Cite as: P. J. Thul *et al.*, *Science* 10.1126/science.aal3321 (2017).

A subcellular map of the human proteome

Peter J. Thul,[1]* Lovisa Åkesson,[1]* Mikaela Wiking,[1] Diana Mahdessian,[1] Aikaterini Geladaki,[2,3] Hammou Ait Blal,[1] Tove Alm,[1] Anna Asplund,[4] Lars Björk,[1] Lisa M. Breckels,[2,5] Anna Bäckström,[1] Frida Danielsson,[1] Linn Fagerberg,[1] Jenny Fall,[1] Laurent Gatto,[2,5] Christian Gnann,[1] Sophia Hober,[6] Martin Hjelmare,[1] Fredric Johansson,[1] Sunjae Lee,[1] Cecilia Lindskog,[4] Jan Mulder,[7] Claire M. Mulvey,[2] Peter Nilsson,[1] Per Oksvold,[1] Johan Rockberg,[6] Rutger Schutten,[1] Jochen M. Schwenk,[1] Åsa Sivertsson,[1] Evelina Sjöstedt,[4] Marie Skogs,[1] Charlotte Stadler,[1] Devin P. Sullivan,[1] Hanna Tegel,[6] Casper Winsnes,[1] Cheng Zhang,[1] Martin Zwahlen,[1] Adil Mardinoglu,[1] Fredrik Pontén,[4] Kalle von Feilitzen,[1] Kathryn S. Lilley,[2] Mathias Uhlén,[1]† Emma Lundberg[1]†

[1]Science for Life Laboratory, School of Biotechnology, KTH - Royal Institute of Technology, SE-171 21 Stockholm, Sweden. [2]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. [4]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden. [5]Computational Proteomics Unit, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [6]Department of Proteomics, School of Biotechnology, KTH - Royal Institute of Technology, SE-106 91 Stockholm, Sweden. [7]Science for Life Laboratory, Department of Neuroscience, Karolinska Institute, SE-171 77 Stockholm, Sweden.

*These authors contributed equally to this work.

†Corresponding author. Email: mathias.uhlen@scilifelab.se (M.U.); emma.lundberg@scilifelab.se (E.L.)

Resolving the spatial distribution of the human proteome at a subcellular level greatly increases our understanding of human biology and disease. Here, we present a comprehensive image-based map of the subcellular protein distribution, the Cell Atlas, built by integrating transcriptomics and antibody-based immunofluorescence microscopy with validation by mass spectrometry. Mapping the in situ localization of 12,003 human proteins at a single-cell level to 30 subcellular structures enabled the definition of 13 major organelle proteomes. Exploration of the proteomes reveals single-cell variations of abundance or spatial distribution, and localization of approximately half of the proteins to multiple compartments. This subcellular map can be used to refine existing protein-protein interaction networks and provides an important resource to deconvolute the highly complex architecture of the human cell.

Spatial partitioning of biological functions is a phenomenon fundamental to life. In humans, this spatial partitioning constitutes a hierarchy of specialized systems ranging across scales, from organs to specialized cells, to subcellular structures, down to macro-molecular complexes. At the cellular level, proteins function at specific times and locations. These subcellular locations such as organelles provide a specific chemical environment and set of interaction partners that are necessary to fulfill the protein's function. Mislocalization of proteins can be associated with cellular dysfunction and disease (*1*, *2*). Thus, knowledge of the spatial distribution of proteins at a subcellular level is essential for understanding protein function, interactions, and cellular mechanisms.

Several approaches for systematic analysis of protein localizations have been described. Quantitative mass-spectrometric readouts allow identification of proteins with similar distribution profiles across fractionation gradients (*3*–*7*) or enzyme-mediated proximity labeled proteins in cells (*8*–*11*). In contrast, imaging-based approaches using tagged proteins (*12*–*14*) or affinity reagents (*15*, *16*) enable exploring the subcellular distribution of proteins in situ in single cells and can also effectively identify cell-to-cell variability and multi-organelle distribution. Complimentary to these experimental methods, a number of *in silico* approaches have been used to predict the subcellular localization in eukaryotic cells (e.g., (*17*, *18*)). The manually curated UniProt database (*19*) is an important resource for protein localization, as it collects subcellular data from literature and external databases for a large number of species. Despite these efforts, experimental data for subcellular localization are lacking for the majority of human proteins. To address this need, pilot studies have been initiated to probe human proteins using immunofluorescence (IF) and high-resolution confocal microscopy (*15*, *20*, *21*) and mass spectrometry (*7*). To date, maps of the subcellular proteome of murine stem cells (*6*), HeLa cells (*7*) and rat liver (*22*) represent the best-characterized data sets for mammals.

Here, we report the establishment of a Cell Atlas within the framework of the Human Protein Atlas (*23*, *24*) to create a comprehensive, proteome-wide knowledge resource for subcellular localization in human cells. By integration of transcriptomics data and an antibody-based image profiling approach, we provide experimental localization data for 12,003 proteins using a panel of 22 human cell lines and 13,993 antibodies. The spatial distribution of these proteins

Downloaded from http://science.sciencemag.org/ on May 11, 2017

is resolved to 30 cellular structures and substructures, altogether representing 13 major organelles. Particular emphases were on defining the organelle proteomes, describing multilocalizing proteins and proteins displaying single-cell variability. We expect the availability of localization information for the human proteome to complement other systematic efforts on the DNA (*25*, *26*), RNA (*27*, *28*), and proteome (*19*, *29*) levels and aid in the molecular understanding of the human cell and its interactions.

Results

Cell lines and transcriptomics analysis

The Cell Atlas aims to define the proteome of organelles and subcellular compartments by IF imaging (Fig. 1). In order to select suitable cell lines for the effort, transcriptomics analysis using RNA-sequencing (RNA-seq) was performed on 56 human cell lines from various origins representing different germ layers and tissues (table S1). A hierarchical clustering analysis based on RNA-seq data (Fig. 2A) shows that cell lines of similar origin or phenotype cluster together, indicating a common pattern of gene expression. Prominent clusters include myeloid and lymphoid cell lines respectively, endothelial cells, and cells immortalized by introduction of telomerase. 22 cell lines were selected for IF imaging; together expressing 84% (16,504) of all protein-coding genes (19,628) predicted by Ensembl (version 83.38 (*26*)) based on a transcripts per million (TPM) cut-off ≥ 1 (table S2). Interestingly, by applying TPM values the average number of expressed genes in the sequenced cell lines is 11,490 genes (table S2) and ranges from 10,136 genes in Daudi cells (B lymphoblast) to 12,816 in SCLC-21H (small cell lung carcinoma). This is notably less than the previously measured average of approximately 14,000 transcripts when using FPKM values (Fragments Per Kilobase of transcript per Million mapped reads) as normalization method. However, the TPM-based number corresponds more accurately to the number of proteins actually detected in the present and other proteomic studies (*30*, *31*).

A classification of the RNA expression levels according to the principle previously described (*24*) was performed to define genes expressed in all cell lines and those expressed in a cell line restricted manner (fig. S1). About one third (6,295) of the protein-coding genes were expressed in all cell lines suggesting a "house-keeping" role, while 45% showed a more variable expression. 11% (2,090) of all protein-coding genes were not detected in any of the analyzed cell lines. Of these genes, 1,225 were detected in tissues suggesting that they code for proteins restricted to a smaller number of specialized cell types or representing specific developmental stages (table S3). Functional annotations from Gene Ontology (GO) support this hypothesis showing enrichment for tissue-restricted proteins, such as receptors in the sensory cells or reproduction-related proteins (table S4).

Creation of a subcellular map

As an integrated part of the Human Protein Atlas effort (*23*), antibodies have been generated, affinity-purified using the antigen, and validated by protein microarray analysis to ensure specific and selective binding to the intended target antigen (*32*). These antibodies cover the majority of all predicted human protein-coding genes. A systematic workflow for subcellular localization of proteins was established using IF and high-resolution confocal microscopy as described in fig. S2 (*15*, *16*). Altogether 13,993 antibodies (13,073 antibodies generated within the HPA project complemented with 920 commercially available antibodies) were selected to be included in the Cell Atlas after reliability analysis. Every antibody was used for immunostaining of the bone osteosarcoma-derived U-2 OS cell line and in two additional cell lines from the panel showing a high expression of the gene. In addition to the antibody of interest, reference markers outlining the nucleus, microtubules and endoplasmic reticulum were included in each sample (fig. S3). For all proteins, the spatial expression pattern observed in the confocal images were assigned to one or more of 30 cellular organelles and substructures (Fig. 1 and table S5), and classified by a location specific reliability score as outlined below. The images and the primary data are presented in the Cell Atlas in a gene-centric manner, including the classification of all images and a description of the validation and reliability of the antibodies and identified locations. Furthermore, the images were annotated by a Citizen Science approach through "Project Discovery" within EVE online, a massive multiplayer online game, resulting in more than 7 million minutes of active participation from over 180,000 players across the world to date. Altogether, the Cell Atlas (version 16.1 of the Human Protein Atlas) contains 82,152 high-resolution annotated images covering 61% of all the human protein-coding genes and 73% of the genes expressed in the IF cell line panel. The complete localization dataset containing the results for all proteins in the Cell Atlas as well as all successful stainings obtained in the different cell lines are found in tables S6 and S7, respectively.

Validation of data in the Cell Atlas

Recently, there have been many articles questioning the quality and use of antibodies in research (e.g., (*33*, *34*)). Since antibody off-target binding can cause false positive results, efforts have gone into manually annotating all antibodies regarding their reliability and quality of the staining. In the Cell Atlas we provide a reliability score for every annotated location and protein at a four-tiered scale: "validated", "supported", "approved" and "uncertain". Locations obtained the score "validated" if the antibody was validated

Downloaded from http://science.sciencemag.org/ on May 11, 2017

according to one of the validation "pillars" proposed by an international working group (*35*) suitable for IF: (i) genetic methods using siRNA silencing (*36*) or CRISPR/Cas9 knockout, (ii) expression of a fluorescent protein-tagged protein at endogenous levels (*37*) or (iii) independent antibodies targeting different epitopes (see fig. S4 for examples). The second tier, "supported" locations, is defined by agreement with external experimental data from the UniProt database. An "approved" location score indicates lack of external experimental information about the protein location. Lastly, an "uncertain" location shows contradictory results compared to complementary information, such as literature or transcriptomics data. In this case, these locations are shown since it still cannot be ruled out that the data are correct. In fig. S5, the distributions of scores for all proteins are shown. 43% of the protein locations provided are in the top two tiers representing a high certainty in the results and half of the proteins are in the "approved" category. Although these proteins have no external evidence to support their location, these antibodies passed our quality tests and show a consistent IF staining. Nevertheless, the likelihood for false results may be higher and should be taken into consideration when looking at individual proteins, while the effect on global proteomic analyses is negligible (fig. S6).

The human organelle proteomes

The spatial information provided by the IF images enabled the development of a subcellular map. The distribution of 12,003 proteins to 30 cellular compartments and substructures is shown in Fig. 2B and detailed in table S8. Taken together, we were able to describe the proteome for 13 major organelles. In addition, we defined a secretome containing proteins secreted through the classical pathway by combining three bioinformatic methods for signal peptide recognition together with seven prediction methods for transmembrane regions (*24*), suggesting that 2,918 proteins are secreted (table S9). Most proteins of the Cell Atlas are found in the nucleoplasm and its substructures (6,245). The number of nuclear proteins exceeds previously identified numbers considerably. Although false nuclear localizations can be observed due to cross-reactivity of antibodies (*21*), the fraction of nuclear locations remain similar in the higher and lower reliability tiers. The second largest number of proteins was identified in the cytosol (4,279) followed by vesicles (1,806), which include transport vesicles as well as small membrane-bound organelles like endosomes or peroxisomes. The nucleoli, including their fibrillar center, contained 1,270 different proteins, which is a more diverse proteome than the mitochondrial or Golgi proteome, although nucleoli are more restricted in their known function. In total, we provide subcellular experimental evidence for 5662 (47%) proteins lacking an experimentally determined GO-term for a cellular compartment. Furthermore, we refined or confirmed subcellular locations for 6341 (53%) proteins already classified by experimentally determined GO-terms (fig. S7).

We further investigated the enrichment of RNA classification categories for the defined organelle proteomes. Figure 2C shows that proteins located to mitochondria, nucleus, nucleoli, and endoplasmic reticulum (ER) are more often expressed in all cell lines, which emphasizes their house-keeping role and important function for cellular survival. On the contrary, proteins with RNA expression pattern categories as "enriched" (expressed 5 times greater than all other cells) and "enhanced" (5 times greater than the mean) are more commonly secretedor located to the plasma membrane, vesicles and the cytoskeleton, which indicate that these compartments play important roles for intercellular communication and adaptation to the surrounding microenvironment. An analogous pattern was seen in the RNA class distribution across 59 human tissues (fig. S8) indicating general similarities in organelle organization between cell lines and tissues.

The goal of proteomic studies lies in the large-scale localization of novel proteins to achieve a complete picture of organelle function. IF-images are in particular advantageous in the identification of protein constituents of compartments that are challenging to purify or have distinct substructures. For example, specialized domains within a compartment such as cell junctions in the plasma membrane are easily visible in IF, like the uncharacterized protein C4orf19 (Fig. 2D). Other compartments such as the cytokinetic bridge correspond to a rare cellular event and are thus challenging for proteomic studies. However, with our high-resolution images we could not only identify 88 proteins located to the cytokinetic bridge (Fig. 2E), but also analyze the underlying components midbody (36 proteins, Fig. 2F) and midbody ring (12 proteins, Fig. 2G). The detection of well-known constituents like CHMP1B to the midbody as well as less characterized proteins like APC2 to the midbody ring or CCSAP to the cytokinetic bridge, results in an enhanced understanding of the final step of cell division. In nucleoli, we identified proteins such as MKI67 that are localized in the rim around the nucleolus and re-organize to line the condensed chromosomes during mitosis (Fig. 2H). As described below, additional tailored assays to complement the Cell Atlas further increase the available information about subcellular locations. The largely uncharacterized dynamic structure rods and rings (R&R) had previously only three known members including IMPDH1 and IMPDH2 (Fig. 2I) (*38*). 21 R&R candidates were discovered and confirmed by actively inducing R&R formation with the compound ribavirin (*38*). The assignment of additional proteins to the R&R sheds new light on

Downloaded from http://science.sciencemag.org/ on May 11, 2017

this structure and provides opportunities for better understanding of its origin, composition and function. In the nucleus, the PML body (marked by SP100, Fig. 2J) was the a prominent substructure. This location can be further explored for selected proteins, as the Cell Atlas contains additional images generated by super-resolution microscopy, allowing a distinction of proteins localizing to the surface (SP100, Fig. 2K,lower image) or to the core (ZBTB8A, Fig. 2K, upper image) of the PML body.

Validation with other proteome-wide data sets

In order to evaluate the overall validity of our data, we controlled its agreement with functional protein information from independent proteome-wide databases. First, we performed a GO "Biological Process" term analysis of the proteome of each organelle. The significantly enriched terms were all related to known key processes of the respective organelle (table S10). Secondly, we performed location enrichment of a set of proteins by a hypergeometric statistical test. In this manner we could demonstrate that the nuclear receptors according to nucleaRDB (*39*) and their co-regulators as defined by NURSA (*40*) were enriched in the nucleus (Fig. 3A and fig. S9) and the group of predicted secreted proteins were enriched in the organelles of the secretory pathway (Golgi apparatus, vesicles, ER) (Fig. 3A). Thirdly, enrichment tests with the mammalian complex database CORUM (*41*) showed similar results (Fig. 3A and fig. S9). Known complexes were significantly enriched in the respective organelle, with exception of the cytoskeleton.

Validation by mass spectrometry

Proteome databases contain information about the subcellular localizations of already characterized proteins; however, our dataset contains a large portion of proteins with a previously uncharacterized location. Therefore, we used an independent approach to reliably validate our annotations. The Cell Atlas data were compared against a high-resolution spatial protein map generated by a mass spectrometry-based method, called hyperplexed localization of organelle proteins by isotope tagging (hyperLOPIT). HyperLOPIT aims to resolve all subcellular compartments in a single experiment by combining biochemical cell fractionation with quantitative mass spectrometry and robust multivariate statistical analysis (*3*, *6*). This enables global identification, quantification and assignment of proteins to their respective subcellular compartments (*42*). The technique does not rely on absolute organelle purification but is based on the measurement of the distribution of cellular proteins across multiple density gradient fractions. Protein localization is assigned by comparing the distribution of proteins of unknown subcellular location with those of unambiguous organelle markers.

The hyperLOPIT pipeline was applied to create a subcellular map of the U-2 OS cell line. Spatial distribution profiles of 5,020 proteins were determined and a Support Vector Machine (SVM) was used to classify 1,971 proteins to 12 discrete subcellular compartments, which were customized to match with the annotations in the Cell Atlas (Fig. 3B). Localization information for a total of 3,626 proteins was available in both the Cell Atlas (U-2 OS only, table S11) and hyperLOPIT (table S12). Out of these, 1,426 proteins were unambiguously classified to a unique location by hyperLOPIT. Within this group, 799 of these proteins were also assigned a unique location in the Cell Atlas, whereas the remaining 627 proteins had Cell Atlas annotations for more than one location.

Two comparisons between the datasets were performed; first, a comparison of proteins shown to be present in only one location in the Cell Atlas data ("unique match", table S13), and secondly, a comparison with all available proteins including those shown to reside in more than one subcellular class in the Cell Atlas, with one unambiguous assignment in the hyperLOPIT dataset ("partial match", table S13). Of the 799 proteins assigned by the Cell Atlas to a single location we found a 76% agreement (unique match) with hyperLOPIT subcellular assignments. For the 1,426 proteins common between the two datasets, an 82% agreement (partial match) was observed between subcellular assignments. However, the overall agreement differs between the four reliability classes given in the Cell Atlas and is only 46% for the "approved" score, which represents 51% of the Cell Atlas dataset (table S13). At the organelle level (table S13), the agreement ranged from 91% and 92% for the ER and mitochondria, respectively, to 60% for vesicles. This lower overlap is expected, since vesicles annotated by the Cell Atlas groups together several organelles and structures that could be analyzed separately using hyperLOPIT. It is clear from Fig. 3C that many Cell Atlas "vesicular" proteins reside in the unclassified intermediate area of the hyperLOPIT dataset when viewed by principal component analysis (PCA). Vesicles are highly dynamic structures which are generated in, and traffic between, different parts of the cell and hence the steady state location of their protein constituents is likely to involve multiple locations, which in the hyperLOPIT data would result in no single, unique classification. The hyperLOPIT workflow involves fractionation of chromatin-associated proteins and nucleoplasm/nucleolus, and this additional fractionation manifests itself as discrete protein correlation patterns. Interrogation of the hyperLOPIT data with Cell Atlas nuclear assignments reveals a nucleolar-like sub cluster in the hyperLOPIT data demonstrating the power of combining data created using orthogonal methods (Fig. 3D).

In the hyperLOPIT dataset, 60% of the proteins identi-

Downloaded from http://science.sciencemag.org/ on May 11, 2017

fied fell into the "unclassified" category. This unclassified category may represent several dynamic scenarios, such as proteins localized to unannotated subcellular structures or multilocalizing proteins. A separate analysis was conducted for the 1,755 proteins, labeled by hyperLOPIT as "unclassified" but which contained subcellular information in the Cell Atlas (fig. S10). Interestingly, the majority of the hyperLOPIT-unclassified proteins are found in the HPA classes 'nucleoplasm', 'vesicles', 'nucleoplasm and cytosol' and 'plasma membrane and cytosol', reflecting the highly dynamic localization of the majority of cellular proteins.

To show the complementary nature of the Cell Atlas and hyperLOPIT at predicting sub-cellular location, we applied a transfer learning method (*43*) to integrate the two data sources. Transfer learning allows one to meaningfully integrate heterogeneous data. By combining labeled marker proteins common in both datasets a significant increase in classifier accuracy was obtained (fig. S11) compared to using the cell atlas alone (p-value < 2e-16). This highlights the strength of integrating the two approaches for the optimal classification of proteins to organelles.

Proteins localized to multiple compartments

In a pilot to this study (*15*), we concluded that many of the studied proteins were not restricted to a single organelle, but localized to one or more additional locations. This observation is supported by the hyperLOPIT data described above and by data for yeast, where 54.3% of the proteins were assigned to multiple localizations (*14*). One of the strengths of imaging-based spatial protein analysis is the ability to localize a protein in situ and simultaneously visualize protein distribution to multiple cellular structures, thus identifying multilocalizing proteins (MLPs). Here, we have classified the main and additional locations for each protein based either on a clear difference in the signal strength or in the occurrence across the tested cell lines. Over 50% (6,163) of the proteins were detected at more than one location, of which 27% (1,649) at three or more locations (table S8). ER and mitochondria mainly contain proteins that are specifically located, while the proteome of the plasma membrane and the nuclear substructures contained mainly MLPs, consistent with the hyperLOPIT data (Fig. 3E and fig. S12). This aspect further agrees with the known biological function of the organelles. While the mitochondria is more self-contained regarding its proteome, the nucleus, plasma membrane and cytosol contain many proteins that operate across organelles in order to regulate metabolic reactions, gene expression, or transmit information from the surrounding environment. The proteome also includes MLPs that vary in their cell-to-cell spatial distribution as well as MLPs that show a cell line-dependent location with different localization in the three cell lines tested, such as ZNF554 (Fig. 4, A to D). In total 3,546 MLPs showed a cell-line dependent localization (table S14).

To investigate if MLPs are organized in superordinate structures, we grouped the individual organelles and substructures into three meta-compartments and searched for distinct patterns within and across these meta-compartments by aligning the proteins on a circular plot (Fig. 4, E to G): Nucleus (nuclear and nucleolar structures), Cytoplasm (cytosol, mitochondria, and the different types of cytoskeleton), and the Secretory Pathway (ER, Golgi apparatus, vesicles, plasma membrane). Within the cytoplasm meta-compartment most MLPs appeared between the cytosol and the cytoskeletal structures and other organelles it embeds (Fig. 4F). Similarly, most MLPs in the nucleus could be identified as a combination of nucleoplasm and the fine structures within, such as nucleoli or nuclear bodies, and likely reflect dynamic translocations of proteins between these proximal compartments (Fig. 4F). The MLPs in the secretory pathway exhibit a sequential pattern likely reflecting the directional protein trafficking (Fig. 4F). This analysis was repeated stratified according to the reliability of locations to control if our results were affected by the data quality (fig. S6). The resulting patterns of multilocalization were highly similar regardless of the dataset used.

Frequent patterns of multilocalization across meta-compartments include "cytosol and nucleus", "cytosol and nucleoli" and "mitochondria and nucleoli" (Fig. 4G). Enrichment analysis of GO "Biological Process" terms of these proteins (table S15) revealed that MLPs of the nucleus and the cytosol are related to transcription and cell cycle regulation, such as UBE2L3 (Fig. 4H); MLPs of the cytosol and nucleoli are enriched for ribosomal proteins such as 60S ribosomal protein L19, which can be also found on the ER (Fig. 4I); and proteins found in both mitochondria and nucleus are related to protein translation and cellular respiration, such as MTIF3 (Fig. 4J) or NDUFA9, respectively. Intriguingly, the meta-compartments secretory pathway and nucleus shared a strikingly high number of MLPs, despite not being in direct physical contact with each other. These MLPs are characterized by their involvement in the regulation of transcription or cell cycle-dependent processes, for example CCAR1 (Fig. 4K). This indicated that the proteomes of the ER, Golgi apparatus, and vesicles are more functional versatile and should not be reduced to their role in protein secretion. In fact, the MLPs create a range of interactions between functionally distant organelles and include them in a network of regulatory processes, which are primarily associated with the nucleus. This may be an indication of the complex network of events surrounding how the cell conveys signals from the exterior to the nucleus.

Downloaded from http://science.sciencemag.org/ on May 11, 2017

Spatial information refines biological networks

The biological function of an organelle is not only defined by the presence or absence of proteins, but by its underlying chain of reactions, which in turn are often conducted by protein-protein interactions. We used the spatial information of the Cell Atlas to examine the relationship between protein interaction partners. For every annotated structure in the Cell Atlas we investigated the subcellular locations for the direct protein interaction partners according to the Reactome database (*44*). Figure 5A shows a heat map for the probability that proteins in one compartment interact directly with proteins in all cellular compartments. In this stringent situation, the majority of the significant enrichments ($p<0.05$) for an interaction pair are found within the same organelle. This compartmental enrichment was even seen for small structures such as nuclear bodies and nucleoli fibrillar centers. The exception was the microtubule-organizing center (MTOC), which showed significant enrichment for interactors found in the centrosome and microtubules. For some structures, proximal structures were also found significantly enriched. Proteins in the plasma membrane for example showed increased probability of directly interacting with proteins in the plasma membrane, cell junctions, the Golgi apparatus, vesicles, focal adhesions, and cytosol. These results support the quality of the locations annotated in the Cell Atlas as direct protein-protein interactions appear in the same or connected compartments. To explore how cellular signaling expands across cellular compartments through reaction pathways the same analysis was performed for the organelle proteomes looking at protein interactions within reaction pathways as defined by Reactome (Fig. 5B). In this analysis, the meta-compartments became more prominent, especially interactions between the organelles of the secretory pathway, and signaling across compartments. Unexpected cross talk across compartments includes apparent interactions between the cytokinetic bridge and nuclear bodies.

We examined whether existing protein-protein interaction networks would benefit from a more comprehensive annotation of a protein's subcellular location, as it constrains the possible number of interaction partners. The localization data was integrated as spatial boundaries to the human interactome that was recently used to systemically uncover the molecular background of human diseases (*45*). The interactome described annotations for 79,020 interactions of 7,827 proteins. By taking the subcellular main location into account, the number decreased to 51,885 (65.7%) interactions of 6,985 proteins that were found likely in vivo (fig. S13). However, a substantial amount of protein interactions could be found when additional locations were included, raising it to 62,352 (78.9%) interactions of 7,494 proteins (fig. S13). This further supports the important functional role of multilocalizing proteins. With this new location-pruned interaction dataset we generated a novel map to identify the most-connective proteins, also called hub proteins, of each compartment (Fig. 5C). The hub proteins of each compartment were mostly different from hubs of the original, non-annotated interactome (table S16), hence our dataset led to the identification of new driver genes within the network. The localization-annotated interactome is available in table S17.

Single-cell variations in protein expression

Protein profiling by IF microscopy allows analysis of expression patterns on a single-cell level to reveal variations of a protein across the analyzed cells. In the Cell Atlas, we labeled proteins with an observed single-cell variation (SCV), like the nucleolar localization of ZNF554 (Fig. 4, A and C). SCV can be observed either in protein expression levels (IF signal intensity) or in the spatial distribution pattern. 1,855 (15%) out of the 12,003 detected proteins showed a SCV (table S18). Further studies are needed to reveal whether the SCV is due to dynamic protein regulation or stochastic events. The majority of these proteins show a variation in protein expression levels (1,671), for example CRYAB (Fig. 6A), whereas 222 proteins showed a variation in spatial distribution (38 proteins fall into both categories). The organelles with most SCV proteins are cytosol (394), nucleoplasm (381), nucleoli (230), and mitochondria (206) (table S8), organelles that also contain most known cell-cycle dependent proteins.

In addition to the subcellular structures that only appear during cell division (mitotic spindle, cytokinetic bridge, midbody and midbody ring), it is plausible to expect a majority of these SCV to also be related to the cell cycle as the cells in the images are growing under asynchronous conditions. To confirm this we used two approaches for a subset of the proteins. First, we stained selected proteins with an observed SCV in the U-2 OS FUCCI cell line (Fluorescence Ubiquitination Cell Cycle Indicator; (*46*)) that allows monitoring of the cell cycle; by this, we could verify a cell cycle-dependent expression of 64 proteins such as ANLN (Fig. 6, B and C; see list of proteins in table S19). The second approach used a computational model to infer the cell cycle position based on features of the microtubule and nucleus reference markers. In this manner the cell cycle position of the cells in the images could be determined on a continuous model and a pseudo-temporal reconstruction allowed the pattern of cell cycle dependency to be modeled. Figure 6D shows such a plot for ANLN that is expressed in cells in the S/G2 phase according to both FUCCI co-localization and the pseudo-temporal computational modeling. Like for SCV, cell cycle-dependent variation could be observed either in a change of the intensity, for example PCNA (Fig. 6E), or in a

Downloaded from http://science.sciencemag.org/ on May 11, 2017

change of the localization, illustrated by the translocation of PSMC6 from nucleoplasm to cytosol (Fig. 6F).

Discussion

Here we present the most comprehensive map of the subcellular distribution of the human proteome, generated by high-resolution IF images on a single-cell level. The results are presented in an interactive resource, the Cell Atlas, as part of the Human Protein Atlas portal (www.proteinatlas.org). This allows exploration of the organelle proteomes, their substructures, single- and multilocalizing proteins, proteins exhibiting single-cell variations in expression or those showing a cell-cycle dependent expression. These defined categories can furthermore be explored in terms of gene expression patterns across a multitude of cell lines based on transcriptome data. To facilitate integration with other biological resources, all data are available for download from the Human Protein Atlas and through collaborations with efforts such as UniProt (*19*), NextProt (*29*), Gene Ontology (*47*), and the pan-European ELIXIR project (*48*).

Spatial partitioning of biological reactions by compartmentalization is an important cellular mechanism for allowing multiple cellular reactions to occur in parallel while avoiding crosstalk. Intriguingly, we identified more than 50% of the analyzed proteins localizing to more than one compartment at the same time. The fact that proteins are localized at multiple sites increases the complexity of the cell from a systems perspective. On one level, it can function as a spatial confinement in order to control the timing of the molecular function in the designated compartment. On another level, multilocalizing proteins are more prone to have diverse protein-protein interactions due to an increased number of potential interaction partners. This is of particular relevance for network analyses and the identification of key hub-proteins that play a crucial role linking complexes to smaller sub-networks leading to a cellular wide network. Moreover, proteins that localize to more than one compartment may have context specific functions increasing the functionality of the proteome. The fact that proteins 'moonlight' in different parts of the cell is now well accepted (*49*, *50*). The high percentage of proteins in multiple locations as indicated by the complementary IF and hyperLOPIT datasets may be an indicator of the scale on which moonlighting occurs. The more complex a system is, the greater number of parts must be sustained in their proper place and with less tolerance for errors, therefore, a high degree of regulation and control is required. To understand cellular function, and in particular in the context of health and disease, detailed knowledge about the cellular system is needed. We demonstrated that current network models benefit from integration of the Cell Atlas localization data as spatial boundaries in order to remove false positive interactions.

Taken together, the proteome of a single cell is compartmentalized and spatiotemporally regulated to a high degree. Protein expression and localization change over time and enable the cell to react to intrinsic or extrinsic factors. Although only presenting a snapshot of the current state for a few cells, our single-cell analysis gives insight to this dynamic process. The high-resolution map of the subcellular localization of 12,003 human proteins provided by the Cell Atlas represents a key resource for a comprehensive understanding of the human cell and its complex underlying molecular machinery, as well as a major step toward modeling the human cell.

Material and methods

Tissue culture cell line cultivation

All cell lines were cultivated at 37°C in a 5% CO_2 humidified environment in the following growth media: Roswell Park Memorial Institute medium (A-431, REH, RH-30, SiHa, SK-MEL-30; Sigma-Aldrich); Dulbecco's Modified Eagle Medium (A549, BJ, HaCaT, HeLa, NTERA, SH-S5Y5; Sigma-Aldrich); Eagle's Minimal Essential Medium (CACO-2, HEK293, HepG2, MCF-7, U-251 MG; Sigma-Aldrich); McCoy's 5A modified (RT-4, U-2 OS; Sigma-Aldrich). Media were always supplemented with 10% fetal bovine serum (FBS, Sigma-Aldrich); additional cell line-specific supplements were: 1% Non-essential amino acids (CACO-2, HeLa, HRK293, HepG2, MCF-7), 1% L-glutamine (CACO-2, HaCaT, HepG2, MCF-7, NTERA, RT-4, U-2 OS), 5% horse serum (NTERA). No antibiotics were used.

AF22 cells were kindly provided by Dr. Falck. They were cultivated in DMEM/F12 supplemented with N-2 (Cat#17502048, Thermo Fisher) and Pen/Strep (Sigma-Aldrich), with freshly added B-27 (1:1000, Cat#12587010, Thermo Fisher), EGF (10 ng/ml, AF-100-15, PeproTech) and FGF2 (10 ng/ml, 100-18B, PeproTech), flask and plates were coated in two steps with poly-N-ornithine (Sigma-Aldrich) and laminin (Sigma-Aldrich).

Telomerase-immortalized cell line HUVEC/TERT2 (Cat# MHT-006-2) and ASC/TERT1 (Cat# MHS-001) were a kind gift by Evercyte GmbH, Vienna, Austria, and were cultured in EndoUp2 and AdipoUp, respectively.

U2-OS FUCCI cells were developed and kindly provided by Dr. Miyawaki (*46*). The cells were cultivated in McCoy's 5A modified medium supplemented with 1% L-glutamine and 10% FBS.

HeLa-Kyoto cell lines stably expressing an EGFP-tagged protein encoded on Bacterial Artificial Chromosome (BAC) were a kind gift from Prof. A. Hyman, Max Planck Institute Dresden, Germany, and were cultivated as described in Skogs *et al.* (*37*, *45*)

Downloaded from http://science.sciencemag.org/ on May 11, 2017

CRISPR/Cas9 knockout and GFP-expressing cells were a kind gift by Horizon Discovery, Cambridge, UK. Their designed HAP1 cell lines were cultivated in IMDM (Iscove's Modified Dulbecco's Medium, Sigma-Aldrich) media supplemented with 10% FBS and 1% Pen/Strep.

All cells were harvested at 60–70% confluency by trypsinization (Trypsin-EDTA solution from Sigma-Aldrich) for splitting or preparing in glass bottom plates.

Antibodies

All antibodies generated and validated within the HPA project were rabbit polyclonal antibodies. They were designed to bind specifically to as many isoforms of the target protein as possible. The antigens consisted of recombinant protein epitope signature tags (PrEST) with a typical length between 50 and 100 amino acids (*51*). The resulting antibodies were affinity purified using the antigen as affinity ligand (*32*). All antibodies used were first approved for sensitivity and lack of cross-reactivity to other proteins, on arrays consisting of glass slides with spotted PrEST fragments. Commercial antibodies were provided by the suppliers and used according to the supplier's recommendations.

Sample preparation for indirect immunofluorescence

A standardized protocol optimized for proteome-wide immunofluorescence localization studies was used, which has previously been described in detail by Stadler *et al.* (*16*). Briefly, cells were seeded in 96-well glass bottom plates (Whatman, Cat# 7716-2370, GE Healthcare, UK, and Greiner Sensoplate Plus, Cat# 655892, Greiner Bio-One, Germany) coated with fibronectin (VWR, Sigma-Aldrich) and grown to a confluency of 60-70% (log-phase growth). PBS-washed cells were fixed in 4% paraformaldehyde (PFA) in growth media supplemented with 10% FBS for 15 min, followed by permeabilization with 0.1% Triton X-100 in PBS for 3x5 min. After a washing step with PBS, cells were incubated with the primary antibody overnight at 4°C. Rabbit polyclonal HPA antibodies were diluted to 2-4 μg/ml in blocking buffer (PBS with 4% FBS) containing 1 μg/ml mouse anti-tubulin (Abcam, ab7291, RRID:AB_2241126, Cambridge, UK), and 1 μg/mL chicken anti-calreticulin (Abcam, ab14234, RRID:AB_2228460) or rat anti-KDEL antibody [MAC 256] (Abcam, ab50601, RRID:AB_880636), respectively. On the next day after 4x10 min washes with PBS, the cells were incubated for 90 min at room temperature with the following secondary antibodies (all from ThermoFisher Scientific) diluted to 1 μg/ml in blocking buffer: goat anti-rabbit AlexaFluor 488 (A11034, RRID:AB_2576217), goat anti-mouse AlexaFluor 555 (A21424, RRID:AB_2535845), and goat anti-chicken AlexaFlour 647 (A-21449, RRID:AB_2535866), or goat anti-rat AlexaFluor 647 (A21247, RRID:AB_1056356), respectively. Cells were subsequently counterstained with DAPI for 10 min. After washing with PBS, the wells were completely filled with 78% glycerol in PBS and sealed.

Fluorescence image acquisition

Fluorescent images were acquired with a Leica SP5 confocal microscope (DM6000CS) equipped with a 63x HCX PL APO 1.40 oil CS objective (Leica Microsystems, Mannheim, Germany). The settings for each image were as followed: Pinhole 1 Airy unit, 16bit acquisition and a pixel size of 0.08 μm. The detector gain measuring the signal of each antibody was adjusted to a maximum of 800 V to avoid strong background noise. The majority of the images were acquired manually from at least two representative field-of-views (FOVs). For proteins displaying single cell variations in their expression pattern, at least six different FOVs were acquired. A small part of the plates were imaged automatically using the MatrixScreener M3 in LAS AF software (Leica Microsystem, Mannheim, Germany). Here, z-stacks at six FOVs were acquired and afterward two images were manually selected for display in the Cell Atlas.

All images on the Cell Atlas are unprocessed with a small compression due to conversion from tiff- to jpg-file format.

IF image annotation

The subcellular location of each protein was manually determined based on the signal pattern and relation to the markers for nucleus (DAPI), microtubules, and endoplasmic reticulum. The annotated locations were as followed: actin filaments, aggresome, cell junctions, centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, endoplasmic reticulum, focal adhesions, Golgi apparatus, intermediate filaments, lipid droplets, microtubule organizing center (MTOC), microtubules, microtubule ends, midbody, midbody ring, mitochondria, mitotic spindle, nuclear bodies, nuclear membrane, nuclear speckles, nucleolar fibrillar center, nucleolar rim, nucleoli, nucleoplasm, nucleus, plasma membrane, rods and rings, and vesicles. If more than one location was detected, they were defined as main or additional location depending on the relative signal strength between the location and the most common location when including all cell lines. Variation between single cells were annotated either as a variation in the intensity or spatial distribution based on a visual inspection. The staining was not annotated if considered negative or unspecific.

Prediction of the human secretome

For the prediction of the human secretome, the analysis was performed as previously described (*24*). Briefly, a majority decision approach was used based on results from three methods for the prediction of signal peptides (SP): SignalP4.0 (*52*), Phobius (*53*) and SPOCTOPUS (*54*). SignalP4.0

Downloaded from http://science.sciencemag.org/ on May 11, 2017

is solely focused on the prediction of SPs whereas the two latter combine the prediction of transmembrane (TM) segments and SPs. In addition, results from the prediction of the human membrane proteome (*55*) were included to classify proteins with a predicted SP as well as one or more TM regions as membrane-spanning. The resulting list of potentially secreted proteins consists of all proteins with a predicted signal peptide by two out of three methods and not including a predicted TM region.

Classification of location reliability

Detected locations were classified based on the reliability of the antibodies and their respective stainings. A score was used for the classification, which incorporated several factors: Reproducibility of the antibody staining in different cell lines (also taken in account when the signal strength correlates with RNA expression); reproducibility of the staining using antibodies binding to different epitopes on the target protein; validation data for the specificity of the antibody (knockdown by siRNA or CRISPR/Cas9 knockout mutants, matching signal with fluorescent-tagged protein); experimental evidence for location described in literature. There were also soft factors such as antibody validation by non IF-related methods like Western blot or immunohistochemistry. The final score led either to the failing of the antibody (approximately 50% of all tested antibodies failed) or to the assignment into one of the following four classes: (i) "validated", if at least one antibody is validated; for example, two independent antibodies show the same localization, that was also observed in experiments outside the HPA or it was supported by e.g., siRNA silencing. (ii) "supported", if there is external experimental data for the location. (iii) "approved", if the localization of the protein has not been previously described and was detected by only one antibody without additional validation. (iv) "uncertain", if the antibody staining is contradicting to experimental data or no expression is detected on the RNA level.

RNA sequencing

Cell lines were selected for IF imaging based on RNA expression of genes (*56*). RNA was extracted from the cells using the RNeasy® kit (Qiagen), generating high quality total RNA (i.e., RIN>8) that was used as input material for library construction with Illumina TruSeq Stranded mRNA reagents. Duplicate samples were sequenced on the Illumina HiSeq2500 platform. Raw sequences were mapped to the Human reference genome GrCh38 and further quantified using the Kallisto software (*57*) to generate normalized Transcript Per Million (TPM) values. TPM values for genes were generated by summing up TPM values for the corresponding transcripts generated by Kallisto. Genes with a TPM value ≥1 were considered expressed.

Location enrichment of protein sets by hypergeometric test

Enrichment of a group of proteins in subcellular locations was examined by hypergeometric tests. In each subcellular location enrichment test, only proteins with subcellular location annotated were considered. Predicted secreted proteins were collected from the Human Protein Atlas (*24*), and nuclear receptors from nucleaRDB (*39*), nuclear receptor coregulators from nuclear receptor signaling atlas (*40*), and subcellular location-specific protein complexes from CORUM (*41*). In CORUM database, nuclear complex proteins were taken from a term "nucleus" in the database; nucleoli complex proteins from "nucleolus"; cytoskeleton complex proteins from "actin cytoskeleton", "microtubule cytoskeleton", and "centrosome" complexes; mitochondria complex proteins from "mitochondrion"; vesicle complex proteins from "intracellular transport vesicle", "peroxisome", and "vacuole or lysosome"; ER complex proteins from "endoplasmic reticulum"; plasma membrane complex proteins from "plasma membrane/membrane attached" and "cell junction"; and cytoplasm complex proteins from "cytoplasm".

HyperLOPIT comparison with Cell Atlas annotations

To compare the subcellular assignments by both methods it was necessary to match the 12 subcellular organelle definitions used by hyperLOPIT to the 30 image categories defined in the Cell Atlas. The comparison was broken down into the following subclasses: all Cell Atlas sub-nuclear categories ("nucleus", "nucleoplasm", "nuclear speckles", "nuclear bodies", "nucleoli", "nucleoli fibrillar center" and "nuclear membrane") were individually compared to a single hyperLOPIT nuclear class encompassing both hyperLOPIT terms "nucleus" and "nuclear chromatin"; the Cell Atlas term for "vesicles" was compared against the combined hyperLOPIT terms for "lysosome" and "peroxisome" (consistent with the Cell Atlas definition for vesicles); and the Cell Atlas class "cell junctions" was compared against the hyperLOPIT term "plasma membrane". For the Cell Atlas terms called "plasma membrane", "mitochondria", "endoplasmic reticulum", "Golgi apparatus" and "cytosol/cytoplasm" the same terms are also available for hyperLOPIT and thus a direct comparison was performed. Proteins that were assigned by hyperLOPIT to the large protein complexes such as ribosomal subunits and proteasome were excluded from the comparison.

Heat maps for protein-protein interaction

Protein-protein interaction pairs were obtained from the independent Reactome database (Downloaded September 20, 2016) (*44*). A binomial test was used to compare the observed frequency of a target protein (Protein B) localizing to

Downloaded from http://science.sciencemag.org/ on May 11, 2017

a given compartment with the expected frequency based on all annotations in the Cell Atlas. Here, the likelihood of localizations of the first protein in the pair (Protein A) can be ignored, as under the null hypothesis it has no impact on the localization of Protein B. The test therefore becomes the probability that we observe at least as many instances of Protein B in a specific compartment given the number of "tries" (instances of Protein A) and the background distribution of proteins over the locations in the Cell Atlas. The background distribution of locations was constructed by taking the frequency of each annotated location for proteins in in the Cell Atlas over the total number of proteins annotated in the Cell Atlas.

The results of the test were visualized using a heat map of p-values (Fig. 5, A and B) where rows represent the location of Protein A and columns represent the location of Protein B. Values are therefore the probability of seeing Protein B in the given compartment at least as frequently as it was actually observed assuming the background distribution. The Bonferroni multiple-hypothesis correction applied per-row to correct for the number of locations being tested for in each pairing. By then considering the correlation of the protein-protein interaction pair locations, key insights into the nature and quality of the data in the Cell Atlas can be gained.

The Reactome database contains several types of protein-protein interactions that can be used to assess different properties of the Cell Atlas annotations. To assess the quality of annotation, we first analyzed direct interactions reasoning that interacting proteins must occupy the same physical space at some point in the cell cycle and therefore should be localized either to the same compartment or adjacent compartments (Fig. 5A).

The same analysis was further performed for protein pairs listed as belonging to the same reaction pathway as defined by the Reactome database to assess what compartments potentially interact through signal cascades (Fig. 5B). This analysis was created using MATLAB2016a.

Figure generation

Plots were generated using R studio (v. 3.3.1) and the additional ggplot2 package. The cell line hierarchical clustering was based on the Spearman correlation of the RNA sequencing data for each cell line. The average distance was used to determine the hierarchical clusters and visualized then by the R package ggdendro. The circular plots showing distribution of multilocalizing proteins were created using the Circos software (v. 0.69) (*58*). The image montages were created using FIJI ImageJ (v 2.0.0-rc-49/1.51f).

Gene Ontology terms and functional enrichment

To check the overlap with Gene Ontology annotations for proteins in the Cell Atlas, the web based tool QuickGO (*59*) was used to acquire GO-annotations for all genes using filters for Cellular Component and information source (downloaded February 15th, 2017). The GO annotations based on data from the Cell Atlas were removed, and the Ensembl IDs for all Cell Atlas genes were then used for checking the overlap of genes with experimental evidence for any GO annotation. The functional annotation clustering for the genes not expressed in the Cell Atlas cell line panel was performed using the web based tool DAVID (Database for Annotation, Visualization, and Integrated Discovery v. 6.8) (*60*). All human genes were used as a background and the GO domain Biological Process terms with Bonferroni value of less than 0.01 were regarded as significantly enriched.

Location-pruned protein-protein interactions

Proteins interactions were obtained from published protein interactome data (*45*); among those protein interactions, only interactions with "signaling", "kinase", "complex", "literature", and "binary" types were taken; this indicates direct protein interactions. Those protein interactions were pruned to proteins localized in the same subcellular locations; in either cytoplasm or plasma membrane; or in either cytoplasm or cytoskeleton. Location-pruned protein interactions were visualized (Fig. 5C) through the edge-weighted spring embedded layout of Cytoscape (*61*) and their nodes were colored by the least frequent one of subcellular locations they have. In each subcellular location, hub proteins from protein interactions of given subcellular locations were examined based on their degree connectivity.

REFERENCES AND NOTES

1. K. Laurila, M. Vihinen, Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10**, 122 (2009). doi:10.1186/1471-2164-10-122 Medline
2. S. Park, J.-S. Yang, Y.-E. Shin, J. Park, S. K. Jang, S. Kim, Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.* **7**, 494 (2011). doi:10.1038/msb.2011.29 Medline
3. T. P. Dunkley, R. Watson, J. L. Griffin, P. Dupree, K. S. Lilley, Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3**, 1128–1134 (2004). doi:10.1074/mcp.T400009-MCP200 Medline
4. L. J. Foster, C. L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. K. Mootha, M. Mann, A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187–199 (2006). doi:10.1016/j.cell.2006.03.022 Medline
5. L. Jakobsen, K. Vanselow, M. Skogs, Y. Toyoda, E. Lundberg, I. Poser, L. G. Falkenby, M. Bennetzen, J. Westendorf, E. A. Nigg, M. Uhlen, A. A. Hyman, J. S. Andersen, Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO J.* **30**, 1520–1535 (2011). doi:10.1038/emboj.2011.63 Medline
6. A. Christoforou, C. M. Mulvey, L. M. Breckels, A. Geladaki, T. Hurrell, P. C. Hayward, T. Naake, L. Gatto, R. Viner, A. Martinez Arias, K. S. Lilley, A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **7**, 9992 (2016). doi:10.1038/ncomms9992 Medline
7. D. N. Itzhak, S. Tyanova, J. Cox, G. H. Borner, Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016). doi:10.7554/eLife.16950 Medline
8. K. J. Roux, D. I. Kim, B. Burke, BioID: A screen for protein-protein interactions. *Curr. Protoc. Protein Sci.* **74**, 19.23.1–19.23.14 (2013). Medline

Downloaded from http://science.sciencemag.org/ on May 11, 2017

9. H.-W. Rhee, P. Zou, N. D. Udeshi, J. D. Martell, V. K. Mootha, S. A. Carr, A. Y. Ting, Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013). doi:10.1126/science.1230593 Medline
10. V. Hung, P. Zou, H.-W. Rhee, N. D. Udeshi, V. Cracan, T. Svinkina, S. A. Carr, V. K. Mootha, A. Y. Ting, Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Mol. Cell* **55**, 332–341 (2014). doi:10.1016/j.molcel.2014.06.003 Medline
11. S.-Y. Lee, M.-G. Kang, J.-S. Park, G. Lee, A. Y. Ting, H.-W. Rhee, APEX fingerprinting reveals the subcellular localization of proteins of interest. *Cell Rep.* **15**, 1837–1847 (2016). doi:10.1016/j.celrep.2016.04.064 Medline
12. J. C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, S. Wiemann, Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1**, 287–292 (2000). doi:10.1093/embo-reports/kvd058 Medline
13. W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O'Shea, Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003). doi:10.1038/nature02026 Medline
14. Y. T. Chong, J. L. Y. Koh, H. Friesen, S. K. Duffy, M. J. Cox, A. Moses, J. Moffat, C. Boone, B. J. Andrews, Yeast proteome dynamics from single cell imaging and automated analysis. *Cell* **161**, 1413–1424 (2015). doi:10.1016/j.cell.2015.04.051 Medline
15. L. Barbe, E. Lundberg, P. Oksvold, A. Stenius, E. Lewin, E. Björling, A. Asplund, F. Pontén, H. Brismar, M. Uhlén, H. Andersson-Svahn, Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* **7**, 499–508 (2008). doi:10.1074/mcp.M700325-MCP200 Medline
16. C. Stadler, M. Skogs, H. Brismar, M. Uhlén, E. Lundberg, A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73**, 1067–1078 (2010). doi:10.1016/j.jprot.2009.10.012 Medline
17. P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, K. Nakai, WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **35**, W585–W587 (2007). doi:10.1093/nar/gkm259 Medline
18. K. C. Chou, Z. C. Wu, X. Xiao, iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **8**, 629–641 (2012). doi:10.1039/C1MB05420A Medline
19. UniProt Consortium, UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015). doi:10.1093/nar/gku989 Medline
20. L. Fagerberg, C. Stadler, M. Skogs, M. Hjelmare, K. Jonasson, M. Wiking, A. Åbergh, M. Uhlén, E. Lundberg, Mapping the subcellular protein distribution in three human cell lines. *J. Proteome Res.* **10**, 3766–3777 (2011). doi:10.1021/pr200379a Medline
21. C. Stadler, E. Rexhepaj, V. R. Singan, R. F. Murphy, R. Pepperkok, M. Uhlén, J. C. Simpson, E. Lundberg, Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods* **10**, 315–323 (2013). doi:10.1038/nmeth.2377 Medline
22. M. Jadot, M. Boonen, J. Thirion, N. Wang, J. Xing, C. Zhao, A. Tannous, M. Qian, H. Zheng, J. K. Everett, D. F. Moore, D. E. Sleat, P. Lobel, Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* **16**, 194–212 (2017). Medline
23. M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, F. Ponten, Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010). doi:10.1038/nbt1210-1248 Medline
24. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi:10.1126/science.1260419 Medline
25. K. D. Pruitt, T. Tatusova, G. R. Brown, D. R. Maglott, NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012). doi:10.1093/nar/gkr1079 Medline
26. A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, P. Flicek, Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016). doi:10.1093/nar/gkv1157 Medline
27. H. Kawaji, J. Severin, M. Lizio, A. R. Forrest, E. van Nimwegen, M. Rehli, K. Schroder, K. Irvine, H. Suzuki, P. Carninci, Y. Hayashizaki, C. O. Daub, Update of the FANTOM web resource: From mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.* **39**, D856–D860 (2011). Medline
28. K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, E. T. Dermitzakis; GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015). doi:10.1126/science.1262110 Medline
29. P. Gaudet, P.-A. Michel, M. Zahn-Zabal, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, A. Gleizes, M. Pereira, D. Teixeira, Y. Zhang, L. Lane, A. Bairoch, The neXtProt knowledgebase on human proteins: Current status. *Nucleic Acids Res.* **43**, D764–D770 (2015). doi:10.1093/nar/gku1178 Medline
30. M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, R. Aebersold, The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549–549 (2011). doi:10.1038/msb.2011.82 Medline
31. T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11**, 014050 (2012). doi:10.1074/mcp.M111.014050 Medline
32. P. Nilsson, L. Paavilainen, K. Larsson, J. Ödling, M. Sundberg, A.-C. Andersson, C. Kampf, A. Persson, C. Al-Khalili Szigyarto, J. Ottosson, E. Björling, S. Hober, H. Wernérus, K. Wester, F. Pontén, M. Uhlen, Towards a human proteome atlas: High-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **5**, 4327–4337 (2005). doi:10.1002/pmic.200500072 Medline
33. J. Bordeaux, A. Welsh, S. Agarwal, E. Killiam, M. Baquero, J. Hanna, V. Anagnostou, D. Rimm, Antibody validation. *Biotechniques* **48**, 197–209 (2010). doi:10.2144/000113382 Medline
34. M. Baker, Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274–276 (2015). doi:10.1038/521274a Medline
35. M. Uhlen, A. Bandrowski, S. Carr, A. Edwards, J. Ellenberg, E. Lundberg, D. L. Rimm, H. Rodriguez, T. Hiltke, M. Snyder, T. Yamamoto, A proposal for validation of antibodies. *Nat. Methods* **13**, 823–827 (2016). Medline
36. C. Stadler, M. Hjelmare, B. Neumann, K. Jonasson, R. Pepperkok, M. Uhlén, E. Lundberg, Systematic validation of antibody binding and protein subcellular localization using siRNA and confocal microscopy. *J. Proteomics* **75**, 2236–2251 (2012). doi:10.1016/j.jprot.2012.01.030 Medline

Downloaded from http://science.sciencemag.org/ on May 11, 2017

37. M. Skogs, C. Stadler, R. Schutten, M. Hjelmare, C. Gnann, L. Björk, I. Poser, A. Hyman, M. Uhlén, E. Lundberg, Antibody validation in bioimaging applications based on endogenous expression of tagged proteins. *J. Proteome Res.* **16**, 147–155 (2017). Medline
38. G. Covini, W. C. Carcamo, E. Bredi, C. A. von Mühlen, M. Colombo, E. K. L. Chan, Cytoplasmic rods and rings autoantibodies developed during pegylated interferon and ribavirin therapy in patients with chronic hepatitis C. *Antivir. Ther.* **17**, 805–811 (2012). doi:10.3851/IMP1993 Medline
39. B. Vroling, D. Thorne, P. McDermott, H. J. Joosten, T. K. Attwood, S. Pettifer, G. Vriend, NucleaRDB: Information system for nuclear receptors. *Nucleic Acids Res.* **40**, D377–D380 (2012). Medline
40. S. A. Ochsner, C. M. Watkins, B. S. LaGrone, D. L. Steffen, N. J. McKenna, Research resource: Tissue-specific transcriptomics and cistromics of nuclear receptor signaling: a web research resource. *Mol. Endocrinol.* **24**, 2065–2069 (2010). doi:10.1210/me.2010-0216 Medline
41. A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H.-W. Mewes, CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010). doi:10.1093/nar/gkp914 Medline
42. L. Gatto, L. M. Breckels, S. Wieczorek, T. Burger, K. S. Lilley, Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* **30**, 1322–1324 (2014). doi:10.1093/bioinformatics/btu013 Medline
43. L. M. Breckels, S. B. Holden, D. Wojnar, C. M. Mulvey, A. Christoforou, A. Groen, M. W. B. Trotter, O. Kohlbacher, K. S. Lilley, L. Gatto, Learning from heterogeneous data sources: An application in spatial proteomics. *PLOS Comput. Biol.* **12**, e1004920 (2016). doi:10.1371/journal.pcbi.1004920 Medline
44. A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The Reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016). doi:10.1093/nar/gkv1351 Medline
45. J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabasi, Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015). doi:10.1126/science.1257601 Medline
46. A. Sakaue-Sawano, H. Kurokawa, T. Morimura, A. Hanyu, H. Hama, H. Osawa, S. Kashiwagi, K. Fukami, T. Miyata, H. Miyoshi, T. Imamura, M. Ogawa, H. Masai, A. Miyawaki, Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008). doi:10.1016/j.cell.2007.12.033 Medline
47. Gene Ontology Consortium, Gene Ontology Consortium: Going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015). doi:10.1093/nar/gku1179 Medline
48. L. C. Crosswell, J. M. Thornton, ELIXIR: A distributed infrastructure for European biological data. *Trends Biotechnol.* **30**, 241–242 (2012). doi:10.1016/j.tibtech.2012.02.002 Medline
49. C. E. Chapple, B. Robisson, L. Spinelli, C. Guien, E. Becker, C. Brun, Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* **6**, 7412 (2015). doi:10.1038/ncomms8412 Medline
50. K.-W. Min, S.-H. Lee, S. J. Baek, Moonlighting proteins in cancer. *Cancer Lett.* **370**, 108–116 (2016). doi:10.1016/j.canlet.2015.09.022 Medline
51. M. Lindskog, J. Rockberg, M. Uhlén, F. Sterky, Selection of protein epitopes for antibody production. *Biotechniques* **38**, 723–727 (2005). doi:10.2144/05385ST02 Medline
52. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). doi:10.1038/nmeth.1701 Medline
53. L. Käll, A. Krogh, E. L. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004). doi:10.1016/j.jmb.2004.03.016 Medline
54. H. Viklund, A. Bernsel, M. Skwark, A. Elofsson, SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928–2929 (2008). doi:10.1093/bioinformatics/btn550 Medline
55. L. Fagerberg, K. Jonasson, G. von Heijne, M. Uhlén, L. Berglund, Prediction of the human membrane proteome. *Proteomics* **10**, 1141–1149 (2010). doi:10.1002/pmic.200900258 Medline
56. F. Danielsson, M. Wiking, D. Mahdessian, M. Skogs, H. Ait Blal, M. Hjelmare, C. Stadler, M. Uhlén, E. Lundberg, RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. *J. Proteome Res.* **12**, 299–307 (2013). doi:10.1021/pr3009308 Medline
57. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016). doi:10.1038/nbt.3519 Medline
58. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). doi:10.1101/gr.092759.109 Medline
59. D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, R. Apweiler, QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009). doi:10.1093/bioinformatics/btp536 Medline
60. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2008). doi:10.1038/nprot.2008.211 Medline
61. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003). doi:10.1101/gr.1239303 Medline
62. R. Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2013); http://www.R-project.org/.
63. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004). doi:10.1186/gb-2004-5-10-r80 Medline
64. L. Gatto, K. S. Lilley, MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **28**, 288–289 (2012). doi:10.1093/bioinformatics/btr645 Medline
65. M. W. Trotter, P. G. Sadowski, T. P. Dunkley, A. J. Groen, K. S. Lilley, Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *Proteomics* **10**, 4213–4219 (2010). doi:10.1002/pmic.201000359 Medline

ACKNOWLEDGMENTS

We acknowledge the staff of the Human Protein Atlas program and Science for Life Laboratory for valuable contributions. We also acknowledge support from Science for Life Laboratory: the National Genomics Infrastructure (NGI) and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure, the National Laboratories for Chemical Biology at Karolinska Institutet (LCBKI) for access to imaging infrastructure, the Advanced Light Microscopy (ALM) for super-resolution microscopy. We acknowledge the following for providing reagents and cell lines to the project: Dr. Hisao Masai (Tokyo Metropolitan Institute of Medical Science) for the stable U-2 OS FUCCI cell line. Dr. Anthony Hyman (Max Planck Institute) for HeLa cells expressing GFP-tagged proteins. Dr. Anna Falk (Karolinska Institute) for the AF22 cells. Evercyte for the ASC TERT1, HUVEC TERT2, HBF TERT88, HTCEpi, LHCN-M2, RPTEC TERT cells. Horizon Discoveries for the Hap1 CRISPR knockout cells, Dr. Rainer Pepperkok and Dr. Beate Neumann for prepared solid-phase siRNA plates. Finally we acknowledge CCP Games and MMOS for creation of Project Discovery, and the players of Eve Online for classification of protein patterns in images through Project Discovery. Funding was provided by the Knut and Alice Wallenberg Foundation and Erling Persson Foundation to M.U, and the Science for Life Laboratory to E.L. CMM and LMB were funded by Wellcome Trust grant 108441/Z/15/Z. A.G. was funded through the Alexander S. Onassis Public Benefit Foundation, the Foundation for Education and European Culture (IPEP) and the Embiricos Trust Scholarship of Jesus College Cambridge. MU, FP, PN and SH are co-founders and MU and FP are on the board of Atlas antibodies. All antibodies are available from various commercial providers, including Atlas antibodies. Correspondence related to this article should be addressed to E.L or MU. The data are available for download in the supplementary material. Images and Cell Atlas transcriptome and proteome data are available in the Human Protein Atlas (www.proteinatlas.org/humancell).

Downloaded from http://science.sciencemag.org/ on May 11, 2017

SUPPLEMENTARY MATERIALS
www.sciencemag.org/cgi/content/full/science.aal3321/DC1
Additional Materials and Methods
Figs. S1 to S13
Tables S1 to S19
References (*62*–*65*)

2 November 2016; accepted 31 March 2017
Published online 11 May 2017
10.1126/science.aal3321

Downloaded from http://science.sciencemag.org/ on May 11, 2017

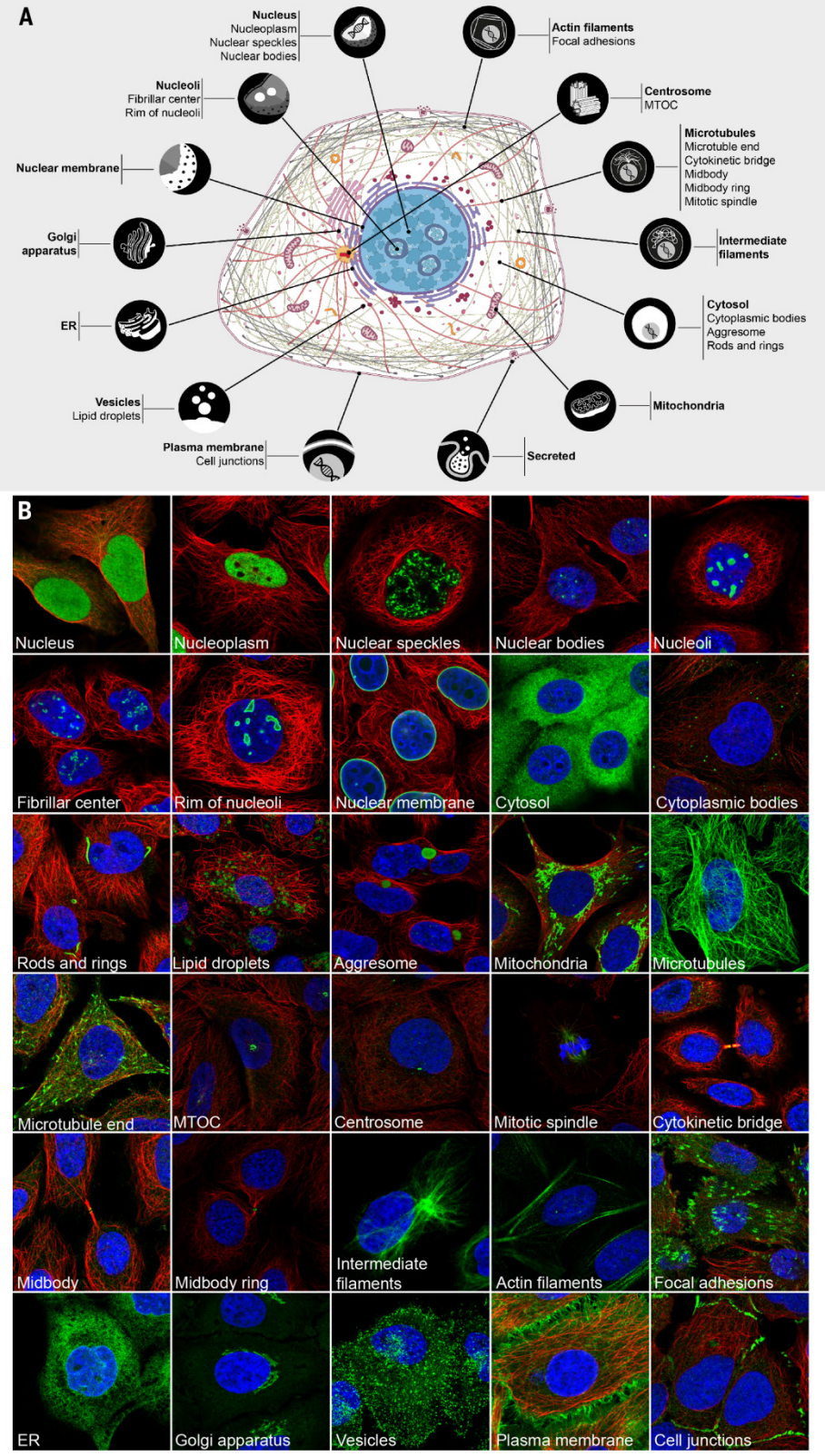
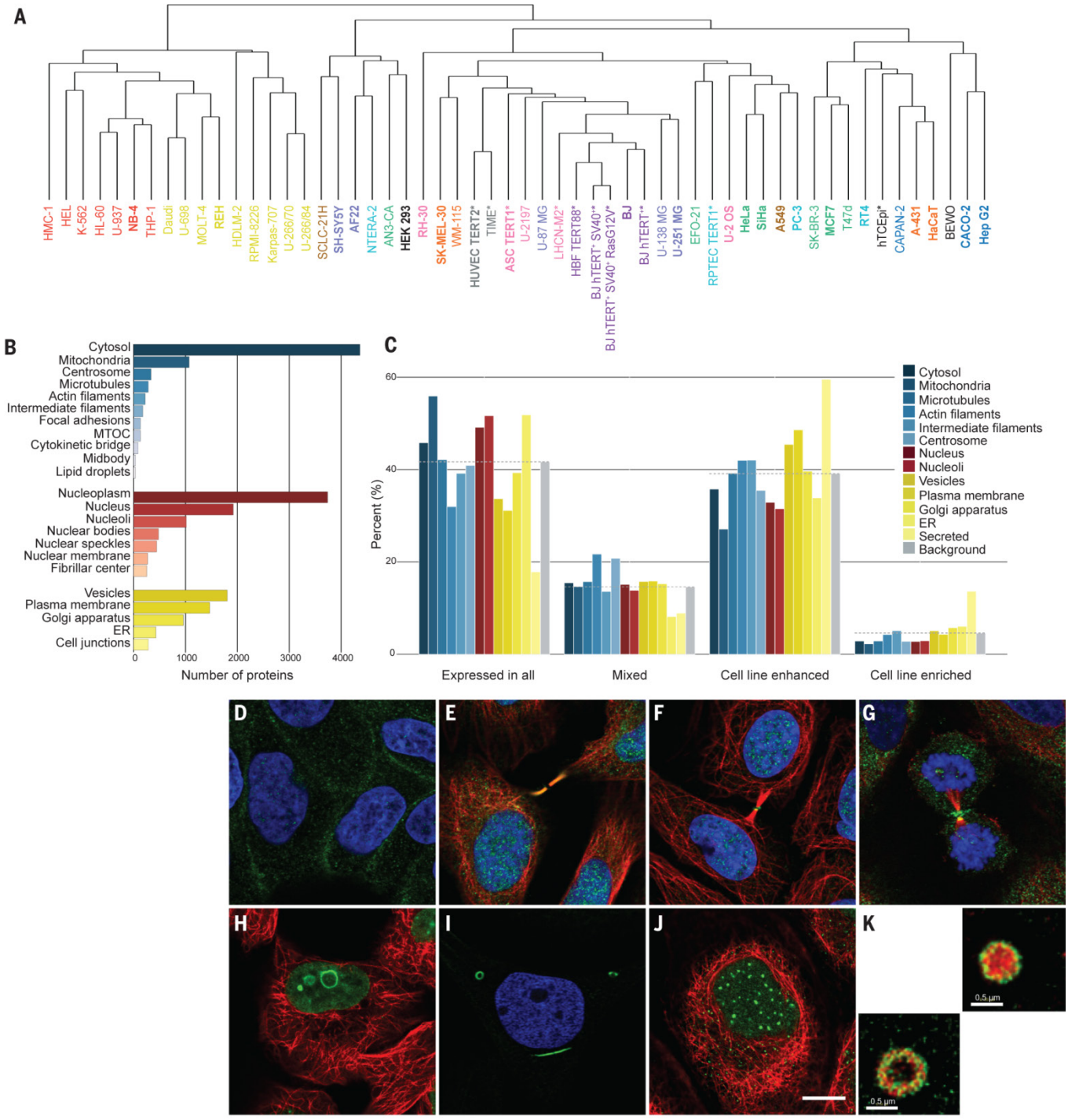


Fig. 1. Subcellular locations in the Cell Atlas. (**A**) Schematic overview of the cell. 13 subcellular proteomes as well as a proteome of secreted proteins were defined in the Cell Atlas by determining the protein localization to 30 subcellular structures. (**B**) Subcellular structures annotated in the Cell Atlas by immunofluorescence microscopy. Examples of proteins (in green) localizing to all annotated structures in a representative set of human cell lines used in the Cell Atlas. Microtubules are marked with an anti-tubulin antibody (red); the nucleus is counterstained with DAPI (blue). The side of an image represents 64 μm. For information about cell lines, antibodies and proteins see table S6.

Downloaded from http://science.sciencemag.org/ on May 11, 2017



Downloaded from http://science.sciencemag.org/ on May 11, 2017

Fig. 2. Transcriptomics and proteomics. (**A**) mRNA deep sequencing was performed for 56 cell lines. The cell lines were clustered based on the gene expression patterns. The color of the cell line name represents its origin: red – myeloid, yellow – lymphoid, brown – lung, dark blue – brain, turquoise - renal, urinary and male reproductive system, green – breast and female reproductive system, pink – sarcoma, purple – fibroblast, dark blue – abdominal, black - miscellaneous. Cells immortalized by the introduction of telomerase are indicated by an asterisk (*). Cell lines in bold are included in the Cell Atlas cell line panel. (**B**) Number of proteins per subcellular location. 12,003 proteins were localized to one or more subcellular compartments in this study. Locations are sorted and color-coded according to the number and the meta-compartments "cytoplasm" (cytosol and embedded organelles), "nucleus" (nuclear and nucleolar structures), and "secretory pathway" (ER, Golgi apparatus, vesicles, and plasma membrane). Note that some locations are merged: aggresomes and R&R to cytosol, microtubule ends and mitotic spindle to microtubules, midbody ring to midbody. (**C**) RNA classification categories per major organelle (nucleus and nuclear membrane are merged) compared to the background of genes in the Cell Atlas. Genes with a TPM value of ≥1 were considered as expressed and classified either as "expressed in all cell lines", "enriched" (expression in one cell line at least five-fold higher than in all other cell lines), "enhanced" (five-fold higher average TPM level in one or more cell lines compared to the mean TPM of all cell lines) or "mixed" (expressed but not one of the other categories). (**D**) C4orf19 (detected by HPA043458 in RT4 cells) localized to cell junctions, a subdomain of the plasma membrane (**E** to **G**) Protein localizing at the final stage of the cytokinesis: CCSAP to the cytokinetic bridge (detected by HPA028402 in U-2 OS cells (E)), CHMP1B to the midbody (detected by HPA061997 in SiHa cells (F)), and APC2 to the midbody ring (detected by HPA078002 in U-2 OS cells (G)). (**H**) MKI67 (detected by CAB000058 in U-251 MG cells) localized to the rim of nucleoli. (**I**) Novel protein C21orf59 (detected by CAB034170 in U-2 OS cells) localized to R&R, whose formation was induced by ribavirin. (**J** and **K**) Conventional IF images in the Cell Atlas (J) and super-resolution images acquired with STED (K) of a PML-body (a class of nuclear bodies) show the surface of the body marked by PML (red) and the shell protein SP100 (HPA016707, green) or the core protein ZBTB8A (HPA031768, green); scale represents 10 µm in (D) to (J) and 0.05 µm in (K).

Downloaded from http://science.sciencemag.org/ on May 11, 2017

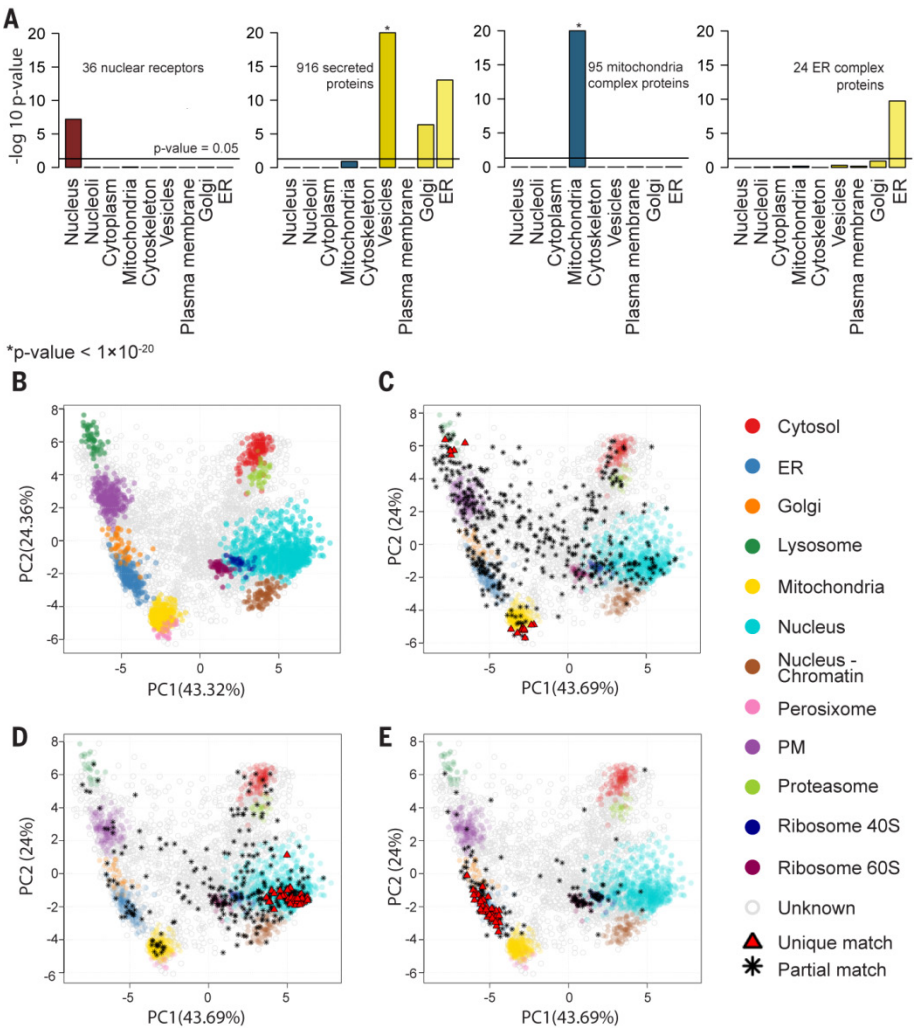
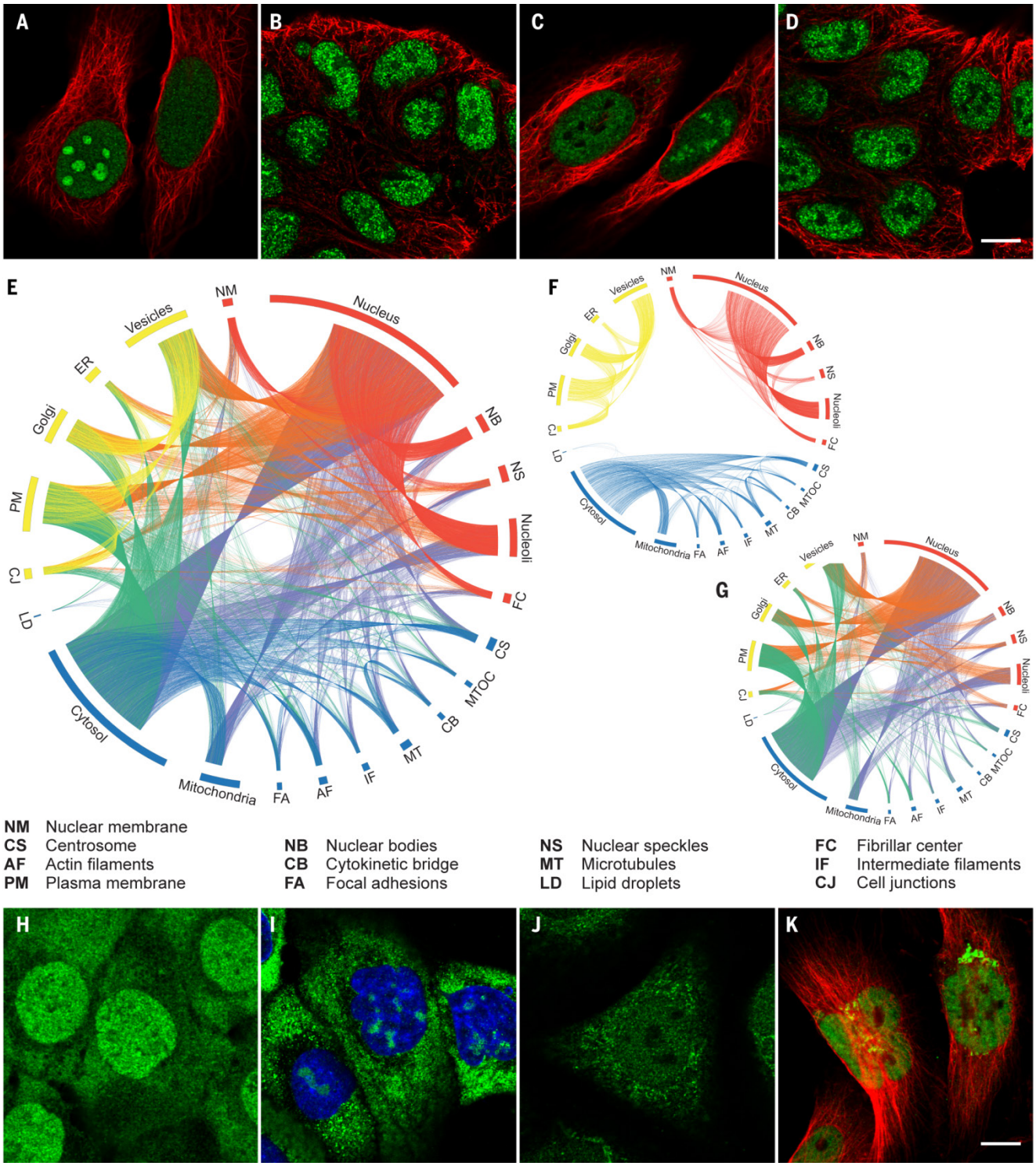


Fig. 3. Validation by proteome-wide databases and hyperLOPIT. (**A**) Location enrichment analyses of different protein sets. Hypergeometric tests were performed to evaluate subcellular locations (P-value < 0.05). Nuclear receptors were enriched in the nuclear meta-compartment. Predicted secreted proteins were enriched in organelles of the secretory pathway, ER, Golgi apparatus and vesicles. Members of known complexes according to the CORUM database were enriched in the respective organelle, for instance mitochondria and ER. (**B**) A principal components analysis (PCA) representation of the human U-2 OS cell hyperLOPIT data (5,020 proteins common across two hyperLOPIT replicates). One point represents one protein and proteins cluster according to their density gradient distribution. Colored circles form subcellular compartments that have been classified by a SVM. For the statistical comparison to the Cell Atlas, hyperLOPIT subcellular annotations were matched with their equivalent Cell Atlas definition. (**C** to **E**) PCA plots of the U-2 OS human dataset for (C) vesicles (D) nucleoli, and (E) the ER. Proteins common in both the Cell Atlas and hyperLOPIT datasets are visualized (3,626 proteins). Black stars represent partial matches and red triangles represent unique matches between organelles in the HPA and hyperLOPIT datasets.

Downloaded from http://science.sciencemag.org/ on May 11, 2017



Downloaded from http://science.sciencemag.org/ on May 11, 2017

Fig. 4. Multilocalizing proteins in the human proteome. (**A** to **D**) ZNF554 is an example of a cell line-dependent subcellular localization. Two antibodies, HPA060247 (left, A and B) and HPA063358 (right, C and D), binding different epitopes detected ZNF554 in both the nucleoplasm and nucleoli in U-2 OS cells, whereas only in the nucleoplasm in RT4 and SH-SY5Y (not shown). The nucleolar expression was only detected in a fraction of the U-2 OS cells and thus showed additionally a single-cell variation. Scale bar represents 10 μm. (**E** to **G**) Circular plots with the identified proteins per compartment presented and sorted by meta-compartments. Multilocalizing proteins appearing more than once in the plots are connected by an edge. (**E**) Circular plot with all meta-compartments and proteins combined. (**F**) Circular plot visualizing connections only within a meta-compartment (**G**) Circular plot visualizing connections only across meta-compartments (**H** to **K**) Examples for dual localizations: (**H**) UBE2L3 in nucleus and cytosol (detected by HPA062415 in A-431 cells), (**I**) 60S ribosomal protein L19 in nucleoli and cytosol (detected by HPA043014 in U-2 OS cells), (**J**) MTIF in nucleus and mitochondria (detected by HPA039791 in U-2 OS cells), (**K**) CCAR1 in Golgi apparatus and nucleoplasm (detected by HPA007856 in U-251 MG cells). Scale bar represents 10 μm.

Downloaded from http://science.sciencemag.org/ on May 11, 2017

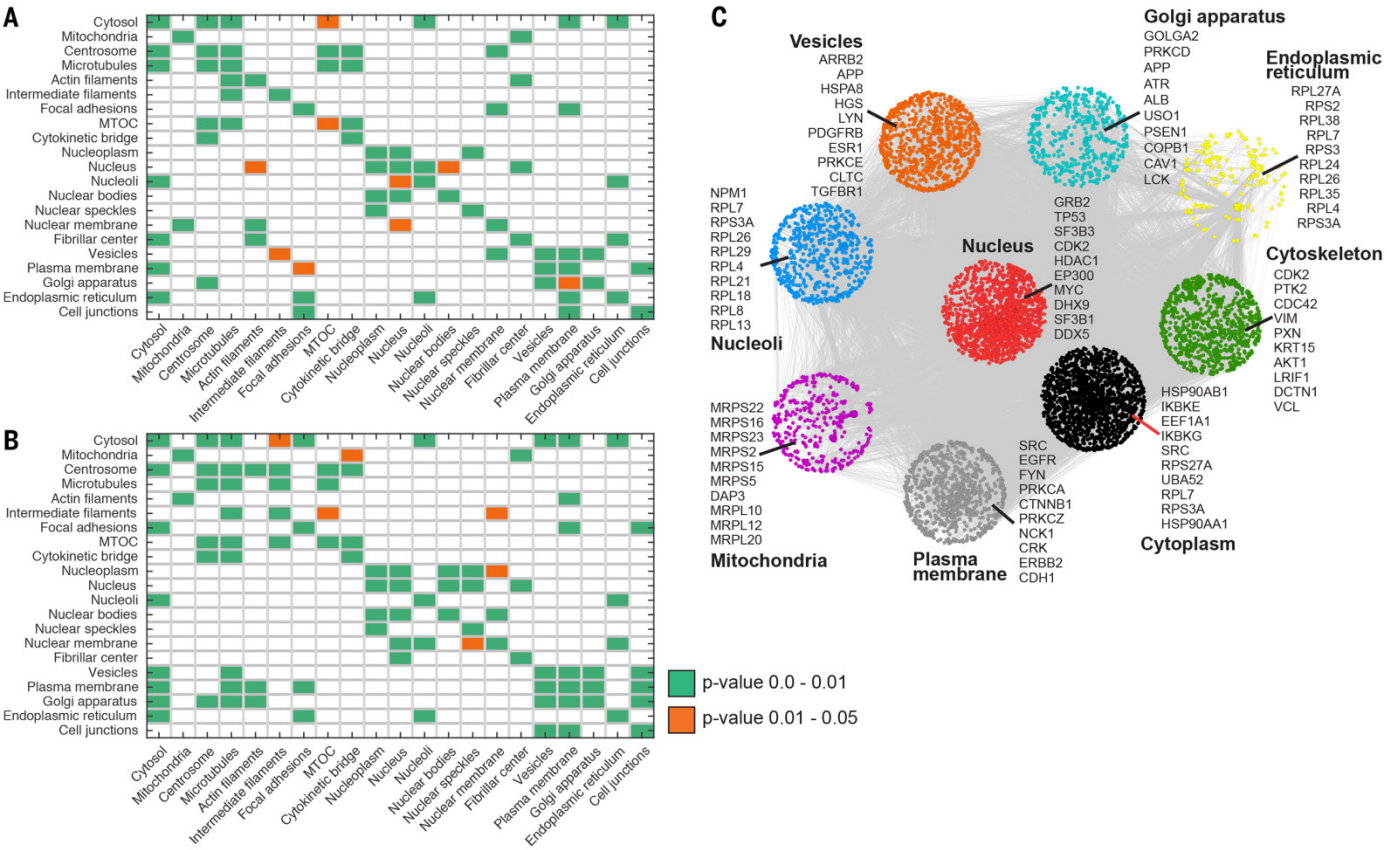


Fig. 5. Protein-protein interactions. (A and B) Information of protein-protein interaction pairs from the independent Reactome database were used to assess the quality of annotations in the Cell Atlas and identify potential interacting compartments. The Bonferroni-corrected binomial test (p-value) heatmaps describe significant overrepresentation of observing proteins in each organelle (y-axis) given the location of its interaction partner (x-axis) and the probability of randomly choosing a protein from each organelle in the Cell Atlas. (A) Analysis of direct protein-protein interactions defined by Reactome (B) Protein-protein interaction within the same reaction defined by the Reactome. (C) Human interactome pruned by the protein subcellular localization reveals hub proteins for each compartment (i.e., top-10 hub proteins, based on their degree connectivity). The full scale of the pruned interactome with nodes colored by one of their subcellular localizations is shown. Edges between same-colored nodes, i.e., same compartment, indicate protein interactions of a compartment of given color; and edges across differently colored nodes, i.e., different compartments, indicate possible linkages across all compartments because of multilocalized proteins.

Downloaded from http://science.sciencemag.org/ on May 11, 2017

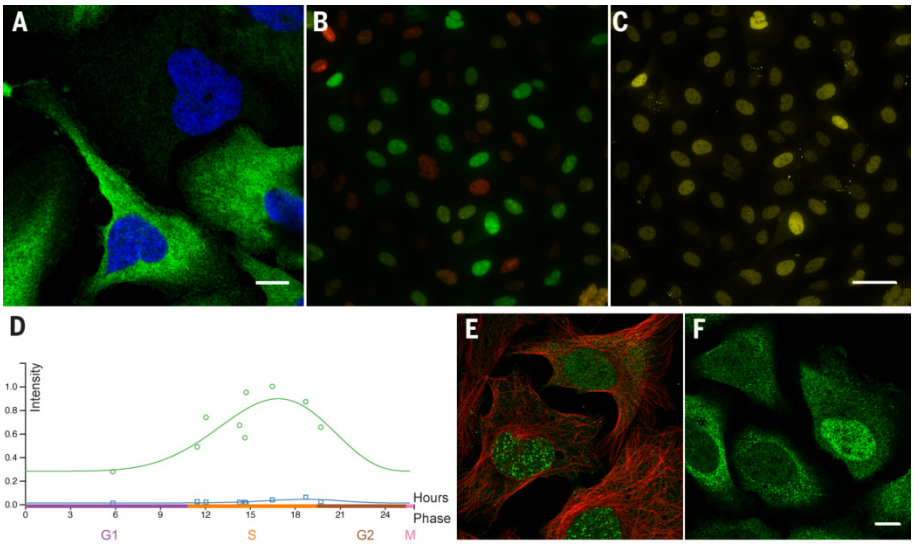


Fig. 6. Single-cell variation of protein expression. (**A**) CRYAB (detected by CAB002053 in U-2 OS cells) showed a single-cell variation in the cytosolic signal strength (**B** and **C**) U-2 OS FUCCI cells expressed the cell cycle regulators Cdt1 (red) during G1 phase and geminin (green) during S/G2 phase. Antibody targeting ANLN (yellow) stained only cells in S/G2 phase marked by the green fluorescence. (**D**) Pattern of expression of ANLN across the cell cycle in U-2 OS cells by pseudo-temporal analysis using a time regressive computational model. (**E**) The protein abundance of PCNA (detected by HPA030522 in U-2 OS cells) at nuclear bodies varied during the cell cycle. (**F**) PSMC6 (detected by HPA042823 in U-2 OS cells) changed its spatial distribution from nucleoplasm to cytosol during the cell cycle based on U-2 OS FUCCI cells. Scale bar represents 10 μm in (A) and (F), and 50 μm in (C).

Downloaded from http://science.sciencemag.org/ on May 11, 2017

A subcellular map of the human proteome

Peter J. Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M. Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M. Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M. Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P. Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S. Lilley, Mathias Uhlén and Emma Lundberg (May 11, 2017)

published online May 11, 2017

Editor's Summary

This copy is for your personal, non-commercial use only.

Article Tools Visit the online version of this article to access the personalization and article tools:
http://science.sciencemag.org/content/early/2017/05/10/science.aal3321

Permissions Obtain information about reproducing this article:
http://www.sciencemag.org/about/permissions.dtl

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2016 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

Downloaded from http://science.sciencemag.org/ on May 11, 2017