratin-associated genes (*brown* in Fig. 1) encoding small, homologous proteins that are involved in the formation of the cross-linked network of the keratin-intermediate filament proteins that support hair fibers (20, 21).

[1] The abbreviation used is: CHAPS, 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonic acid.

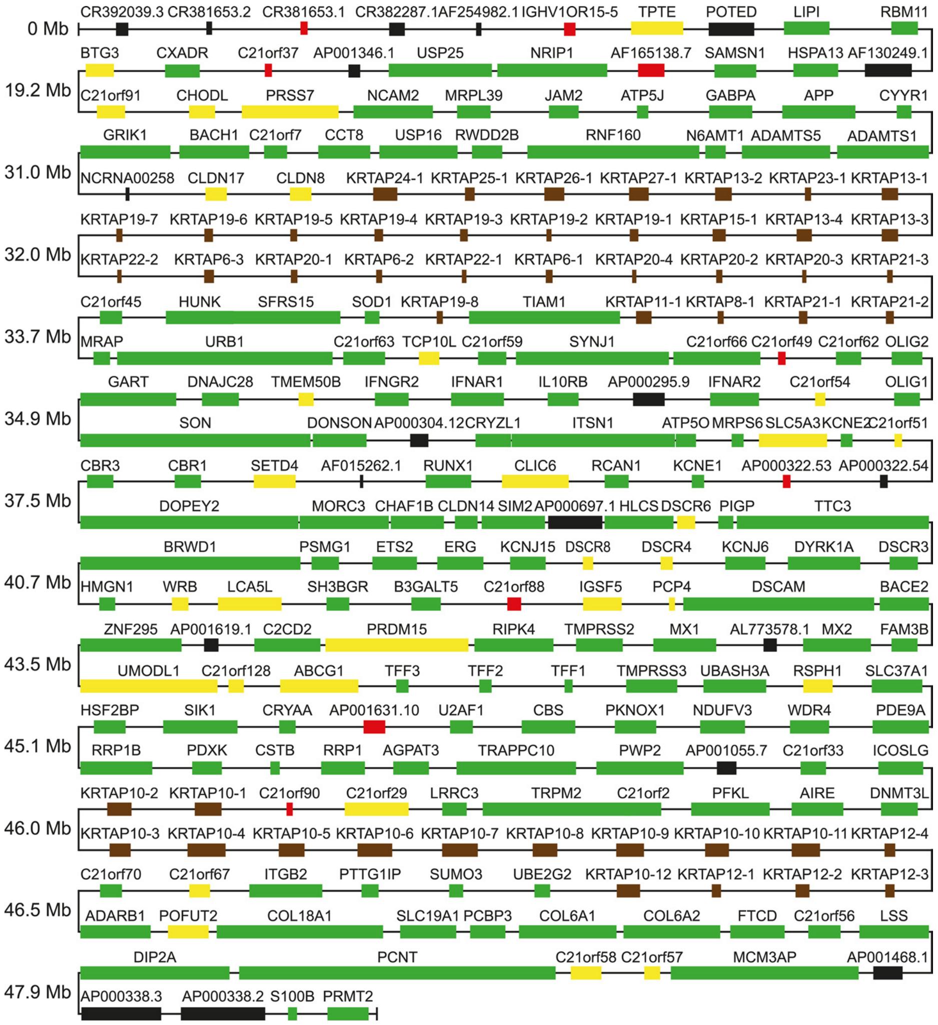

FIG. 1. **Overview and evidence level of the putative protein-coding genes on human chromosome 21.** The 240 protein-coding genes annotated in Ensembl release 59 have been color coded according to the annotation of the corresponding protein in the UniProt database. The size of each box corresponds to the number of amino acids of the largest splice variant of the corresponding gene, but the distances between genes are not drawn in scale; instead 10 genes are shown across each *horizontal line* (except for the last two lines). The chromosomal position from the left arm of the chromosome is shown to the *left*. The color codes are as follows: *green*, UniProt category 1 (evidence at protein level); *yellow*, UniProt category 2 (evidence at transcript level); *red*, UniProt category 4 and 5 (predicted and uncertain); *black*, no reviewed data available in UniProt; *brown*, keratin-associated genes.

Excluding the keratin-associated proteins, there are 192 putative protein-coding genes, and according to UniProt (17–19), there is evidence at the protein level for 69% (n = 133) of these genes, whereas another 31 genes have been found only at the transcriptional level. For another nine genes, there is no evidence either on the transcript or protein level (class 4 and 5 genes), and for 19 genes there are no reviewed data in the UniProt portal. The large fraction of protein-coding genes lacking evidence at the protein level demonstrates the need for systematic strategies to characterize the putative proteins and presents chromosome 21 as an appropriate target for a gene-centric approach. In supplemental Table 1, a list of all 240 genes, including keratin-associated, are presented with data predicted from the genome sequence, including molecular weight, signal peptides, transmembrane regions, and number of splice variants. Another 41 proteins defined by Uniprot are not included in the Ensembl list of genes for this chromosome and therefore are excluded from this study (see supplemental Table 2). These genes might be included in extended studies of the chromosome 21 genes in the future.

Antibody-based Protein Profiling—As part of the Human Protein Atlas project, we have generated antibodies in a systematic effort, and this has been complemented with antibodies from more than 60 commercial providers. A summary of the overall status for the chromosome 21 gene products can be seen in Fig. 2*a*. Antibody-based protein profiling data are provided for 68% of the protein-coding genes, and for one-third of these proteins, knowledge-based annotated protein expression level data (8), based on at least two separate antibodies, are available. Antibodies approved by a multiplex microarray assay (7) exist for an additional 20% of the genes, and recombinant antigens were verified by mass spectrometry for another 5%. Thus, at present more than 90% of the putative genes on chromosome 21 have either antibodies or mass spectrometry-approved antigens.

Molecular Characterization—The protein products of 167 genes were characterized by Western blot analysis (22) of protein lysates from selected human cell lines, tissues, and a pooled mixture of plasma. A summary of the results is presented in supplemental Fig. 1. 50% of the analyzed proteins displayed a band corresponding to the predicted size in one

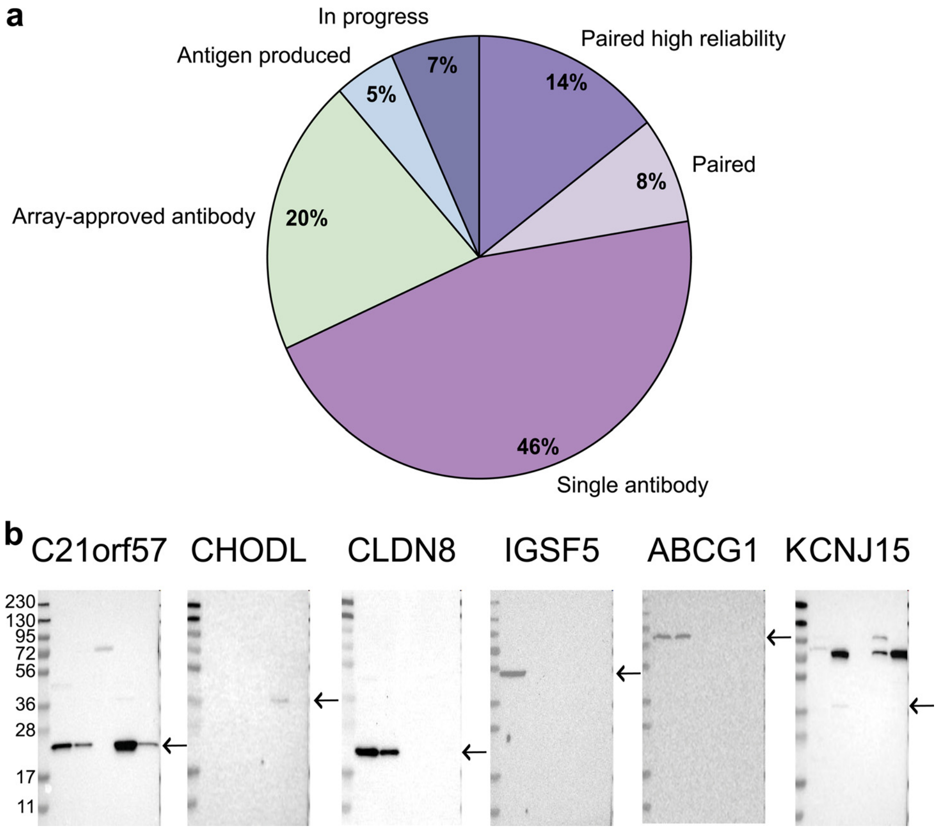

FIG. 2. **Status and molecular characterization of the chromosome 21 encoded proteins.** *a*, the pie chart shows the status of the chromosome 21 genes as the fraction of genes having two published antibodies for which knowledge-based annotation show high reliability (*Paired high reliability*), having two published antibodies for which knowledge-based annotation is not conclusive (*Paired*), having one published antibody (*Single antibody*), having an antibody specifically recognizing its target on a protein array (*Array-approved antibody*), having an MS-verified antigen produced (*Antigen produced*), and being in an earlier stage in the workflow (*In progress*). *b*, the first five Western blots from the *left* show proteins that were detected as a single major band of predicted size and had no previous evidence at the protein level. The *right-most Western blot* is an example where there are strong indications that glycosylation of the protein is responsible for the shift of the band to a larger size than predicted.

or more of the samples. According to Uniprot, there is no previous evidence on the protein level for many of these genes and in Fig. 2*b*, five examples of such genes (C21orf57, CHODL, CLDN8, IGSF5, and ABCG1) are shown. The results from the Western blot analysis show that a protein of the expected size is detected with the antibody in at least one of the analyzed cells or tissues for each of these genes, thus providing evidence on the protein level based on the appearance of a single band of the predicted size in the Western blot assay.

For 21% of the proteins, the results from the Western blot analysis were considered not conclusive, in most cases because of detection of either several bands or a single band of other than expected size. This might to some extent reflect the presence of yet uncharacterized splice variants or alternative isoforms resulting from protein modifications. An example of this is the human protein KCNJ15, which is predicted to be a multi-pass membrane protein belonging to the inward rectifier-type potassium channel family (23). The molecular mass according to the antibody-based Western blot analysis is ~75 kDa (Fig. 2*b*), whereas the theoretical size predicted from the genome sequence is 43 kDa. This is not unexpected, because integral membrane proteins frequently are *N*-glycosylated (24), yielding glycoproteins of higher molecular mass, and the results therefore suggest that human KCNJ15 is indeed glycosylated. It is reassuring that the specificity of the antibody is supported by confocal microscopy analysis showing a subcellular localization in the plasma membrane (supplemental Fig. 2).

Subcellular Localization—The antibodies were subsequently used to determine the subcellular localization of each gene product as part of the Human Protein Atlas effort. Antibodies corresponding to 97 human genes were analyzed on the subcellular level using high resolution confocal microscopy of immunostained cell lines. 41% of the analyzed proteins were detected in a single subcellular compartment and 59% in multiple compartments (data not shown). Fig. 3*a* shows examples from the subcellular localization analysis with *images I–III* displaying proteins expressed in cytoplasmic/membranous locations, including the centrosomal protein PCNT, which is localized to centrosomes (*image I*), the receptor protein CXAR specifically localized to cell junctions (*image II*), and the mitochondrial protein ATP5J localized to the mitochondria (*image III*). *Images IV–VI* demonstrate three different types of nuclear distribution patterns with the transcription factor BACH1 localized to the nucleus, the ribosomal protein RRP1 localized to nucleoli, and the SON DNA-binding protein localized to specific patches of the nucleus, with a speckled pattern typical for DNA-binding proteins.

Normal Tissue Profiles—The validated antibodies have also been used to generate protein profiles corresponding to 131 of the 192 non-keratin-associated protein-coding genes defined by the Ensembl database, using immunohistochemistry-based protein detection in tissue microarrays as part of the

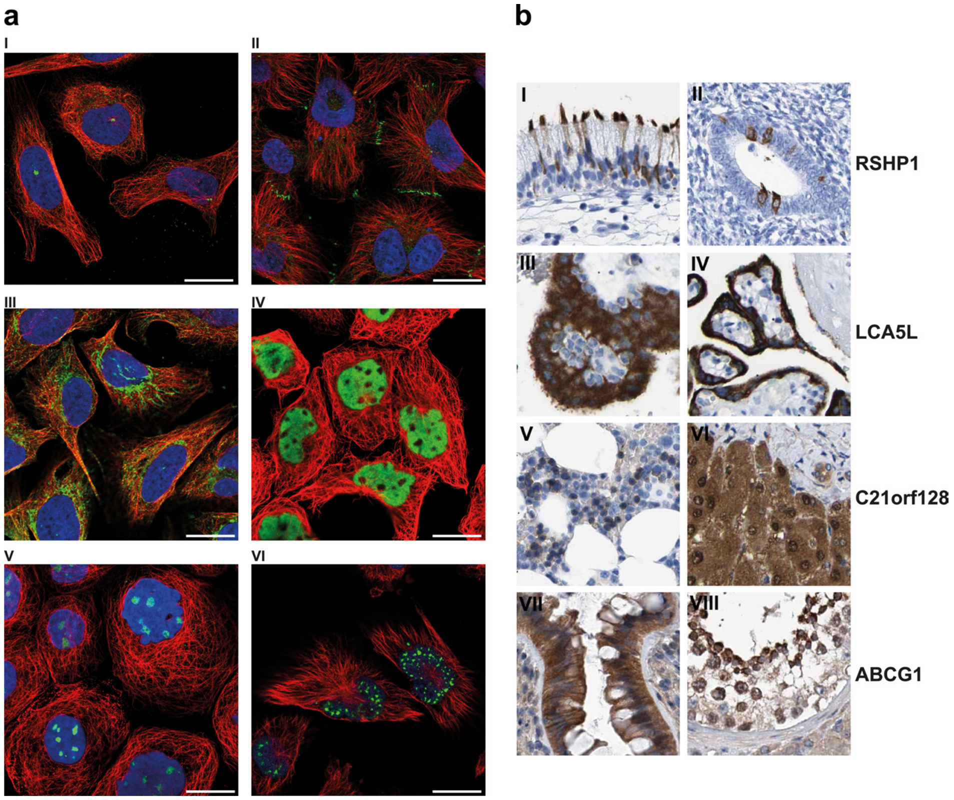

FIG. 3. **Antibody-based profiling to determine subcellular localization and tissue distribution.** *a*, examples of proteins localized to different subcellular compartments. The protein of interest is shown in *green*, the nucleus in *blue*, and microtubules are in *red*. The *scale bar* indicates 20 μm. *Image I*, image of the centrosomal protein PCNT, in U-2 OS cells using the antibody HPA016820. *Image II*, image of the receptor CXAR, localized to cell junctions in U-251 MG cells using the antibody HPA030411. *Image III*, image of the mitochondrial protein ATP5J in U-2 OS cells, using the antibody HPA031069. *Image IV*, image of the transcription factor BACH1 shown to localize to the nucleus but not to the nucleoli in A-431 cells using the antibody HPA003175. *Image V*, image of the ribosomal protein RRP1 localized to the nucleoli in A-431 cells using the antibody HPA018166. *Image VI*, image of the DNA-binding protein SON that localizes to the nucleus with a speckled pattern in U-251 MG cells using the antibody HPA023535. *b*, examples of immunohistochemistry stained tissue sections from various normal human tissues and cancer tissues. The previously unknown RSPH1 (Radial spoke head 1 homolog) protein was found expressed in ciliated cells in various tissues. The pattern of expression included a distinct positivity in cilia within the respiratory mucosa (*image I*). In hormonally active corpus mucosa, a small subset of normal glandular cells shows clear cytoplasmic expression of RSPH1 (*image II*). The unknown LCA5L (Leber congenital amaurosis 5-like) protein was found to be highly cell type-specific and in normal tissues expressed at high levels in only placental trophoblasts from both early (*image III*) and late stage placenta (*image IV*). The C21orf128 was found to be expressed in a selective pattern. Although two different antibodies showed a partly different pattern of immunohistochemical positivity, both antibodies showed a similar expression pattern in a subset of hematopoietic cells in normal bone marrow (*image V*) and widespread cytoplasmic and nuclear expression in hepatocytes from normal liver tissue (*image VI*). The ATP-binding cassette, subfamily G (*WHITE*), member 1 (ABCG1), belonging to the UniProt class of potential transmembrane proteins with no evidence at protein level, showed ubiquitous expression in epithelial cell types, whereas cells within the central nervous system and hematopoietic system were essentially negative. In normal colonic mucosa, ABCG1 expression was accentuated in surface epithelium and differentiating glandular cells lining upper parts of the colonic crypts (*image VII*). ABCG1 was also expressed in germinal cells from normal testis, with an apparent gradient, so that the most mature spermatocytes and spermatids showed a relative higher level of expression as compared with more basal layers in the seminiferous duct (*image VIII*).

Human Protein Atlas effort. In this way, the protein profiles in 46 normal tissues, and organs were determined, including liver, kidney, pancreas, gastrointestinal tract, lung, and various regions in the brain. For 22 of these gene products, no previous evidence on the protein level exists according to UniProt, and therefore this antibody-based effort contributes to the functional annotation of the corresponding proteins. For a subset of these proteins, annotated protein expression patterns were obtained using two or more paired antibodies to the same target (8). One such example is the RSPH1 gene with an interesting tissue-specific and highly selective expression pattern localized to cilia in ciliated epithelium, exemplified in respiratory epithelia (Fig. 3*b*, *image I*) and maturing spermatids in testis. The expression pattern is in agreement with earlier reports suggesting a role in ciliary function for other members of radial spoke head genes (25). Our data also suggest that the RSPH1 protein is expressed in a subset of ciliated glandular cells in the endometrium (Fig. 3*b*, *image II*). The putative protein LCA5L was also found to be expressed in a highly specific manner, with protein expression restricted to trophoblasts of the placenta, both early, immature placental tissue (Fig. 3*b*, *image III*), as well as fully matured, end stage placenta (*image IV*). The putative protein C21orf128 was found to be expressed in a selective manner with the highest expression levels in a subset of hematopoietic cells (Fig. 3*b*, *image V*) and liver hepatocytes (*image VI*). An example of a more ubiquitously expressed protein was the putative protein ABCG1, expressed abundantly in epithelial cell types, as exemplified by widespread cytoplasmic expression in glandular cells lining crypts in colon mucosa (Fig. 3*b*, *image VII*) and maturing germinal cells in seminiferous ducts of testis (*image VIII*). In addition to these examples of previously unknown proteins, there are several known proteins for which we report an in-depth analysis of protein expression across a multitude of human cells, tissues, and organs, such as the protein OLIG2, showing expression in a subset of glial cells in normal cerebral cortex and malignant gliomas of oligodendrocytic subtype. A summary view with additional examples of protein profiles in normal tissues and cancer tissues for the above five examples are shown in supplemental Fig. 3.

RNA Expression Analysis—An important complement to the antibody-based profiling described above is to perform transcript profiling using RNA-seq. We have recently shown (15), using a comparison of mass spectrometry-based stable isotope labeling with amino acids in cell culture (SILAC) analysis (26, 27) and quantitative transcript profiling using RNA-seq, that there is a strong correlation between changes of RNA and protein levels when differences in levels between human cell lines are analyzed. This enforces the need to characterize expression levels on both the transcript and protein levels to use as validation of the respective results, but also to pinpoint genes with low correlation between protein and RNA changes. The RNA-seq data (15) were reanalyzed for all the putative genes on chromosome 21, and ~56% of the genes showed strong evidence at the transcriptional level (Fig. 4*a*) as judged by high or medium level transcripts in at least one of the three human cell lines analyzed. A comparison of the expression in the three analyzed cell lines show that more than 50% of the genes (n = 111) show a "housekeeping" expression pattern with similar or slightly changed transcript levels in all three cell lines, whereas 7% (n = 13) are cell type-specific (supplemental Fig. 4). These 13 genes are interesting starting points to understand the biology of phenotypes corresponding to cells of brain, epithelial, and mesenchymal origin, respectively.

The RNA-seq analysis can also be used to validate putative genes with no evidence on the protein level. A bar plot can be made showing the read coverage for each nucleotide in the region of the chromosome for the putative protein-coding gene, including exons and introns. Obviously, the read count should be larger for the exons as compared with the introns if the gene transcript is spliced to form a functional mRNA. In Fig. 4 (*b*–*d*), three examples of chromosome 21 genes with no previous evidence on either protein or transcript level are shown, with the predicted exons and introns together with the read coverage shown across the whole chromosomal region. For all three genes, the bar plots suggest efficient splicing of the exons with low number of reads in the intron regions. These results strongly support that the three putative genes are indeed coding for proteins and re-enforces that efforts should be made to generate antibodies to allow for characterization of the corresponding proteins.

Overall Status of the Antibody-based Profiling—In Fig. 5, a detailed matrix is shown with the status of the experimental characterization of the proteins encoded by chromosome 21 genes, here excluding the keratin-associated proteins, with further details presented in supplemental Table 3. The first column shows the status of the annotation performed by Uniprot for the 192 putative genes with the color code according to Fig. 1. The second column shows the status of the generation of antibodies with a *green box* representing genes with at least one antibody approved by the Human Protein Atlas program and available to the public. Most of the remaining genes have a *yellow* color code, indicating that at least one antigen corresponding to a unique region of the corresponding protein target has been expressed, purified, and verified by mass spectrometry as part of the Human Protein Atlas program. At present, only four putative genes have failed attempts to generate antigens: APOO1346.1, AFO15262.1, DSCR8, and C21orf33. Two of these lack transcript evidence according to UniProt, and corresponding transcripts have not been detected in any of the three diverse cell lines assayed using RNA-seq, which calls for more in-depth studies to confirm the annotation of these putative genes as protein coding.

The next three columns in the status matrix show the results of the antibody-based molecular, subcellular, and tissue profiling, respectively. The molecular profiling was done using Western blot analysis, the subcellular profiling was done using immunofluorescent-based confocal microscopy, and the tissue profiling was done using immunohistochemistry on tissue microarrays (see above for details). The color codes show the status of these applications for each gene, with the results displayed as supportive (*green*), uncertain (*yellow*), nonsupportive (*red*), or not done (*black*). The final column shows the results of transcript profiling in three functionally different cell lines using next generation sequencing (RNA-seq), and the color code indicates how well the results support actual transcription of the gene with *green* as supportive (high or medium expression in at least one cell line), *yellow* as uncertain (no more than low expression in any cell line), and *red* as unsupportive (not detected in any cell line).

Isoform Characterization—An important part of the molecular characterization is to determine the status of post-translational modifications, such as phosphorylation, acetylation, methylation, glycosylation, and proteolytic truncations. Because these modifications cause changes in the

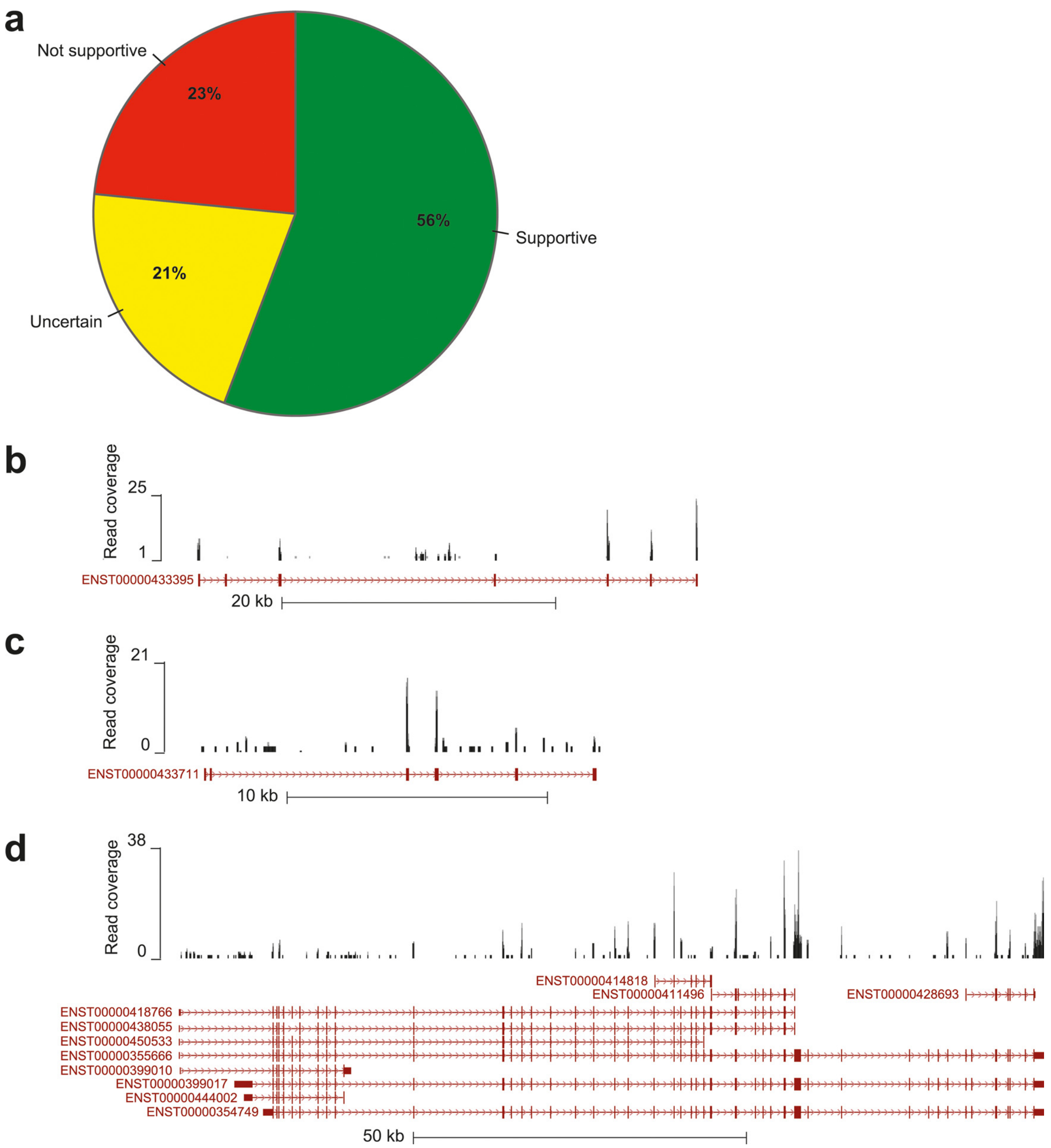

FIG. 4. **RNA-seq to determine transcript abundance in three cell lines.** *a*, the number of genes present in the three categories as defined by transcript expression levels in the three cell lines: supportive (detected at medium or high level in one or more cell lines), uncertain (detected at no more than low level in any cell line), and not supportive (not detected in any cell line). *b*, read coverage of RNA-seq reads mapping to the protein coding transcript of the gene AP000295.9 (ENSG00000249624). *c*, read coverage of RNA-seq reads mapping to the protein coding transcript of the gene AP001055.7 (ENSG00000248354). *d*, read coverage of RNA-seq reads mapping to the protein coding transcripts of the gene TTC3 (ENSG00000182670). The reads align almost exclusively to the exons of the genes.

isolectric point of the target protein, it is possible to explore the protein modification landscape of each gene product by isoelectric focusing followed by Western blotting analysis (28, 29). This allows a rapid and systematic analysis of the modification landscape for each protein, and here we decided to perform the analysis using recombinant proteins expressed from full-length clones of the respective putative genes.

In Fig. 6, 14 examples of chromosome 21 genes are shown with the theoretical pI values for each putative protein, including splice variants, given by *red arrowheads*. We examined the proteins over a pH gradient from 3 to 10, because a majority of the proteins have isoelectric points in this range. For all proteins analyzed, the observed isoelectric points based on the electrophoretic migration were somewhat different from the theoretical pI calculated from genome sequence data. Of the 14 examined proteins, six resulted in a single band, whereas eight were detected as multiple bands. Although the analysis is based on few samples only,

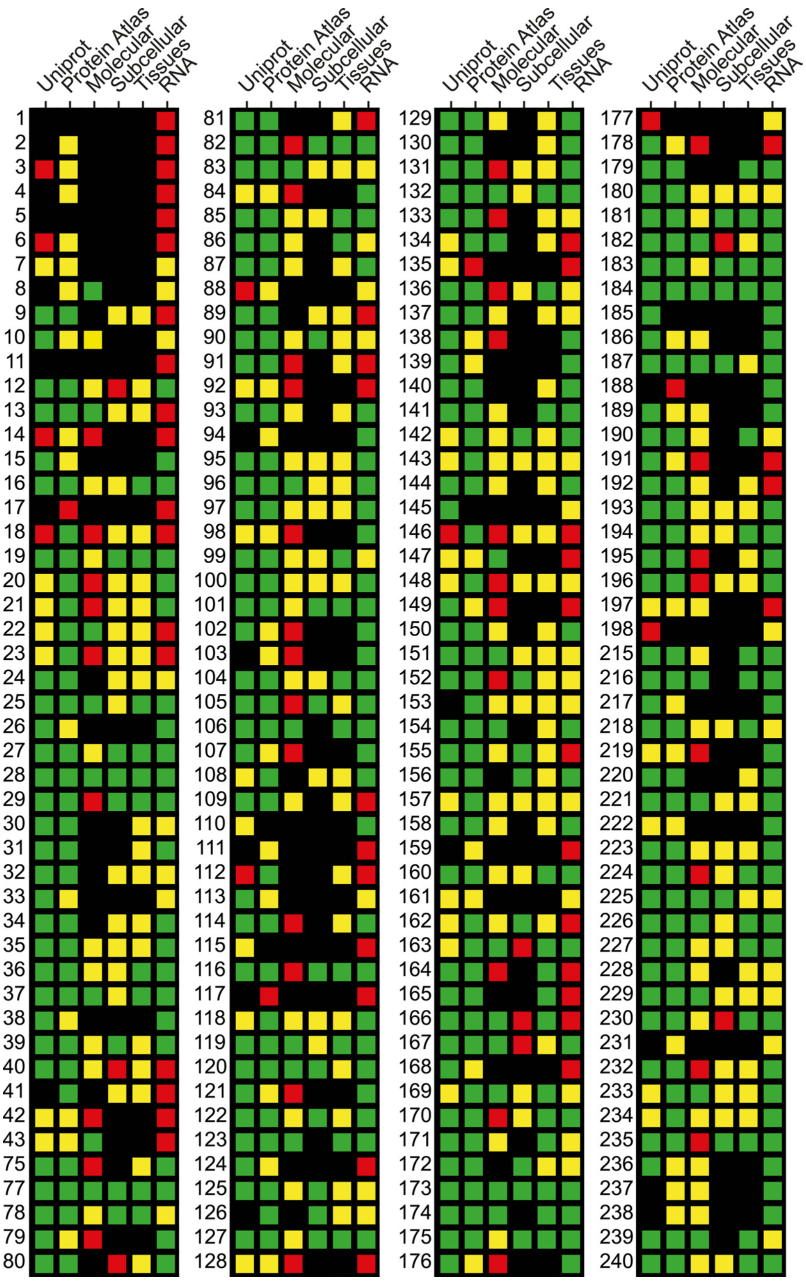

FIG. 5. **Overview of data for proteins on chromosome 21.** A status matrix is shown where, for each gene, the evidence levels according to UniProt (*first columns*), the current status in the Human Protein Atlas (*second columns*), the result of molecular characterization using Western blot (*third columns*), the reliability of the results from subcellular location analysis (*fourth columns*), and tissue profiling (*fifth columns*), and the evidence for existence on transcriptional level in cell lines (*sixth columns*) are represented using four colors as follows. *First columns*: *green*, protein evidence; *yellow*, transcript evidence; *red*, uncertain; *black*, not reviewed. *Second columns*: *green*, at least one antibody approved by the Human Protein Atlas; *yellow*, MS-verified antigen generated; *red*, failed or not started; *black*, in progress. *Third* through *sixth columns*: *green*, supportive; *yellow*, uncertain; *red*, nonsupportive; *black*, not done.

the results suggest that at least half of the analyzed genes encode proteins that are post-translationally modified. These proteins are interesting starting points for systematic analysis to explore the molecular basis of these modifications. In summary, this pilot study shows that an antibody-based isoelectric analysis can be used for rapid and convenient identification of potential targets for further isoform analysis to explore the degree of modification of each gene product.

DISCUSSION

Here, we report on a gene-centric approach aimed to experimentally annotate all protein-coding genes of the human chromosome 21 using antibody-based profiling. The genome sequence analysis by the Ensembl group has in release 59 identified 192 non-keratin-associated putative genes coding for proteins on this chromosome, and these genes have been characterized on the protein level by antibody-based profiling, and the status was reported in a matrix. The overall aim is to fill this matrix with information on all levels to generate experimental evidence for molecular characterization, isoforms, subcellular localization, tissue profiles, and cell and tissue specificity and to contribute to the functional annotation of the proteome by identifying faulty annotated genes that do not code for proteins.

The study presented here has contributed to several insights of both general and specific interest. Five genes with no

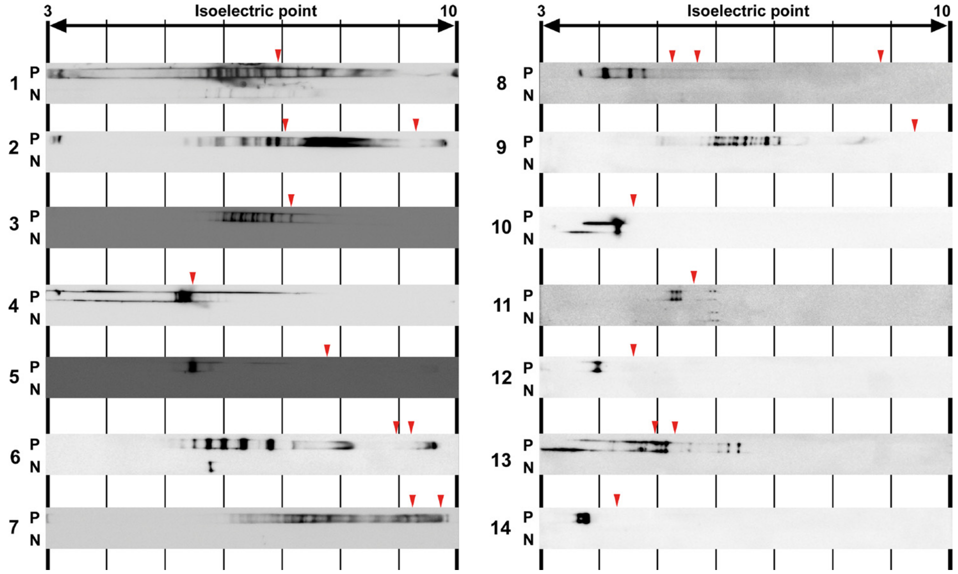

FIG. 6. **Analysis of protein isoforms using isoelectric focusing.** Protein lysates for 14 chromosome 21 genes were separated according to isoelectric point on a pH gradient gel and subsequently reacted with corresponding antibodies. *Red triangles* indicate the theoretically calculated isoelectric point(s) for each protein, including known splice variants. To verify antibody specificity, each antibody is analyzed against one lysate from cells transfected with a chromosome 21 gene (*rows P*) and one lysate from cells infected with control vector (*rows N*). Pairs of protein and antibody are as follows: *pair 1*, C21orf59/HPA019055; *pair 2*, C21orf56/HPA018979; *pair 3*, CHAF1B/HPA016698; *pair 4*, CRYZL1/HPA019120; *pair 5*, DSCR4/HPA018460; *pair 6*, FAM3B/HPA015885; *pair 7*, NDUFV3/HPA020463; *pair 8*, PIGP/HPA026921; *pair 9*, RRP1/HPA018166; *pair 10*, RSPH1/HPA016816; *pair 11*, RWDD2B/HPA018316; *pair 12*, S100B/HPA015768; *pair 13*, SAMSN1/HPA017055; *pair 14*, TFF1/HPA003425.

previous evidence on the protein level have been identified by molecular characterization (Fig. 2*b*), and the level of protein modifications has been studied using a new approach for isoelectric focusing based Western blot analysis. Although this analysis was performed only on a small number of genes, the results indicate that a large fraction of the analyzed proteins have multiple isoforms or post-translational modifications. In addition, the tissue profiling using immunohistochemistry has revealed several proteins with highly selective expression patterns.

The protein analysis has been complemented with transcript profiling using next generation sequencing. The results from this analysis provide a useful tool to yield evidence for protein-coding genes as demonstrated by the ratio of reads across introns and exons for a number of chromosome 21 putative genes with no previous evidence on the protein level. The power of the RNA-seq method for transcript analysis can also be further extended to define and characterize the alternative splice variants from each gene locus and to determine the quantitative levels of RNA expression in different cells, tissues, and organs.

At present, we report annotated protein expression using two or more (paired) antibodies for 22% of the genes on chromosome 21. An important priority for the future is to add additional antibodies to allow the results from one antibody to be validated by the other. It will also be important to extend the analysis with renewable antibodies, such as monoclonal antibodies or recombinant affinity binders to complement the polyclonal antibodies generated within the Human Protein Atlas program. In this context, it is reassuring that several programs have been initiated recently to develop new methods for systematic generation of renewable binders to human proteins (30, 31). Another important objective is to extend the validation of the molecular and subcellular localization to include analysis of cell lines in which the gene has been knocked down using siRNA technology. The combination of gene knockdowns and antibody-based profiling is a powerful approach for generating profiling data with high reliability.

In conclusion, we describe a human proteome project to perform a systematic characterization of all the protein-coding genes on human chromosome 21 using antibody-based protein profiling. Through collaboration with research groups utilizing several complementary technologies, this effort can be integrated with similar efforts as part of a Human Proteome Project to characterize the proteins in normal cells, tissues and organs to generate a proteome-wide knowledge-based resource. The objective is to ultimately create an experimentally validated resource covering all proteins encoded by the human genome.

Acknowledgments—We 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 program PROSPECTS. 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.

S This article contains supplemental Tables 1–3 and Figs. 1–4.

¶ To whom correspondence should be addressed: Science for Life Laboratory, Royal Institute of Technology, box 1031, SE-17121 Solna, Sweden. Tel.: 46855378325; E-mail: mathias.uhlen@scilifelab.se.

REFERENCES

1. (2010) A gene-centric human proteome project: HUPO–the Human Proteome Organisation. *Mol. Cell. Proteomics* **9,** 427–429
2. (2010) The call of the human proteome. *Nat. Methods* **7,** 661
3. Hochstrasser, D. (2008) Should the human proteome project be gene- or protein-centric? *J. Proteome Res.* **7,** 5071
4. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., Srivastava, S., Uhlen, M., Wu, C. H., Yamamoto, T., Paik, Y. K., and Omenn, G. S. (2011) The human proteome project: Current state and future direction. *Mol. Cell. Proteomics* **10,** 10.1074/mcp.M111.009993
5. Agaton, C., Galli, J., Höidén Guthenberg, I., Janzon, L., Hansson, M., Asplund, A., Brundell, E., Lindberg, S., Ruthberg, I., Wester, K., Wurtz, D., Höög, C., Lundeberg, J., Ståhl, S., Pontén, F., and Uhlén, M. (2003) Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol. Cell. Proteomics* **2,** 405–414
6. Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A., Amini, B., Andersen, E., Andersson, A. C., Angelidou, P., Asplund, A., Asplund, C., Berglund, L., Bergström, K., Brumer, H., Cerjan, D., Ekström, M., Elobeid, A., Eriksson, C., Fagerberg, L., Falk, R., Fall, J., Forsberg, M., Björklund, M. G., Gumbel, K., Halimi, A., Hallin, I., Hamsten, C., Hansson, M., Hedhammar, M., Hercules, G., Kampf, C., Larsson, K., Lindskog, M., Lodewyckx, W., Lund, J., Lundeberg, J., Magnusson, K., Malm, E., Nilsson, P., Odling, J., Oksvold, P., Olsson, I., Oster, E., Ottosson, J., Paavilainen, L., Persson, A., Rimini, R., Rockberg, J., Runeson, M., Sivertsson, A., Sköllermo, A., Steen, J., Stenvall, M., Sterky, F., Strömberg, S., Sundberg, M., Tegel, H., Tourle, S., Wahlund, E., Waldén, A., Wan, J., Wernérus, H., Westberg, J., Wester, K., Wrethagen, U., Xu, L. L., Hober, S., and Pontén, F. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4,** 1920–1932
7. Nilsson, P., Paavilainen, L., Larsson, K., Odling, J., Sundberg, M., Andersson, A. C., Kampf, C., Persson, A., Al-Khalili Szigyarto, C., Ottosson, J., Björling, E., Hober, S., Wernérus, H., Wester, K., Pontén, F., and Uhlen, M. (2005) Towards a human proteome atlas: High-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **5,** 4327–4337
8. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28,** 1248–1250
9. Barbe, L., Lundberg, E., Oksvold, P., Stenius, A., Lewin, E., Björling, E., Asplund, A., Pontén, F., Brismar, H., Uhlén, M., and Andersson-Svahn, H. (2008) Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* **7,** 499–508
10. Stadler, C., Skogs, M., Brismar, H., Uhlén, M., and Lundberg, E. (2010) A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73,** 1067–1078
11. Pontén, F., Jirström, K., and Uhlen, M. (2008) The Human Protein Atlas: A tool for pathology. *J. Pathol.* **216,** 387–393
12. Kampf, C., Andersson, A. C., Wester, K., Björling, E., Uhlén, M., and Pontén, F. (2004) Antibody-based tissue profiling as a tool for clinical proteomics. *Clin. Proteomics* **1,** 285–299
13. Paavilainen, L., Edvinsson, A., Asplund, A., Hober, S., Kampf, C., Pontén, F., and Wester, K. (2010) The impact of tissue fixatives on morphology and antibody-based protein profiling in tissues and cells. *J. Histochem. Cytochem.* **58,** 237–246
14. Björling, E., Lindskog, C., Oksvold, P., Linné, J., Kampf, C., Hober, S., Uhlén, M., and Pontén, F. (2008) A web-based tool for in silico biomarker discovery based on tissue-specific protein profiles in normal and cancer tissues. *Mol. Cell. Proteomics* **7,** 825–844
15. Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6,** 450
16. Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Vogel, J., and Searle, S. M. (2011) Ensembl 2011. *Nucleic Acids Res.* **39,** D800–D806
17. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38,** D142–D148
18. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39,** D214–D219
19. Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., McGarvey, P., and Gasteiger, E. (2009) Infrastructure for the life sciences: Design and implementation of the UniProt website. *BMC Bioinformatics* **10,** 136
20. Shimomura, Y., and Ito, M. (2005) Human hair keratin-associated proteins. *J. Investig. Dermatol. Symp. Proc.* **10,** 230–233
21. Rogers, M. A., Langbein, L., Winter, H., Beckmann, I., Praetzel, S., and Schweizer, J. (2004) Hair keratin associated proteins: Characterization of a second high sulfur KAP gene domain on human chromosome 21. *J. Invest. Dermatol.* **122,** 147–158
22. Renart, J., Reiser, J., and Stark, G. R. (1979) Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: A method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. U.S.A.* **76,** 3116–3120
23. Khanna, R., Myers, M. P., Lainé, M., and Papazian, D. M. (2001) Glycosylation increases potassium channel stability and surface expression in mammalian cells. *J. Biol. Chem.* **276,** 34028–34034
24. Ben-Dor, S., Esterman, N., Rubin, E., and Sharon, N. (2004) Biases and complex patterns in the residues flanking protein *N*-glycosylation sites. *Glycobiology* **14,** 95–101
25. Castleman, V. H., Romio, L., Chodhari, R., Hirst, R. A., de Castro, S. C., Parker, K. A., Ybot-Gonzalez, P., Emes, R. D., Wilson, S. W., Wallis, C., Johnson, C. A., Herrera, R. J., Rutman, A., Dixon, M., Shoemark, A., Bush, A., Hogg, C., Gardiner, R. M., Reish, O., Greene, N. D., O'Callaghan, C., Purton, S., Chung, E. M., and Mitchison, H. M. (2009) Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am. J. Hum. Genet.* **84,** 197–209
26. Mann, M. (2006) Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7,** 952–958
27. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
28. Righetti, P. G., and Bossi, A. (1997) Isoelectric focusing in immobilized pH gradients: Recent analytical and preparative developments. *Anal. Biochem.* **247,** 1–10
29. Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R., and Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: Principle, methodology and some applications. *J. Biochem. Biophys. Methods* **6,** 317–339
30. Taussig, M. J., Stoevesandt, O., Borrebaeck, C. A., Bradbury, A. R., Cahill, D., Cambillau, C., de Daruvar, A., Dübel, S., Eichler, J., Frank, R., Gibson, T. J., Gloriam, D., Gold, L., Herberg, F. W., Hermjakob, H., Hoheisel, J. D., Joos, T. O., Kallioniemi, O., Koegl, M., Konthur, Z., Korn, B., Kremmer, E., Krobitsch, S., Landegren, U., van der Maarel, S., McCafferty, J., Muyldermans, S., Nygren, P. A., Palcy, S., Pluckthun, A., Polic, B., Przybylski, M., Saviranta, P., Sawyer, A., Sherman, D. J., Skerra, A., Templin, M., Ueffing, M., and Uhlén, M. (2007) ProteomeBinders: Planning a European resource of affinity reagents for analysis of the human proteome. *Nat. Methods* **4,** 13–17
31. Uhlen, M., Gräslund, S., and Sundström, M. (2008) A pilot project to generate affinity reagents to human proteins. *Nat. Methods* **5,** 854–855