RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

OpenCell: Endogenous tagging for the cartography of human cellular organization

Nathan H. Cho†, Keith C. Cheveralls†, Andreas-David Brunner†, Kibeom Kim†, André C. Michaelis†, Preethi Raghavan†, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y. S. Kim, Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, Manuel D. Leonetti*

INTRODUCTION: Proteins are the product of gene expression and the molecular building blocks of cells. Examples include enzymes that orchestrate the cell's chemistry, filaments that shape the cell's structure, or the pharmacological targets of drugs. The genome sequence provides us with the complete set of proteins that give rise to the human cell. However, systematically characterizing how proteins organize within the cell to sustain its operation remains an important goal of modern cell biology. A comprehensive map of the human proteome's organization will serve as a reference to explore gene function in health and disease.

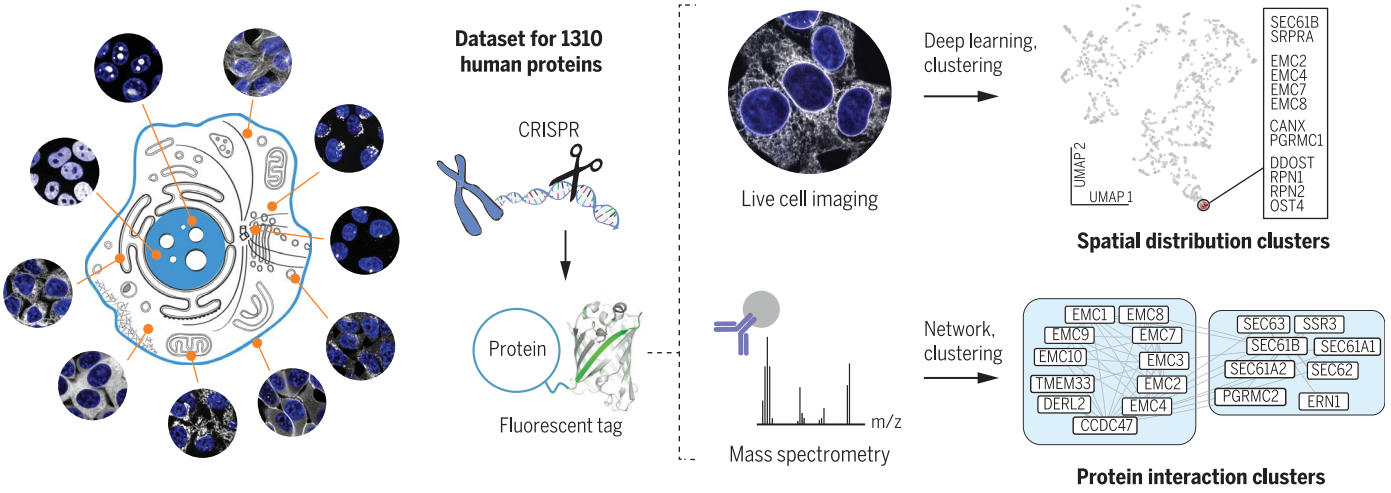
RATIONALE: Subcellular localization and physical interactions are key aspects tightly related to the function of any given protein. Proteins localize to different subcellular compartments, which enables a spatial separation of cellular functions. Proteins also physically interact with one another, forming molecular networks that connect proteins involved in the same processes. Therefore, mapping the cell's molecular organization requires a comprehensive description of where different proteins localize and how they interact. Among other strategies, a powerful approach to map cellular architecture is to visualize individual proteins using fusions with fluorescent protein "tags." These tags allow us not only to image protein localization in live cells, but also to measure protein interactions by serving as handles for immunopurification–mass spectrometry (IP-MS). Recent advances in genome engineering facilitate tagging of endogenous human genes, so that the corresponding proteins can be characterized in their native cellular environment.

RESULTS: Using high-throughput CRISPR-mediated genome editing, we constructed a library of 1310 fluorescently tagged cell lines. By performing paired IP-MS and live-cell imaging using this library, we generated a large dataset that maps the cellular localization and physical interactions of the corresponding 1310 proteins. Applying a combination of unsupervised clustering and machine learning for image analysis allowed us to objectively identify proteins that share spatial or interaction signatures. Our data provide insights into the function of individual proteins, but also enable us to derive some general principles of human cellular organization. In particular, we show that proteins that bind RNA form a separate subgroup defined by specific localization and interaction signatures. We also show that the precise spatial distribution of a given protein is very strongly correlated with its cellular function, such that fine-grained molecular insights can be derived from the analysis of imaging data. Our open-source dataset can be explored through an interactive web interface at opencell.czbiohub.org.

CONCLUSION: Our results show that endogenous tagging coupled with interactome and microscopy analysis provides new systems-level insights about the organization of the human proteome. The information contained within the subcellular distribution of each protein is highly specific and can be paired with advances in machine learning to extrapolate fine-grained functional information using microscopy alone. This opens exciting avenues for the characterization of understudied proteins, high-throughput screening, and modeling of complex cellular states during differentiation and disease. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: manuel.leonetti@czbiohub.org
†These authors contributed equally to this work.
Cite this article as N. H. Cho *et al.*, *Science* **375**, eabi6983 (2022). DOI: 10.1126/science.abi6983

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abi6983



OpenCell: Combining endogenous tagging, live-cell imaging, and interaction proteomics to map the architecture of the human proteome. We created a library of engineered cell lines by using CRISPR to introduce fluorescent tags into 1310 individual human proteins. This allowed us to image the localization of each protein in live cells, as well as the interactions between a given target and other proteins within the cell. This large dataset enables a systems-level description of the organization of the human proteome.

RESEARCH ARTICLE

CELL BIOLOGY

OpenCell: Endogenous tagging for the cartography of human cellular organization

Nathan H. Cho[1]†, Keith C. Cheveralls[1]†, Andreas-David Brunner[2]†, Kibeom Kim[1]†, André C. Michaelis[2]†, Preethi Raghavan[1]†, Hirofumi Kobayashi[1], Laura Savy[1], Jason Y. Li[1], Hera Canaj[1], James Y. S. Kim[1], Edna M. Stewart[1], Christian Gnann[1,3], Frank McCarthy[1], Joana P. Cabrera[1], Rachel M. Brunetti[4], Bryant B. Chhun[1], Greg Dingle[5], Marco Y. Hein[1], Bo Huang[1,4,6], Shalin B. Mehta[1], Jonathan S. Weissman[7,8], Rafael Gómez-Sjöberg[1], Daniel N. Itzhak[1], Loïc A. Royer[1], Matthias Mann[2,9], Manuel D. Leonetti[1]*

Elucidating the wiring diagram of the human cell is a central goal of the postgenomic era. We combined genome engineering, confocal live-cell imaging, mass spectrometry, and data science to systematically map the localization and interactions of human proteins. Our approach provides a data-driven description of the molecular and spatial networks that organize the proteome. Unsupervised clustering of these networks delineates functional communities that facilitate biological discovery. We found that remarkably precise functional information can be derived from protein localization patterns, which often contain enough information to identify molecular interactions, and that RNA binding proteins form a specific subgroup defined by unique interaction and localization properties. Paired with a fully interactive website (opencell.czbiohub.org), our work constitutes a resource for the quantitative cartography of human cellular organization.

Sequencing the human genome has transformed cell biology by defining the protein parts list that forms the canvas of cellular operation (*1*, *2*). This paves the way for elucidating how the ~20,000 proteins encoded in the genome organize in space and time to define the cell's functional architecture (*3*, *4*). Where does each protein localize within the cell? Can we comprehensively map how proteins assemble into larger functional communities? A main challenge to answering these fundamental questions is that cellular architecture is organized along multiple scales. Therefore, several approaches need to be combined for its elucidation (*5*). In a series of pioneering studies, human protein-protein interactions have been mapped using ectopic expression strategies with yeast two-hybrid (*6*) or epitope tagging coupled to immunoprecipitation–mass spectrometry (IP-MS) (*7*, *8*), whereas protein localization has been charted using immunofluorescence in fixed samples (*9*). A complementary approach is to directly modify genes in a genome by appending sequences that illuminate specific aspects of the corresponding proteins' function [commonly referred to as "endogenous tagging" (*10*)]. For example, endogenously tagging a gene with a fluorescent reporter enables imaging of protein subcellular localization in live cells and supports functional characterization in a native cellular environment (*10*, *11*). The use of endogenous tagging to study the organization of a eukaryotic cell is illustrated by seminal work in the budding yeast *Saccharomyces cerevisiae*. There, libraries of tagged strains have enabled the comprehensive mapping of protein localization and molecular interactions across the yeast proteome (*12*–*14*). These libraries were made possible by the relative simplicity of homologous recombination and genome engineering in yeast (*15*). In human cells, earlier work has leveraged alternative strategies including expression from bacterial artificial chromosomes (*16*) or central-dogma tagging (*17*) because of the difficulty of site-specific gene editing. CRISPR-mediated genome engineering now allows for homologous recombination–based endogenous tagging to be applied for the interrogation of the human cell (*10*, *11*, *18*).

Here, we combined experimental and analytical strategies to create OpenCell, a proteomic map of human cellular architecture. We generated a library of 1310 CRISPR-edited human embryonic kidney (HEK) 293T cell lines harboring fluorescent tags on individual proteins, which we characterized by pairing confocal microscopy and mass spectrometry. Our dataset constitutes the most comprehensive live-cell image collection of human protein localization to date. In addition, integration of IP-MS using the fluorescent tags for affinity capture enables measurement of localization and interactions from the same samples. For a quantitative description of cellular architecture, we developed a data-driven framework to represent protein interactions and localization features, supported by a new machine learning algorithm for image encoding.

This approach allows us to delineate communities of functionally related proteins by unsupervised clustering and facilitates the generation of mechanistic hypotheses, including for proteins that had so far remained uncharacterized. We further demonstrate that the localization pattern of each protein is defined by unique and specific features that can be used for functional interpretation, to the point that spatial relationships often contain enough information to predict interactions at the molecular scale. Finally, our analysis enables an unsupervised description of the human proteome's organization; in particular, we found that RNA binding proteins exhibit unique functional signatures that shape the proteome's network.

Engineered cell library

Fluorescent protein (FP) fusions are versatile tools that can enable the measurement of protein localization (by microscopy) as well as protein-protein interactions (by acting as affinity handles for IP-MS) (*18*, *19*) (fig. S1A). Here, we constructed a library of fluorescently tagged HEK293T cell lines by targeting human genes with the split-mNeonGreen2 system (*20*) (Fig. 1A). Split FPs greatly simplify CRISPR-based genome engineering by circumventing the need for molecular cloning (*18*) and allowed us to generate endogenous genomic fusions (Fig. 1B) that preserve native expression regulation. A full description of our pipeline is available in (*21*) and is summarized in Fig. 1, C to E. In brief, FP insertion sites (N or C terminus) were chosen on the basis of information from the literature or structural analysis (fig. S1B and table S1). For each tagged target, we isolated a polyclonal pool of CRISPR-edited cells, which was then characterized by live-cell three-dimensional (3D) confocal microscopy, IP-MS, and genotyping of tagged alleles by next-generation sequencing. Open-source software development and advances in instrumentation supported scalability (Fig. 1C). In particular, we developed crispycrunch, a CRISPR design software that enables guide RNA selection and homology donor sequence design (github.com/czbiohub/crispycrunch). We also fully automated the acquisition of microscopy data in Python for

[1]Chan Zuckerberg Biohub, San Francisco, CA, USA. [2]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [3]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH–Royal Institute of Technology, Stockholm, Sweden. [4]Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA. [5]Chan Zuckerberg Initiative, Redwood City, CA, USA. [6]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA. [7]Whitehead Institute, Koch Institute, Howard Hughes Medical Institute, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, CA, USA. [9]NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
***Corresponding author. Email: manuel.leonetti@czbiohub.org**
†These authors contributed equally to this work.

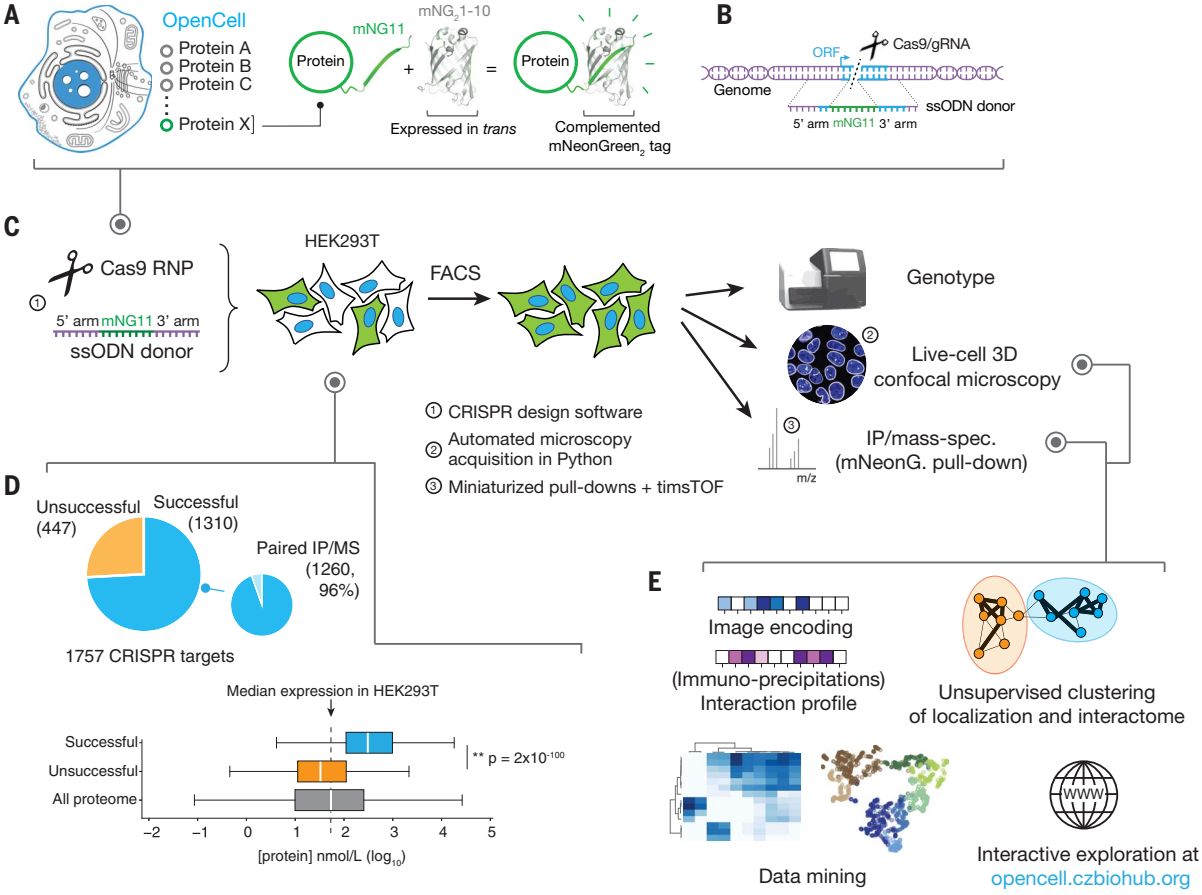


Fig. 1. The OpenCell library. (**A**) Functional tagging with split-mNeonGreen$_2$. In this system, mNeonGreen$_2$ is separated into two fragments: a short mNG11 fragment, which is fused to a protein of interest, and a large mNG$_2$1-10 fragment, which is expressed separately in trans (that is, tagging is done in cells that have been engineered to constitutively express mNG$_2$1-10). (**B**) Endogenous tagging strategy: mNG11 fusion sequences are inserted directly within genomic open reading frames (ORFs) using CRISPR-Cas9 gene editing and homologous recombination with single-stranded oligodeoxynucleotide donors (ssODNs). (**C**) The OpenCell experimental pipeline. See text for details. (**D**) Successful detection of fluorescence in the OpenCell library. Top: Of 1757 genes that were originally targeted, fluorescent signal was successfully detected for 1310. Bottom: Low protein abundance is the main obstacle to successful detection. The graph shows the distribution of abundance for all proteins expressed in HEK293T versus successfully or unsuccessfully detected OpenCell targets; boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5 times the interquartile range. Median is indicated by a white line. *P* value: Student's *t* test. (**E**) The OpenCell data analysis pipeline.

on-the-fly computer vision and selection of desirable fields of view imaged in 96-well plates (github.com/czbiohub/2021-opencell-microscopy-automation). Our mass spectrometry protocols used the high sensitivity of trapped ion mobility spectrometry time-of-flight (timsTOF) instruments (*22*), which allowed miniaturization of IP-MS down to 0.8×10^6 cells of starting material [fig. S1C; about one-tenth of the material required in previous approaches (*7*, *8*)].

In total, we targeted 1757 genes, of which 1310 (75%) could be detected by fluorescence imaging and form our current dataset (full library details in table S1). From these, we obtained paired IP-MS measurements for 1260 targets (96%; Fig. 1D). The 1310-protein collection includes a balanced representation of the pathways, compartments, and functions of the human proteome (fig. S1D), with the exception of processes specific to mitochondria, organellar lumen, or extracellular matrix. Indeed, the split-FP system tags a gene of interest with a short sequence (mNG11) while a larger FP fragment (mNG$_2$1-10) is expressed separately (Fig. 1A). In the version used here, the mNG$_2$1-10 fragment is expressed in the nucleocytoplasm and prevents access to proteins inside organellar compartments. Membrane proteins can be tagged as long as one terminus extends into the nucleocytoplasm. In future iterations, other split systems that contain compartment-specific signal sequences could be used to target organellar lumen (*23*).

Fluorescent tagging was readily successful for essential genes, which suggests that FP fusions are well tolerated (fig. S2A). To evaluate other factors contributing to successful fluorescent detection, we measured RNA and protein concentration in HEK293T cells (fig. S2B) using a 24-fraction scheme for deep proteome quantification (see fully annotated proteome in table S2). This revealed that protein abundance is the main limitation to detection (Fig. 1D and fig. S2C; see details for unsuccessful targets in table S3); most successful targets are among the top 50% most abundant (fig. S2D). Gene-editing efficiency was another important factor: Among well-expressed targets, failure was correlated with significantly lower rates of homologous recombination (fig. S2E), which would impair the selection of edited cells by fluorescence-activated cell sorting (FACS). Training a regression model revealed that the combination of protein abundance and editing efficiency could predict successful detection with 82% accuracy.

To maximize throughput, we used a polyclonal strategy to select genome-edited cells by FACS. Polyclonal pools contain cells with distinct

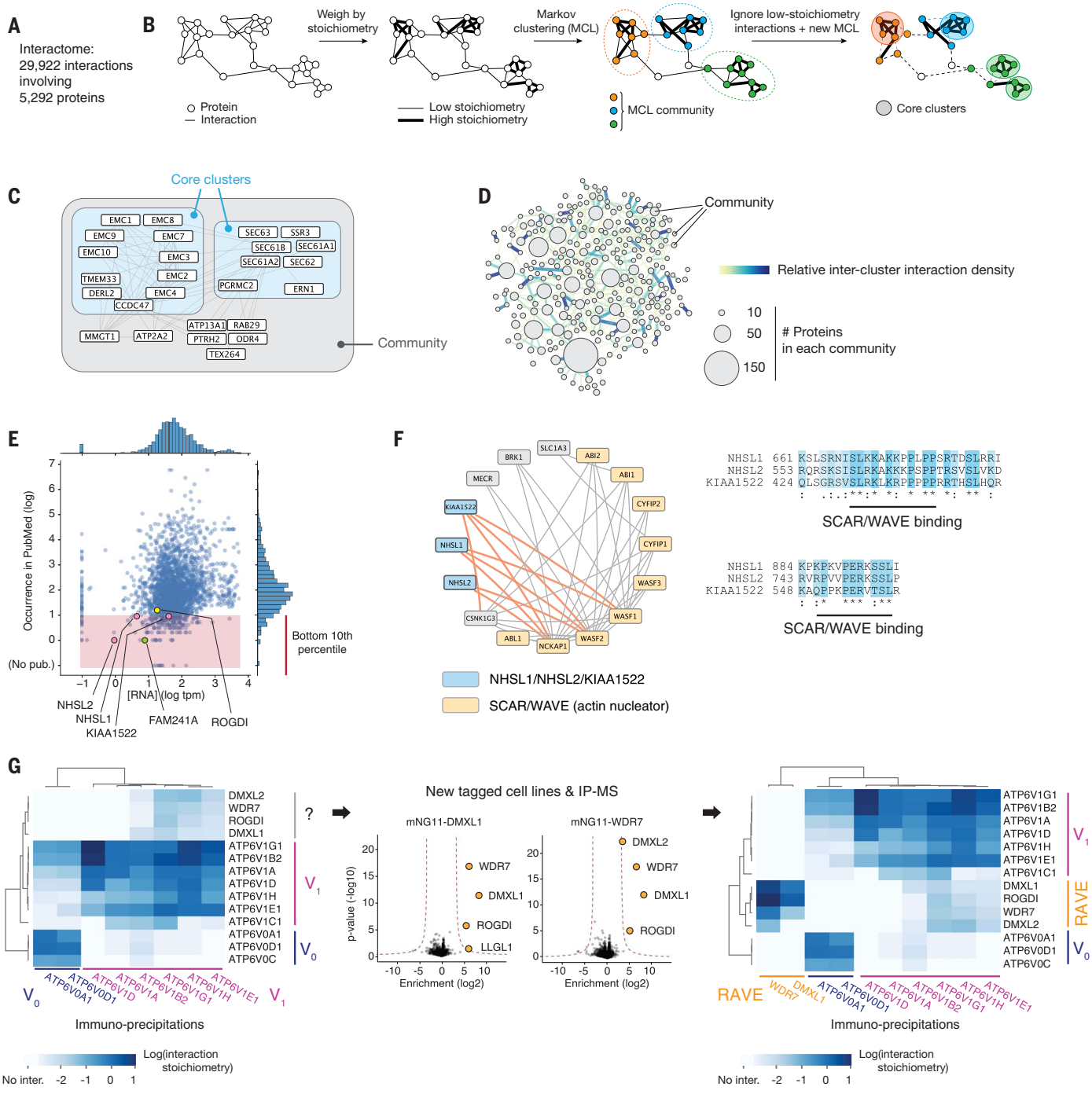


Fig. 2. Protein interactome. (**A**) Overall description of the interactome. (**B**) Unsupervised Markov clustering of the interactome graph. (**C**) Example of community and core cluster definition for the translocon/EMC community. (**D**) The complete graph of connections between interactome communities. The density of protein-protein interactions between communities is represented by increased edge width. The numbers of targets included in each community are represented by circles of increasing diameter. (**E**) Distribution of occurrence in PubMed articles versus RNA expression for all proteins found within interactome communities. The bottom 10th percentile of publication count (poorly characterized proteins) is highlighted. (**F**) NHSL1, NHSL2, and KIAA1522 are part of the SCAR/WAVE community and share amino acid sequence homology. (**G**) DMXL1/2, WDR7, and ROGDI form the human RAVE complex. Heatmaps represent the interaction stoichiometry of preys (lines) in the pull-downs of specific OpenCell targets (columns). See text for details.

genotypes. HEK293T cells are pseudo-triploid (*24*) and a single edited allele is sufficient to confer fluorescence. Moreover, various DNA repair mechanisms compete with homologous recombination for the resolution of CRISPR-induced genomic breaks (*25*) so that alleles containing nonfunctional mutations can be present in addition to the desired fusion alleles. However, such alleles do not support fluorescence and are therefore unlikely to have an impact on other measurements, especially in the context of a polyclonal pool. We developed a stringent selection scheme to significantly enrich for fluorescent fusion alleles (fig. S3A). Our final cell library has a median

61% of mNeonGreen-integrated alleles, 5% wild-type, and 26% other nonfunctional alleles (fig. S3B; full genotype information in table S1).

Finally, we verified that our engineering approach maintained the endogenous abundance of the tagged target proteins. For this, we quantified protein expression by Western blotting using antibodies specific to proteins targeted in 12 different cell pools (fig. S3C) and by single-shot mass spectrometry in 63 tagged lines (fig. S3D). Both approaches revealed a median abundance of tagged targets in engineered lines at ~80% of untagged HEK293T control, with five outliers (8% of total) identified by proteomics (fig. S3D, all within a factor of 3.5 of control). The overall proteome composition was unchanged in all tagged lines (fig. S3, E and F).

Overall, our gene-editing strategy preserves near-endogenous abundances and circumvents the limitations of ectopic overexpression (*11*, *26*, *27*), which include aberrant localization, changes in organellar morphology, and masking effects (see the examples of SPTLC1, TOMM20, and MAP1LC3B in fig. S3G). Therefore, OpenCell supports the functional profiling of tagged proteins in their native cellular context.

Interactome analysis and stoichiometry-driven clustering

Affinity enrichment coupled to mass spectrometry is an efficient and sensitive method for the systematic mapping of protein interaction networks (*28*). We isolated tagged proteins ("baits") from cell lysates solubilized in digitonin, a mild nonionic detergent that preserves the native structure and properties of membrane proteins (*29*). Specific protein interactors ("preys") were identified by proteomics from biological triplicate experiments [see fig. S4, A and B, and (*21*) for a detailed description of our statistical analysis, which builds upon established methods (*7*)]. In total, the full interactome from our 1260 OpenCell baits includes 29,922 interactions among a total of 5292 proteins (baits and preys, Fig. 2A; full interactome data in table S4).

To assess the quality of our interactome, we estimated its precision (the fraction of true positive interactions over all interactions) and recall (the fraction of interactions identified relative to a ground truth set) using reference data (fig. S4B). For recall analysis, we quantified the coverage in our data of interactions included in CORUM (*30*), a compendium of protein interactions manually curated from the literature. To estimate precision, we quantified how many of our interactions involved protein pairs expected to localize to the same broad cellular compartment (*31*) (fig. S4B). To benchmark OpenCell against other large-scale interactomes, we compared its precision and recall to Bioplex [overexpression of hemagglutinin-tagged baits (*8*, *32*)], the yeast two-hybrid human reference interactome [HuRI (*6*)], and our own previous data [green fluorescent protein fusions expressed from bacterial artificial chromosomes (*7*)] (fig. S4, C to E). We also calculated compression rates for each dataset as a measure of the overall richness in network patterns and motifs distinguishable from noise, which correlates with overall network quality: Real-world networks contain redundant information that can be compressed, whereas pure noise is not compressible (*33*) (fig. S4F). Across all metrics, OpenCell outperformed previous approaches. OpenCell also includes many interactions not reported in previous datasets (fig. S4, E and G). Our interactome may better reflect biological interactions because it preserves near-endogenous protein expression.

A powerful way to interpret interactomes is to identify communities of interactors (*8*, *13*). To this end, we applied unsupervised Markov clustering (MCL) (*34*) to the graph of interactions defined by our data (5292 proteins total, baits and preys). We first measured the stoichiometry of each interaction, using a quantitative approach we previously established (*7*). Interaction stoichiometry measures the abundance of a protein interactor relative to the abundance of the bait in a given immunoprecipitation sample. We have shown that stoichiometry can be interpreted as a proxy for interaction strength and that interactions can be classified between core (i.e., high) and low stoichiometries (*7*). In our current data, both high- and low-stoichiometry interactions were significantly enriched for protein pairs sharing Gene Ontology (GO) annotations (fig. S4H). Using stoichiometry to assign weights to the edges in the interaction graph (Fig. 2B), a first round of MCL delineated interconnected protein communities and led to better clustering performance than clustering based on connectivity alone (fig. S4I). To better delineate stable complexes, we further refined each individual MCL community by additional clustering while removing low-stoichiometry interactions. The resulting subclusters outline core interactions within existing communities (Fig. 2B). Figure 2C illustrates how this unsupervised approach enables delineation of functionally related proteins: All subunits of the machinery responsible for the translocation of newly translated proteins at the endoplasmic reticulum (ER) membrane (SEC61/62/63) and of the ER membrane complex (EMC) are grouped within respective core interaction clusters, but both are part of the same larger MCL community. This mirrors the recently appreciated cotranslational role of EMC for insertion of transmembrane domains at the ER (*35*). Additional proteins that have only recently been shown to act cotranslationally are found clustering with translocon or EMC subunits, including ERN1 (IRE1) (*36*) and CCDC47 (*37*, *38*). Thus, clustering can facilitate mechanistic exploration by grouping proteins involved in related pathways. Overall, we identified 300 communities including a total of 2096 baits and preys (full details in table S4). Ontology analysis revealed that these communities are significantly enriched for specific cellular functions, supporting their biological relevance (82% of all communities are significantly enriched for specific biological process or molecular function GO terms; see table S5 for complete analysis). A graph of interactions between communities reveals a richly interconnected network (Fig. 2D), the structure of which outlines the global architecture of the human interactome (discussed further below).

Interactome clustering can be directly applied to help elucidate the cellular roles of the many human proteins that remain poorly characterized (*39*). We identified poorly characterized proteins by quantifying their occurrence in article titles and abstracts from PubMed (Fig. 2E). Empirically, we determined that proteins in the bottom 10th percentile of publication count (corresponding to fewer than 10 publications) are very poorly annotated (Fig. 2E). This set encompasses a total of 251 proteins found in interaction communities for which our dataset offers potential mechanistic insights. For example, the proteins NHSL1, NHSL2, and KIAA1522 are all found as part of a community centered around SCAR/WAVE, a large multisubunit complex nucleating actin polymerization (Fig. 2F). All three proteins share sequence homology and are homologous to NHS (fig. S5A), a protein mutated in patients with Nance-Horan syndrome. NHS interacts with SCAR/WAVE components to coordinate actin remodeling (*40*). Thus, NHSL1, NHSL2, and KIAA1522 might also act to regulate actin assembly. A recent mechanistic study supports this hypothesis: NHSL1 localizes at the cell's leading edge and directly binds SCAR/WAVE to negatively regulate its activity, reducing F-actin content in lamellipodia and inhibiting cell migration (*41*). The authors identified NHSL1's SCAR/WAVE binding sites, and we found these sequences to be conserved in NSHL2 and KIA1522 (Fig. 2F). Therefore, our data suggest that both NHSL2 and KIAA1522 are also direct SCAR/WAVE binders and possible modulators of the actin cytoskeleton.

Our data also shed light on the function of ROGDI, whose variants cause Kohlschuetter-Toenz syndrome [a recessive developmental disease characterized by epilepsy and psychomotor regression (*42*)]. ROGDI appears in the literature because of its association with disease, but no study, to our knowledge, has specifically determined its molecular function. We first observed that ROGDI's interaction pattern closely matched that of three other proteins in our dataset: DMXL1, DMXL2, and WDR7 (Fig. 2G). This set exhibited a specific

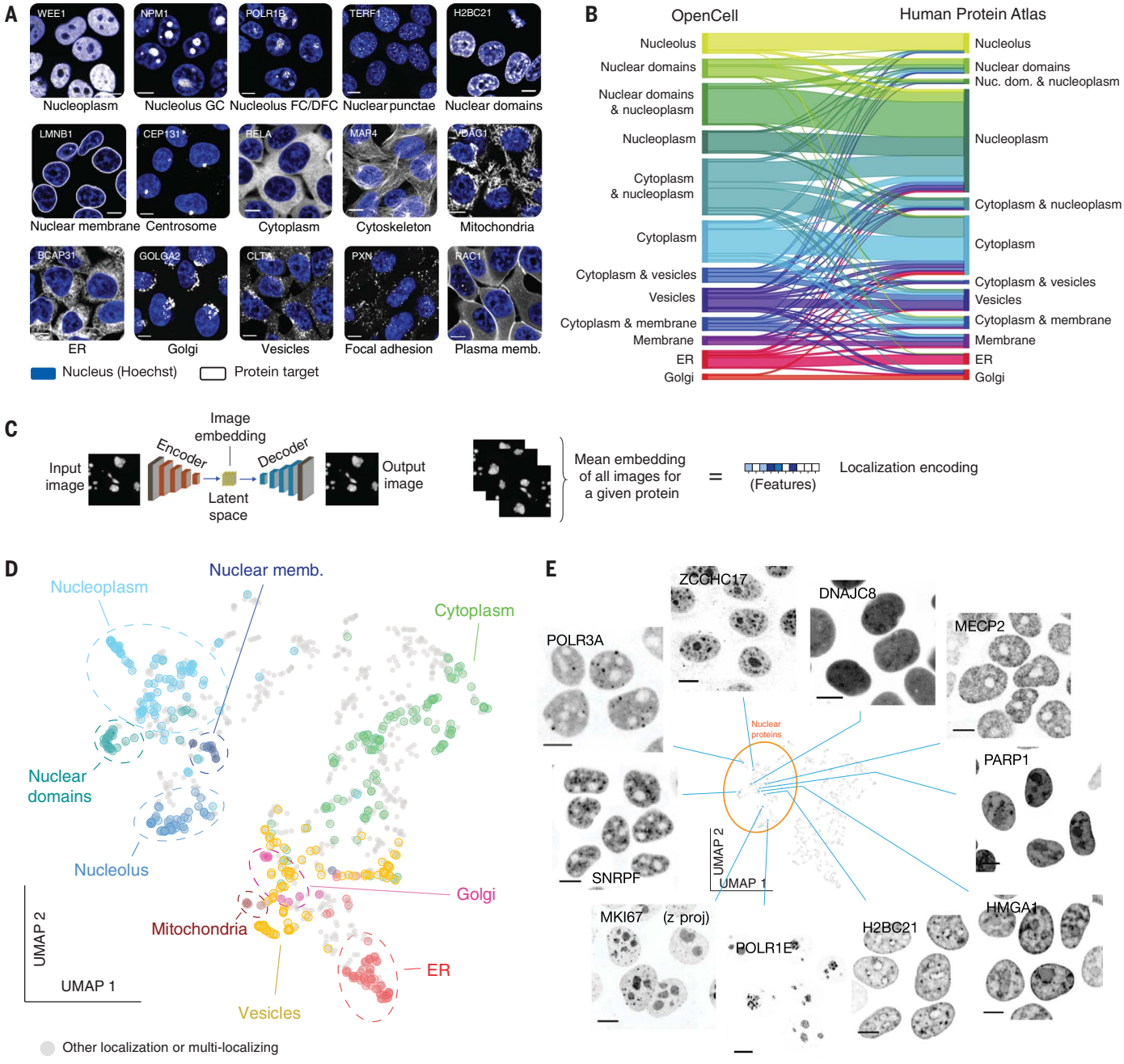


Fig. 3. Live-cell image collection. (**A**) The 15 cellular compartments segregated for annotation of localization patterns. The localization of a representative protein belonging to each group is shown (grayscale, gene names in top left corners; scale bar, 10 μm). Nuclear stain (Hoechst) is shown in blue. "Nuclear domains" designate proteins with pronounced nonuniform nucleoplasmic localization, such as chromatin-binding proteins. (**B**) Comparison of annotated localizations for proteins included in both OpenCell and Human Protein Atlas datasets. In this flow diagram, colored bands represent groups of proteins that share the same localization annotation in OpenCell, and the width of the band represents the number of proteins in each group. For readability, only the 12 most common localization groups are shown. Some multilocalization groups are included (e.g., "cytoplasm & nucleoplasm"). (**C**) Principle of localization encoding by self-supervised machine learning. See text for details. (**D**) UMAP representation of the OpenCell localization dataset, highlighting targets found to localize to a unique cellular compartment. (**E**) Representative images for 10 nuclear targets that exemplify the nuanced diversity of localization patterns across the proteome. Scale bars, 10 μm.

interaction signature with the v-ATPase lysosomal proton pump. All four proteins interact with soluble v-ATPase subunits (ATP6-V1), but not its intramembrane machinery (ATP6-V0). DMXL1 and WDR7 interact with V1 v-ATPase, and their depletion in cells compromises lysosomal reacidification (*43*). Sequence analysis showed that DMXL1 or 2, WDR7, and ROGDI are homologous to proteins from yeast and *Drosophila* involved in the regulation of assembly of the soluble V1 subunits onto the V0 transmembrane ATPase core (*44*, *45*) (fig. S5B). In yeast, Rav1 and Rav2 (homologous to DMXL1/2 and ROGDI, respectively) form the stoichiometric RAVE complex, a soluble chaperone that regulates

v-ATPase assembly (45). To assess the existence of a human RAVE-like complex, we generated new tagged cell lines for DMXL1 and 2, WDR7, and ROGDI. Because of the low abundance of these proteins, the localization of DMXL2 and ROGDI was not detectable, but pull-downs of DMXL1 and WDR7 confirmed a stoichiometric interaction between DMXL1 and 2, WDR7, and ROGDI (Fig. 2G, right panels). No direct interaction between DXML1 and DMXL2 was detected, suggesting that they might nucleate two separate subcomplexes. Therefore, our data reveal a human RAVE-like complex comprising DMXL1 or 2, WDR7, and ROGDI, which we propose acts as a chaperone for v-ATPase assembly based on its yeast homolog. Altogether, these results illustrate how our data can facilitate the generation of new mechanistic hypotheses by combining quantitative analysis and literature curation.

Image dataset: Localization annotation and self-supervised machine learning

A key advantage of our cell engineering approach is to enable the characterization of each tagged protein in live, unperturbed cells. To profile localization, we performed spinning-disk confocal fluorescence microscopy (63×, 1.47 NA objective) under environmental control (37°C, 5% CO_2) and imaged the 3D distribution of proteins in consecutive z-slices. Microscopy acquisition was fully automated in Python to enable scalability (fig. S6, A and B). In particular, we trained a computer vision model to identify fields of view (FOVs) with homogeneous cell density on-the-fly, which reduced experimental variation between images. Our dataset contains a collection of 6375 3D stacks (five different FOVs for each target) and includes paired imaging of nuclei with live-cell Hoechst 33342 staining.

We manually annotated localization patterns by assigning each protein to one or more of 15 separate cellular compartments such as the nucleolus, centrosome, or Golgi apparatus (Fig. 3A). Because proteins often populate multiple compartments at steady state (9), we graded annotations using a three-tier system: Grade 3 identifies prominent localization compartment(s), grade 2 represents less pronounced localizations, and grade 1 annotates weak localization patterns nearing our limit of detection (see fig. S7A for two representative examples; full annotations in table S6). Ignoring grade 1 annotations, which are inherently less precise, 55% of proteins in our library were detected in multiple locations consistent with known functional relationships. For example, clear connections were observed between secretory compartments (ER, Golgi, vesicles, plasma membrane), or between cytoskeleton and plasma membrane (fig. S7B and table S6). Many proteins were found in both nucleus and cytoplasm (21% of our library), highlighting the importance of the nucleocytoplasmic import and export machinery in shaping global cellular function (46, 47). Because our split-FP system does not enable the detection of proteins in the lumen of organelles, multilocalization involving translocation across an organellar membrane (which is rare but does happen for mitochondrial or peroxisomal proteins) cannot be detected in our data.

To benchmark our dataset, we compared our localization annotations against the Human Protein Atlas (HPA), the reference antibody-based compendium of human protein localization (9). This revealed significant agreement between datasets: 75% of proteins shared at least one localization annotation in common (Fig. 3B; this includes 25% of all proteins that shared the exact same set of annotations, see full description in table S7A). Because HPA mostly reports on cell lines other than HEK293T, a perfect overlap was not expected, as proteins might differentially localize between related compartments in different cell types. However, the annotations for 147 proteins (11% of our data) were fully inconsistent between the two datasets (fig. S7C). An extensive curation of the literature on the localization of those proteins allowed us to resolve discrepancies for 115 proteins (i.e., 78% of that set; full curation in table S8). Of these, existing literature evidence supported the OpenCell results for 113 (98.3%) of the 115 cases (fig. S7D). This confirms the usefulness of endogenous tagging as an aid to refining the curation of localization in the human proteome. Finally, our dataset included 350 targets that have orthologs in *S. cerevisiae*. Comparison between OpenCell and yeast localization annotations (48) revealed a high degree of concordance (fig. S7E and table S7B; 81% of proteins share at least one annotation in common, including 36% perfect matches).

Although expert annotation remains the best-performing strategy to curate protein localization (49, 50), the low-dimensional description it allows is not well suited for quantitative comparisons. Recent developments in image analysis and machine learning offer new opportunities to extract high-dimensional features from microscopy images (50, 51). Therefore, we developed a deep learning model to quantitatively represent the localization pattern of each protein in our dataset (52). Briefly, our model is a variant of an autoencoder (Fig. 3C): a form of neural network that learns to vectorize an image through paired tasks of encoding (from an input image to a vector in a latent space) and decoding (from the latent space vector to a new output image). After training, a consensus representation for a given protein can be obtained from the average of the encodings from all its associated images. This generates a high-dimensional "localization encoding" (Fig. 3C) that captures the complex set of features that define the spatial distribution of a protein at steady state and across many individual cells. One of the main advantages of this approach is that it is self-supervised. Therefore, as opposed to supervised machine learning strategies that are trained to recognize pre-annotated patterns [for example, manual annotations of protein localization (50)], our method extracts localization signatures from raw images without any a priori assumptions or manually assigned labels. To visualize the relationships between these high-dimensional encodings, we embedded the encodings for all 1310 OpenCell targets in two dimensions using UMAP, an algorithm that reduces high-dimensional datasets to two dimensions (UMAP 1 and UMAP 2) while attempting to preserve the global and local structures of the original data (53). The resulting map is organized in distinct territories that closely match manual annotations (Fig. 3D, highlighting monolocalizing proteins). This shows that the encoding approach yields a quantitative representation of the biologically relevant information in our microscopy data. The separation of different protein clusters in the UMAP embedding (discussed further below) mirrors the fascinating diversity of localization patterns across the full proteome. Images from nuclear proteins offer compelling illustrative examples of this diversity and reveal how fine-scale details can define the localization of proteins within the same organelle (Fig. 3E).

Functional specificity of protein localization in the human cell

Extracting functional insights directly from cellular images is a major goal of modern cell biology and data science (54). In this context, our image library and associated machine learning encodings enable us to explore what degree of functional relationship can be inferred between proteins solely based on their localization. For this, we first used an unsupervised Leiden clustering strategy commonly used to identify cell types in single-cell RNA sequencing datasets (55). Clusters group proteins that share similar localization properties (every protein in the dataset is included in a cluster); these groups can then be analyzed for how well they match different sets of ground-truth annotations (Fig. 4A). The average size of clusters is controlled by varying a hyperparameter called resolution (fig. S8A). Systematically varying clustering resolution in our dataset revealed that not only did low-resolution clusters delineate proteins belonging to the same organelles (Fig. 4, A and B), clustering at higher resolution also enabled us to delineate functional pathways and even molecular complexes of interacting proteins (Fig. 4, A to C). This demonstrates that the spatial distribution of each protein in the cell is highly specific, to the point that proteins sharing closely related

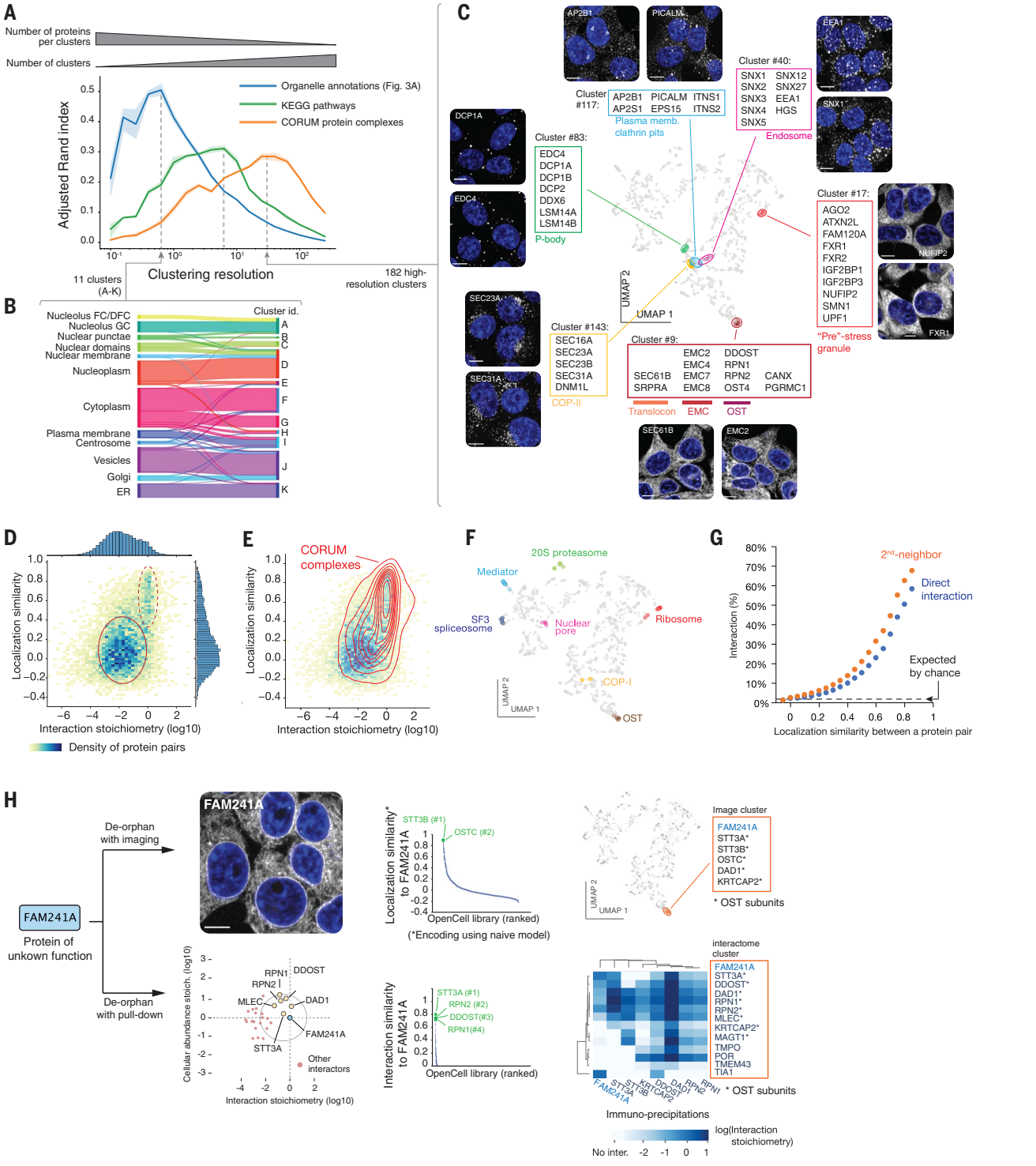


Fig. 4. Protein functional features derived from unsupervised image analysis. (**A**) Comparison of image-based Leiden clusters with ground-truth annotations. The adjusted Rand index [ARI (*88*)] of clusters relative to three ground-truth datasets is plotted as a function of the Leiden clustering resolution. ARI [a metric between 0 and 1 (*21*)] measures how well the groups from a given partition (in our case, the groups of proteins delineated at different clustering resolutions) match groups defined in a reference set. The amplitude of the ARI curves is approximately equal to the number of pairs of elements that partition similarly between sets; the resolution at which each curve reaches its maximum corresponds to the resolution that best captures the information in each ground-truth dataset. At a low resolution, Leiden clustering delineates groups that recapitulate about half of the organellar localization annotations; at increasing resolutions, clustering recapitulates about one-third of pathways annotated in KEGG, or molecular protein complexes annotated in CORUM. Shaded regions show standard deviations calculated from nine separate repeat rounds of clustering, and average values are shown as solid lines. (**B**) High correspondence between low-resolution image clusters and cellular organelles. (**C**) Examples of functional groups delineated by high-resolution image clusters, highlighted on the localization UMAP. (**D**) Heatmap distribution of localization similarity (defined as the Pearson correlation between two deep learning–derived

encoding vectors) versus interaction stoichiometry between all interacting pairs of OpenCell targets. Two discrete subgroups are outlined: low stoichiometry/low localization similarity pairs (solid oval) and high stoichiometry/high localization similarity pairs (dashed oval). In this representation, the distributions of values across *x* and *y* axes have been binned, and the density of protein pairs within each bin is color-coded. (**E**) Probability density distribution of CORUM interactions mapped on the graph from (D). Contours correspond to isoproportions of density thresholds for each 10th percentile. (**F**) Localization patterns of different subunits from example stable protein complexes, represented on the localization UMAP. (**G**) Frequency of direct (first neighbor) or once-removed (second neighbor, having a direct interactor in common) protein-protein interactions between any two pairs of OpenCell targets sharing localization similarities above a given threshold (*x* axis). (**H**) Parallel identification of FAM241A as a new OST subunit by imaging or mass spectrometry. See text for details.

functions can be identified on the sole basis of the similarity between their spatial distributions. This is further illustrated by how finely high-resolution clusters encapsulate proteins specialized in defined cellular functions (Fig. 4C). For example, our analysis not only separated P-body proteins (cluster #83) from other forms of punctate cytoplasmic structures, but also unambiguously differentiated vesicular trafficking pathways despite their very similar localization patterns: The endosomal machinery (#40), plasma membrane endocytic pits (#117), and COP-II vesicles (#143) were all delineated with high precision (Fig. 4C). Among ER proteins, the translocon formed clusters with the SRP receptor, EMC subunits, and the OST glycosylation complex, all responsible for cotranslational operations (#9). This performance extended to cytoplasmic (fig. S8A) and nuclear clusters (fig. S8B), revealing that spatial patterning is not limited to membrane-bound organelles and that subcompartments also exist in the nucleocytoplasm. An illustrative example is a cytoplasmic cluster (#17) formed by a group of RNA binding proteins (including ATXN2L, NUFIP2, and FXR1; Fig. 4C) that separate into granules upon stress conditions (*56*–*59*). Stress granules are not formed under the standard growth conditions used in our experiments, but the ability of our analysis to cluster these proteins together reveals an underlying specificity to their cytoplasmic localization (i.e., "texture") even in the absence of stress.

A direct comparison between imaging and interactome data allowed us to further examine the extent to which molecular-level relationships (that is, protein interactions) can be derived from a comparison of localization patterns. For OpenCell targets that directly interact, we compared the correlation between their localization encodings derived from machine learning (defining a "localization similarity") and the stoichiometry of their interaction. This "localization similarity" measures the similarity between the global steady-state distributions of two proteins, as opposed to a direct measure of colocalization. We found that most proteins interact with low stoichiometry [as we previously described (*7*)] and without strong similarities in their spatial distribution (Fig. 4D, solid oval). This means that although low-stoichiometry interactors colocalize at least partially to interact, their global distribution within the cell is different at steady state. On the other hand, high-stoichiometry interactors share very similar localization signatures (Fig. 4D, dashed oval). Indeed, proteins interacting within stable complexes annotated in CORUM fall into this category (Fig. 4E), and the localization signatures of different subunits from large complexes are positioned very closely in UMAP embedding (Fig. 4F). In an important correlate, we found that a high similarity of spatial distribution is a strong predictor of molecular interaction. Across the entire set of target pairs (predicted to interact or not), proteins that share high localization similarities are also very likely to interact (Fig. 4G). For example, target pairs with a localization similarity greater than 0.85 have a 58% chance of being direct interactors and a 68% chance of being second neighbors (i.e., sharing a direct interactor in common). This suggests that protein-protein interactions could be identified from a quantitative comparison of spatial distribution alone.

To test this, we focused on FAM241A (C4orf32), a protein of unknown function that was not part of our original library, and asked whether we could predict its interactions using imaging data alone, rather than the classical deorphaning approach that uses interaction proteomics. We thus generated a FAM241A endogenous fusion that was analyzed with live imaging and IP-MS separately. Encoding its localization pattern, using a "naïve" machine learning model that was never trained with images of this new target, revealed a very high localization similarity with two subunits of the ER oligosaccharyl transferase OST (>0.85 similarity to STT3B and OSTC), and high-resolution Leiden clustering placed FAM241A in an image cluster containing only OST subunits (Fig. 4H, top). This analysis suggested that FAM241A is a high-stoichiometry interactor of OST. IP-MS revealed that FAM241A was indeed a stoichiometric subunit of the OST complex (Fig. 4H, bottom). Although the specific function of FAM241A in protein glycosylation remains to be fully elucidated, this proof-of-concept example establishes that live-cell imaging can be used as a specific readout to predict molecular interactions.

Collectively, our analyses establish that the spatial distribution of a given protein contains highly specific information from which precise functional attributes can be extracted by modern machine learning algorithms. In addition, we show that whereas high-stoichiometry interactors share very similar localization patterns, most proteins interact with low stoichiometry and share different localization signatures. This reinforces the importance of low-stoichiometry interactions for defining the overall structure of the cellular network, not only providing the "glue" that holds the interactome network together (*7*) but also connecting different cellular compartments.

RNA binding proteins form a unique group in both interactome and spatial networks

To gain insight into global signatures that organize the proteome, we further examined the structures of our imaging and interactome datasets. First, we reduced the dimensionality of each dataset by grouping proteins into their respective spatial clusters (as defined by the high-resolution localization-based clusters in Fig. 4, A and C) or interaction communities (as defined in Fig. 2B). We then separately clustered these spatial groups (fig. S9A) and interaction communities (fig. S9B) to formalize paired hierarchical descriptions of the human proteome organization. These hierarchies are highly structured and delineate clear groups of proteins (see comparison to hierarchies expected by chance, fig. S9C). In both hierarchies, groups isolated at an intermediate hierarchical layer outline "modules" that are enriched for specific cellular functions or compartments (fig. S9, A and B; full ontology analysis in tables S5 and S9). At a higher layer, each dataset is partitioned into three "branches," which represent core signatures that shape the proteome's architecture from a molecular or spatial perspective (fig. S9, A and B). The structure of the localization-based hierarchy (fig. S9A) recapitulates the human cell's architecture across its three key compartments (nucleus, cytoplasm, membrane-bound organelles; fig. S10, A and B), which reinforces the relevance of our unsupervised hierarchical analysis. This motivated a deeper examination of the hierarchical architecture of the interactome (fig. S9B; ontology analysis in table S5). We found that intermediate-layer modules of the interactome delineate specific cellular functions such as transcription or vesicular transport (fig. S9B), reflecting, as expected, that functional pathways are formed by groups of proteins that physically interact (*60*, *61*). More

strikingly, the highest-layer structure showed that two of the three interactome branches were defined by clear functional signatures (fig. S10, C to E): Branch B is significantly enriched in proteins that reside in or interact with lipid membranes, whereas branch C is significantly enriched in RNA binding proteins (RNA-BPs) (Fig. 5B). This indicates that both membrane-related proteins and RNA-BPs interact more preferentially with each other than with other kinds of proteins in the cell.

That membrane-related proteins form a specific interaction group is perhaps not surprising, as the membrane surfaces that sequester them within the 3D cell will be partially maintained upon detergent solubilization. On the other hand, the fact that RNA-BPs also form a specific interaction group is unexpected, because our protein interactions were measured in nuclease-treated samples (*21*) in which most RNAs are degraded. This suggests that protein features beyond binding to RNAs themselves might drive the preferential interactions of RNA-BPs with each other. Therefore,

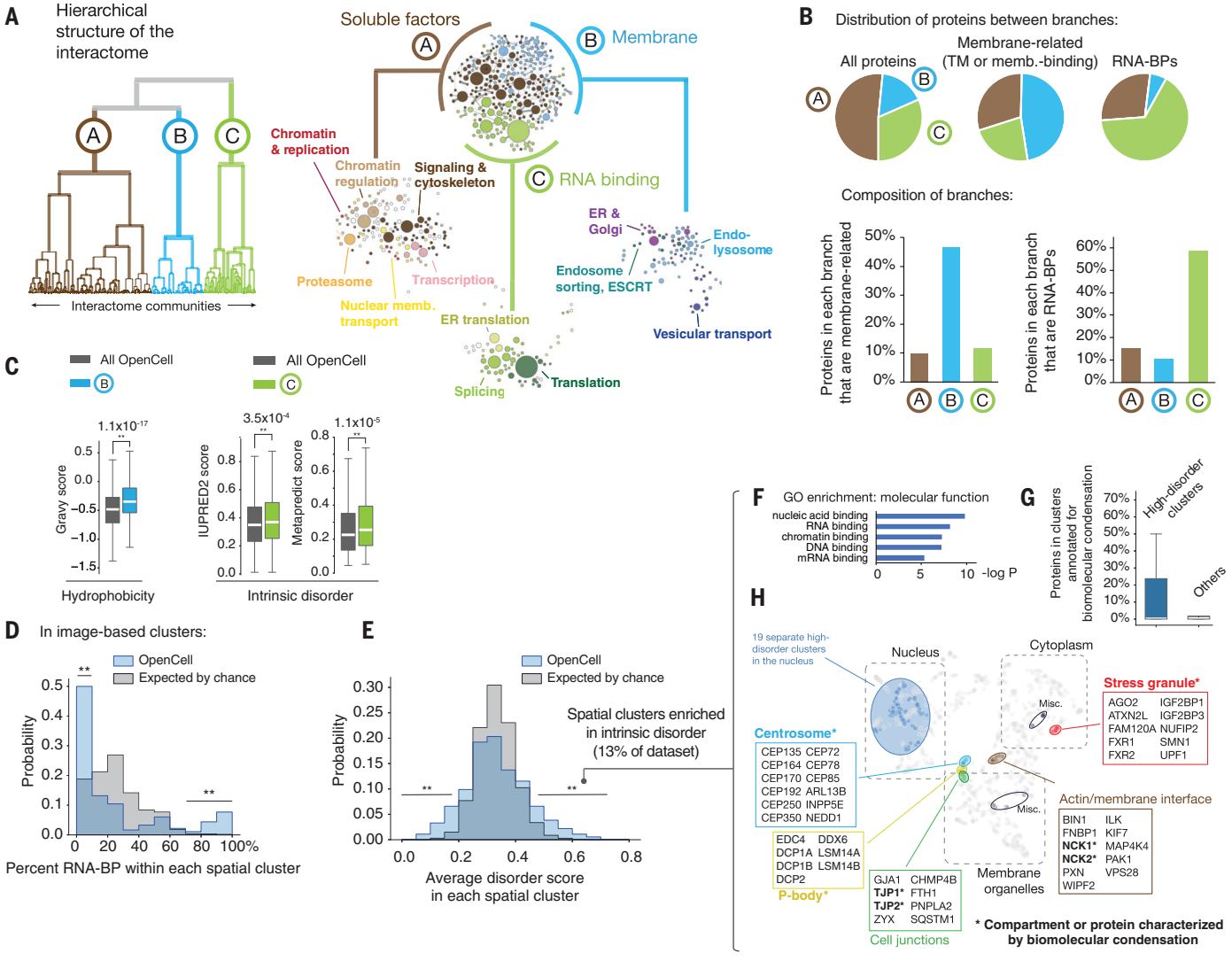


Fig. 5. Segregation of RNA-BPs in both interactome and imaging datasets. (**A**) Hierarchical structure of the interactome dataset; see full description in fig. S9B. (**B**) Distribution of membrane-related (transmembrane or membrane-binding) proteins and RNA-BPs within the three interactome branches. (**C**) Distribution of hydrophobicity and intrinsic disorder in the membrane and RNA-BP branches of the interactome hierarchy, respectively (see full analysis in fig. S10). For intrinsic disorder, two separate scores are shown for completeness: IUPRED2 (*89*) and metapredict (*90*), a new aggregative disorder scoring algorithm. Boxes represent 25th, 50th, and 75th percentiles; whiskers represent 1.5 times the interquartile range. Median is represented by a white line. $^{**}P < 10^{-3}$ (Student's *t* test); exact *P* values are shown. (**D**) Distribution of RNA-BP percentage across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1000 times. Lines indicate parts of the distribution overrepresented in our data versus control ($^{**}P < 2 \times 10^{-3}$, Fisher's exact *t* test). (**E**) Distribution of disorder score (IUPRED2) across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1000 times. Lines indicate parts of the distribution overrepresented in our data versus control ($^{**}P < 2 \times 10^{-3}$, Fisher's exact *t* test). (**F**) Ontology enrichment analysis of proteins contained in high-disorder spatial clusters (average disorder score >0.45). Enrichment compares to the whole set of OpenCell targets (*P* value: Fisher's exact test). (**G**) Prevalence of proteins annotated to be involved in biomolecular condensation in high-disorder versus other spatial clusters. Boxes represent 25th, 50th, and 75th percentiles; whiskers represent 1.5 times the interquartile range. Median is represented by a white line. Note that for both distributions, the median is zero. (**H**) Distribution of high-disorder spatial clusters in the UMAP embedding from Fig. 3D. Individual nuclear clusters are not outlined for readability. Multiple high-disorder spatial clusters include compartments or proteins characterized by biomolecular condensation behaviors, which are marked by an asterisk.

we reasoned that the biophysical properties of proteins within each interactome branch might underlie their segregation. Indeed, an analysis of protein sequence features revealed a separation of different biophysical properties in each branch (fig. S10, F and G). Branch B was enriched for hydrophobic sequences (Fig. 5C), consistent with its enrichment for membrane-related proteins, whereas branch C was enriched for intrinsic disorder (Fig. 5C). This is consistent with the fact that RNA-BPs are significantly more disordered than other proteins in the proteome (fig. S11A) (*62*). RNA-BPs are also among the most abundant in the cell (fig. S11B) and form a higher number of interactions than other proteins (fig. S11, C and D).

IP-MS measures protein interactions in vitro after lysis and therefore does not directly address the spatial relationship between interacting proteins. Thus, we sought to further examine how RNA-BPs distribute in our live-cell imaging data. If RNA-BPs segregate into interacting groups in vivo, this should also manifest at the level of their intracellular localization: They should enrich in the same spatial clusters derived from our unsupervised machine learning analysis. Indeed, the distribution of RNA-BP content within spatial clusters revealed a significant overrepresentation of clusters that were either strongly enriched or depleted for RNA-BPs (Fig. 5D). Because spatial clusters can be interpreted as defining "microcompartments" within the cell, both enrichment and depletion have functional implications: Not only are RNA-BPs enriched within the same microcompartments, they also tend to be excluded from others. Of the 26 spatial clusters (62%) that are highly enriched in RNA-BPs, 16 include at least one protein involved in biomolecular condensation [as curated in PhaSepDB (*63*)], which might reflect a prevalent role for biomolecular condensation in shaping the RNA-BP proteome. Collectively, both interactome and imaging data underscore that RNA-BPs (a prevalent group of proteins representing 13% of proteins expressed in HEK293T cells; see table S2) form a distinct subgroup within the proteome characterized by unique properties.

These results motivated a broader analysis of the contribution of intrinsic disorder to the spatial organization of the proteome in our dataset. Plotting the distribution of mean intrinsic disorder within spatial clusters revealed a significant overrepresentation of clusters both enriched and depleted in disordered proteins (Fig. 5E). Of 182 total spatial clusters, 26 were enriched for disordered proteins, covering 13% of the proteins in our imaging dataset. Overall, the extent to which disordered proteins segregated spatially was similar to the degree of segregation found for hydrophobic proteins: An analogous analysis revealed that 10% of proteins in our dataset are found within clusters significantly enriched for high hydrophobicity (fig. S12E), which map to membrane-bound organelles (fig. S12F). This supports the hypothesis that intrinsic disorder is as important a feature as hydrophobicity in organizing the spatial distribution of the human proteome. Consistent with our previous analysis, high-disorder clusters were enriched for RNA-BPs (Fig. 5F), with 15 of these 26 clusters containing more than 50% of RNA-BPs. High-disorder clusters were also enriched for proteins annotated to participate in biomolecular condensation (Fig. 5G) and were predominantly found in the nucleus (19 clusters, 73% of total, Fig. 5H). Five of seven high-disorder clusters found in the cytosol delineate compartments for which biomolecular condensation has been proposed to play an important role (Fig. 5G), namely P-bodies (*64*), stress granules (*59*), centrosome (*65*), cell junctions (*66*), and the interface between cell surface and actin cytoskeleton (*67*).

Interactive data sharing at opencell.czbiohub.org

To enable widespread access to the OpenCell datasets, we built an interactive web application that provides side-by-side visualizations of the 3D confocal images and of the interaction network for each tagged protein, together with RNA and protein abundances for the whole proteome (Fig. 6). Our web interface is fully described in fig. S12.

Discussion

OpenCell combines three strategies to augment the description of human cellular architecture. First, we present an integrated experimental pipeline for high-throughput cell biology, fueled by scalable methods for genome engineering, live-cell microscopy, and IP-MS. Second, we provide an open-source resource of well-curated localization and interactome measurements, easily accessible through an interactive web interface at opencell.czbiohub.org. Third, we describe an analytical framework for the representation and comparison of interaction or localization signatures (including a self-supervised machine learning approach for image encoding). Finally, we demonstrate how our dataset can be used for fine-grained mechanistic exploration (to explore the function of multiple proteins that were previously uncharacterized) as well as for investigating the core organizational principles of the proteome.

Our current strategy that combines split FPs and HEK293T—a cell line that is heavily transformed but easily manipulatable—is mostly constrained by scalability considerations. Technological advances are quickly broadening the set of cellular systems that can be engineered and profiled at scale. Advances in stem cell technologies enable the generation of libraries that can be differentiated in multiple cell types (*11*), while innovations in genome engineering [for example, by modulating DNA repair (*68*)] pave the way for the scalable insertion of gene-sized payload, for the combination of multiple edits in the same cell, or for increased homozygosity in polyclonal pools. In addition, recent developments in high-throughput light-sheet microscopy (*69*) might soon enable the systematic description of 4D intracellular dynamics (*70*).

A central feature of our approach is the use of endogenous fluorescent tags to study protein function. Genome-edited cells enable protein function to be examined at near-native expression levels [which can circumvent some limitations of overexpression (*71*)] and enable the measurement of protein localization in live cells [which can avoid artifacts caused by fixation or antibody labeling (*72*)]. Comparing our data to the current reference datasets of protein-protein interactions (fig. S4, C to F) or localization (fig. S7, C and D) highlights the performance of our strategy. In addition, our high success rate tagging essential genes [fig. S2A; see also (*73*) in yeast] and the successful tagging of the near-complete yeast proteome (*14*, *73*) indicate that fluorescent tagging generally preserves normal protein physiology.

However, limitations exist for specific protein targets. FPs are as big as an average human protein, and their insertion can impair function or localization—for example, by occluding important interaction interfaces or impairing subcellular targeting sequences. In other cases, tags can affect expression or degradation rates, which might explain why we find tagged proteins being expressed at 80% of their endogenous abundance, and 8% of targets in our dataset having outlier abundances at steady state (Fig. S3D). Further, tagging often cannot discriminate between different isoforms of a protein (such as splicing or posttranslationally modified variants). Finally, relying on endogenous expression can be an obstacle given the low concentration of most proteins in the human cell: Detection of poorly abundant proteins is difficult (especially those in the bottom half of the abundance distribution) even when using a very bright FP such as mNeonGreen (*74*) (fig. S2D). Solutions to this obstacle include using FP repeats to increase signal (*18*, *23*) or using tags that bind chemical fluorophores [e.g., HaloTag (*75*)], which can be brighter than FPs or operate at wavelengths where cellular autofluorescence is decreased (*76*). Overall, the full description of human cellular architecture remains a formidable challenge that will require complementary methods being applied in parallel. The diversity of large-scale cell biology approaches is a solution to this problem (*6*, *8*, *9*, *11*, *31*, *70*, *77*–*80*). Mirroring the advances in genomics following the human genome sequence (*2*), open-source systematic datasets will likely play an important role in how the growth of cell biology measurements can be transformed

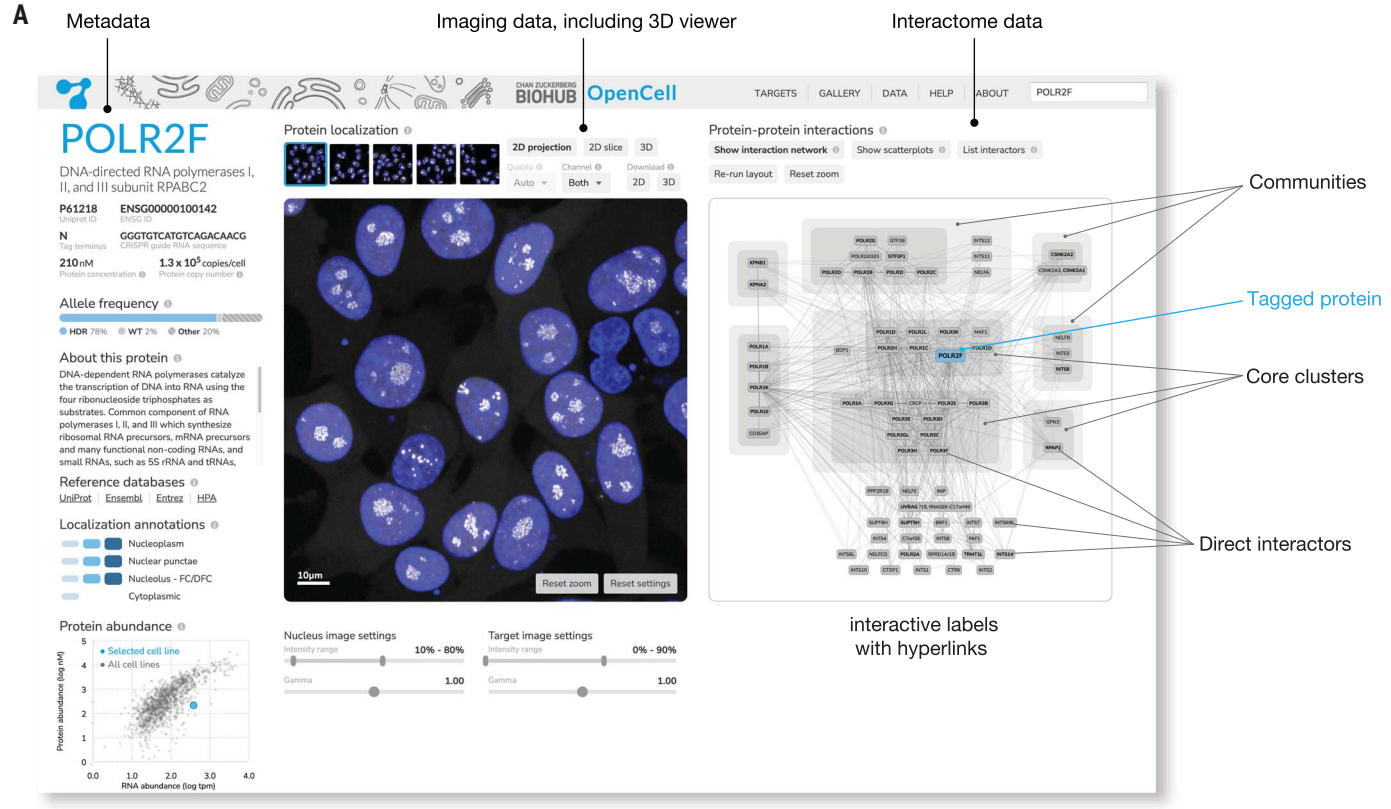


Fig. 6. The OpenCell website. Shown is an annotated screenshot from our web-app at https://opencell.czbiohub.org, which is described in more detail in fig. S12.

into fundamental discoveries by an entire community (*81*).

In addition to presenting a resource of measurements and protocols, we also demonstrate how our data can be used to study the global signatures that pattern the proteome. Our analysis reveals that RNA binding proteins, which form one of the biggest functional families in the cell, are characterized by a unique set of properties and segregate from other proteins in terms of both interactions and spatial distribution. It would be fascinating to explore the extent to which RNA itself might act as a structural organizer of the cellular proteome (*62*, *82*). This is, for example, the case for some noncoding RNAs whose main function is to template protein interactions to form nuclear bodies (*83*). High intrinsic disorder is one of the distinguishing features of RNA-BPs that likely contributes to their unique properties. Beyond RNA-BPs, our data support a general role for intrinsic disorder in shaping the spatial distribution of human proteins. For example, 13% of proteins in our dataset are found in spatial clusters that are significantly enriched for disordered proteins. This adds to the growing appreciation that intrinsic disorder, which is much more prevalent in eukaryotic than in prokaryotic proteomes (*84*, *85*), plays a key role in the functional subcompartmentalization of the eukaryotic nucleo- and cytoplasm in the context of biomolecular condensation (*86*).

Lastly, we show that the spatial distribution of each human protein is very specific, to the point that remarkably detailed functional relationships can be inferred on the sole basis of similarities between localization patterns, including the prediction of molecular interactions [which complements other studies (*87*)]. This highlights that intracellular organization is defined by fine-grained features that go beyond membership to a given organelle. Our demonstration that self-supervised deep learning models can identify complex but deterministic signatures from light microscopy images opens exciting avenues for the use of imaging as an information-rich method for deep phenotyping and functional genomics (*51*). Because light microscopy is easily scalable, can be performed live, and enables measurements at the single-cell level, this should offer rich opportunities for the full quantitative description of cellular diversity in normal physiology and disease.

Methods summary

See (*21*) for a complete description of methods for cell culture and CRISPR engineering, immunoprecipitation and mass spectrometry, live-cell imaging, and data analysis of both interactome and imaging datasets. In brief, HEK-293T cells (ATCC CRL-3216) were engineered to express in-frame fluorescent gene fusions using the split-mNeonGreen2 system (see Fig. 1A). In total, we targeted 1757 human genes in separate experiments and could successfully detect fluorescence for 1310 of these gene targets, which constitute our current dataset. CRISPR-based genomic insertions were performed by nucleofection of purified Cas9 protein precomplexed with synthetic guide RNAs, delivered together with single-stranded oligodeoxynucleotide donors to template homologous recombination. Fluorescent cells were selected by flow cytometry as a polyclonal pool, which was genotyped by next-generation amplicon sequencing of the edited alleles. These selected cell pools were used for functional characterization by microscopy and IP-MS. To image protein localization in live cells, we performed 3D spinning disk confocal microscopy (63× objective, 1.47 NA) under environmental control (37°C, 5% CO_2). Microscopy acquisition was fully automated in 96-well plates using a custom acquisition script, written in Python (github.com/czbiohub/2021-opencell-microscopy-automation). To measure protein interactions, we performed IP-MS from cell lysates solubilized using digitonin detergent in the presence

of benzonase (nuclease for DNA and RNA). After immunoprecipitation of target proteins using anti-mNeonGreen nanobody resin, captured proteins were digested into peptides by LysC protease. Bottom-up mass spectrometry analysis was then performed on a timsTOF instrument. Mass spectrometry data were quantified using MaxQuant. Data analysis was performed in Python as detailed in (*21*). The code and data used to generate the figures can be found on GitHub at github.com/czbiohub/2021-opencell-figures.

REFERENCES AND NOTES

1. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi: 10.1038/nature03001; pmid: 15496913
2. L. Hood, L. Rowen, The Human Genome Project: Big science transforms biology and medicine. *Genome Med.* **5**, 79 (2013). doi: 10.1186/gm483; pmid: 24040834
3. P. Nurse, J. Hayles, The cell in an era of systems biology. *Cell* **144**, 850–854 (2011). doi: 10.1016/j.cell.2011.02.045; pmid: 21414476
4. F. D. Mast, A. V. Ratushny, J. D. Aitchison, Systems cell biology. *J. Cell Biol.* **206**, 695–706 (2014). doi: 10.1083/jcb.201405027; pmid: 25225336
5. E. Lundberg, G. H. H. Borner, Spatial proteomics: A powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019). doi: 10.1038/s41580-018-0094-y; pmid: 30659282
6. K. Luck *et al*., A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020). doi: 10.1038/s41586-020-2188-x; pmid: 32296183
7. M. Y. Hein *et al*., A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015). doi: 10.1016/j.cell.2015.09.053; pmid: 26496610
8. E. L. Huttlin *et al*., Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017). doi: 10.1038/nature22366; pmid: 28514442
9. P. J. Thul *et al*., A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017). doi: 10.1126/science.aal3321; pmid: 28495876
10. H. Bukhari, T. Müller, Endogenous Fluorescence Tagging by CRISPR. *Trends Cell Biol.* **29**, 912–928 (2019). doi: 10.1016/j.tcb.2019.08.004; pmid: 31522960
11. B. Roberts *et al*., Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Mol. Biol. Cell* **28**, 2854–2874 (2017). doi: 10.1091/mbc.e17-03-0209; pmid: 28814507
12. S. Ghaemmaghami *et al*., Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003). doi: 10.1038/nature02046; pmid: 14562106
13. S. R. Collins *et al*., Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6**, 439–450 (2007). doi: 10.1074/mcp.M600381-MCP200; pmid: 17200106
14. U. Weill *et al*., Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods* **15**, 617–622 (2018). doi: 10.1038/s41592-018-0044-9; pmid: 29988094
15. A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute, C. Cullin, A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Res.* **21**, 3329–3330 (1993). doi: 10.1093/nar/21.14.3329; pmid: 8341614
16. I. Poser *et al*., BAC TransgeneOmics: A high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415 (2008). doi: 10.1038/nmeth.1199; pmid: 18391959
17. A. Sigal *et al*., Generation of a fluorescently labeled endogenous protein library in living human cells. *Nat. Protoc.* **2**, 1515–1527 (2007). doi: 10.1038/nprot.2007.197; pmid: 17571059
18. M. D. Leonetti, S. Sekine, D. Kamiyama, J. S. Weissman, B. Huang, A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3501–E3508 (2016). doi: 10.1073/pnas.1606731113; pmid: 27274053
19. N. C. Hubner *et al*., Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739–754 (2010). doi: 10.1083/jcb.200911091; pmid: 20479470
20. S. Feng *et al*., Improved split fluorescent proteins for endogenous protein labeling. *Nat. Commun.* **8**, 370 (2017). doi: 10.1038/s41467-017-00494-8; pmid: 28851864
21. See supplementary materials.
22. F. Meier *et al*., Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device. *J. Proteome Res.* **14**, 5378–5387 (2015). doi: 10.1021/acs.jproteome.5b00932; pmid: 26538118
23. D. Kamiyama *et al*., Versatile protein tagging in cells with split fluorescent protein. *Nat. Commun.* **7**, 11046 (2016). doi: 10.1038/ncomms11046; pmid: 26988139
24. Y.-C. Lin *et al*., Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* **5**, 4767 (2014). doi: 10.1038/ncomms5767; pmid: 25182477
25. S. Lin, B. T. Staahl, R. K. Alla, J. A. Doudna, Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, e04766 (2014). doi: 10.7554/eLife.04766; pmid: 25497837
26. J. B. Doyon *et al*., Rapid and efficient clathrin-mediated endocytosis revealed in genome-edited mammalian cells. *Nat. Cell Biol.* **13**, 331–337 (2011). doi: 10.1038/ncb2175; pmid: 21297641
27. T. J. Gibson, M. Seiler, R. A. Veitia, The transience of transient overexpression. *Nat. Methods* **10**, 715–721 (2013). doi: 10.1038/nmeth.2534; pmid: 23900254
28. E. C. Keilhauer, M. Y. Hein, M. Mann, Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14**, 120–135 (2015). doi: 10.1074/mcp.M114.041012; pmid: 25363814
29. J. A. Thomas, C. G. Tate, Quality control in eukaryotic membrane protein overproduction. *J. Mol. Biol.* **426**, 4139–4154 (2014). doi: 10.1016/j.jmb.2014.10.012; pmid: 25454020
30. M. Giurgiu *et al*., CORUM: The comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2018). doi: 10.1093/nar/gky973
31. 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; pmid: 27278775
32. E. L. Huttlin *et al*., Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *Cell* **184**, 3022–3040.e28 (2021). doi: 10.1101/2020.01.19.905109
33. L. Royer, M. Reimann, A. F. Stewart, M. Schroeder, Network compression as a quality measure for protein interaction networks. *PLOS ONE* **7**, e35729 (2012). doi: 10.1371/journal.pone.0035729; pmid: 22719828
34. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002). doi: 10.1093/nar/30.7.1575; pmid: 11917018
35. M. J. Shurtleff *et al*., The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *eLife* **7**, e37018 (2018). doi: 10.7554/eLife.37018; pmid: 29809151
36. D. Acosta-Alvear *et al*., The unfolded protein response and endoplasmic reticulum protein targeting machineries converge on the stress sensor IRE1. *eLife* **7**, e43036 (2018). doi: 10.7554/eLife.43036; pmid: 30582518
37. P. T. McGilvray *et al*., An ER translocon for multi-pass membrane protein biogenesis. *eLife* **9**, e56889 (2020). doi: 10.7554/eLife.56889; pmid: 32820719
38. P. J. Chitwood, R. S. Hegde, An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature* **584**, 630–634 (2020). doi: 10.1038/s41586-020-2624-y; pmid: 32814900
39. T. Stoeger, M. Gerlach, R. I. Morimoto, L. A. Nunes Amaral, Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biol.* **16**, e2006643 (2018). doi: 10.1371/journal.pbio.2006643; pmid: 30226837
40. S. P. Brooks *et al*., The Nance-Horan syndrome protein encodes a functional WAVE homology domain (WHD) and is important for co-ordinating actin remodelling and maintaining cell morphology. *Hum. Mol. Genet.* **19**, 2421–2432 (2010). doi: 10.1093/hmg/ddq125; pmid: 20332100
41. A.-L. Law *et al*., Nance-Horan Syndrome-like 1 protein negatively regulates Scar/WAVE-Arp2/3 activity and inhibits lamellipodia stability and cell migration. *Nat. Commun.* **12**, 5687 (2021). doi: 10.1101/2020.05.11.083030
42. A. Schossig *et al*., Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. *Am. J. Hum. Genet.* **90**, 701–707 (2012). doi: 10.1016/j.ajhg.2012.02.012; pmid: 22424600
43. M. Merkulova *et al*., Mapping the H^+ (V)-ATPase interactome: Identification of proteins involved in trafficking, folding, assembly and phosphorylation. *Sci. Rep.* **5**, 14827 (2015). doi: 10.1038/srep14827; pmid: 26442671
44. Y. Yan, N. Denef, T. Schüpbach, The vacuolar proton pump, V-ATPase, is required for notch signaling and endosomal trafficking in Drosophila. *Dev. Cell* **17**, 387–402 (2009). doi: 10.1016/j.devcel.2009.07.001; pmid: 19758563
45. T. Vasanthakumar, J. L. Rubinstein, Structure and Roles of V-type ATPases. *Trends Biochem. Sci.* **45**, 295–307 (2020). doi: 10.1016/j.tibs.2019.12.007; pmid: 32001091
46. D. Görlich, U. Kutay, Transport between the cell nucleus and the cytoplasm. *Annu. Rev. Cell Dev. Biol.* **15**, 607–660 (1999). doi: 10.1146/annurev.cellbio.15.1.607; pmid: 10611974
47. C. P. Lusk, M. C. King, The nucleus: Keeping it together by keeping it apart. *Curr. Opin. Cell Biol.* **44**, 44–50 (2017). doi: 10.1016/j.ceb.2017.02.001; pmid: 28236735
48. M. Breker, M. Gymrek, M. Schuldiner, A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *J. Cell Biol.* **200**, 839–850 (2013). doi: 10.1083/jcb.201301120; pmid: 23509072
49. D. P. Sullivan *et al*., Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018). doi: 10.1038/nbt.4225; pmid: 30125267
50. W. Ouyang *et al*., Analysis of the Human Protein Atlas Image Classification competition. *Nat. Methods* **16**, 1254–1261 (2019). doi: 10.1038/s41592-019-0658-6; pmid: 31780840
51. S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, A. E. Carpenter, Image-based profiling for drug discovery: Due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021). doi: 10.1038/s41573-020-00117-w; pmid: 33353986
52. H. Kobayashi, K. C. Cheveralls, M. D. Leonetti, L. A. Royer, Self-Supervised Deep-Learning Encodes High-Resolution Features of Protein Subcellular Localization. bioRxiv 437595 [preprint] (2021). doi: doi:10.1101/2021.03.29.437595
53. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv 1802.03426 (2018).
54. E. Meijering, A. E. Carpenter, H. Peng, F. A. Hamprecht, J.-C. Olivo-Marin, Imagining the future of bioimage analysis. *Nat. Biotechnol.* **34**, 1250–1255 (2016). doi: 10.1038/nbt.3722; pmid: 27926723
55. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019). doi: 10.1038/s41598-019-41695-z; pmid: 30914743
56. S. Markmiller *et al*., Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590–604.e13 (2018). doi: 10.1016/j.cell.2017.12.032; pmid: 29373831
57. J.-Y. Youn *et al*., High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol. Cell* **69**, 517–532.e11 (2018). doi: 10.1016/j.molcel.2017.12.020; pmid: 29395067
58. H. Marmor-Kollet *et al*., Spatiotemporal Proteomic Analysis of Stress Granule Disassembly Using APEX Reveals Regulation by SUMOylation and Links to ALS Pathogenesis. *Mol. Cell* **80**, 876–891.e6 (2020). doi: 10.1016/j.molcel.2020.10.032; pmid: 33217318
59. P. Yang *et al*., G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**, 325–345.e28 (2020). doi: 10.1016/j.cell.2020.03.046; pmid: 32302571
60. M. Costanzo *et al*., A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016). doi: 10.1126/science.aaf1420; pmid: 27708008
61. M. A. Horlbeck *et al*., Mapping the Genetic Landscape of Human Cells. *Cell* **174**, 953–967.e22 (2018). doi: 10.1016/j.cell.2018.06.010; pmid: 30033366
62. A. Balcerak, A. Trebinska-Stryjewska, R. Konopinski, M. Wakula, E. A. Grzybowska, RNA-protein interactions: Disorder, moonlighting and junk contribute to eukaryotic complexity. *Open Biol.* **9**, 190096 (2019). doi: 10.1098/rsob.190096; pmid: 31213136
63. K. You *et al*., PhaSepDB: A database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* **48** (D1), D354–D359 (2020). doi: 10.1093/nar/gkz847; pmid: 31584089

64. Y. Luo, Z. Na, S. A. Slavoff, P-Bodies: Composition, Properties, and Functions. *Biochemistry* **57**, 2424–2431 (2018). doi: 10.1021/acs.biochem.7b01162; pmid: 29381060
65. J. B. Woodruff *et al*., The Centrosome Is a Selective Condensate that Nucleates Microtubules by Concentrating Tubulin. *Cell* **169**, 1066–1077.e10 (2017). doi: 10.1016/j.cell.2017.05.028; pmid: 28575670
66. O. Beutel, R. Maraspini, K. Pombo-García, C. Martin-Lemaitre, A. Honigmann, Phase Separation of Zonula Occludens Proteins Drives Formation of Tight Junctions. *Cell* **179**, 923–936.e11 (2019). doi: 10.1016/j.cell.2019.10.011; pmid: 31675499
67. S. Banjade *et al*., Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E6426–E6435 (2015). doi: 10.1073/pnas.1508778112; pmid: 26553976
68. S. Riesenberg *et al*., Simultaneous precise editing of multiple genes in human cells. *Nucleic Acids Res.* **47**, e116 (2019). doi: 10.1073/pnas.1508778112; pmid: 26553976
69. B. Yang *et al*., Epi-illumination SPIM for volumetric imaging with high spatial-temporal resolution. *Nat. Methods* **16**, 501–504 (2019). doi: 10.1038/s41592-019-0401-3; pmid: 31061492
70. Y. Cai *et al*., Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018). doi: 10.1038/s41586-018-0518-z; pmid: 30202089
71. C. von Mering *et al*., Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002). doi: 10.1038/nature750; pmid: 12000970
72. U. Schnell, F. Dijk, K. A. Sjollema, B. N. G. Giepmans, Immunolabeling artifacts and the need for live-cell imaging. *Nat. Methods* **9**, 152–158 (2012). doi: 10.1038/nmeth.1855; pmid: 22290187
73. W.-K. Huh *et al*., Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003). doi: 10.1038/nature02026; pmid: 14562095
74. N. C. Shaner *et al*., A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nat. Methods* **10**, 407–409 (2013). doi: 10.1038/nmeth.2413; pmid: 23524392
75. G. V. Los *et al*., HaloTag: A novel protein labeling technology for cell imaging and protein analysis. *ACS Chem. Biol.* **3**, 373–382 (2008). doi: 10.1021/cb800025k; pmid: 18533659
76. L. D. Lavis, Chemistry Is Dead. Long Live Chemistry! *Biochemistry* **56**, 5165–5170 (2017). doi: 10.1021/acs.biochem.7b00529; pmid: 28704030
77. C. D. Go *et al*., A proximity-dependent biotinylation map of a human cell. *Nature* **595**, 120–124 (2021). doi: 10.1038/s41586-021-03592-2; pmid: 34079125
78. G. Gut, M. D. Herrmann, L. Pelkmans, Multiplexed protein maps link subcellular organization to cellular states. *Science* **361**, eaar7042 (2018). doi: 10.1126/science.aar7042; pmid: 30072512
79. J. R. A. Hutchins *et al*., Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**, 593–599 (2010). doi: 10.1126/science.1181348; pmid: 20360068
80. P. C. Havugimana *et al*., A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012). doi: 10.1016/j.cell.2012.08.011; pmid: 22939629
81. J. Ellenberg *et al*., A call for public archives for biological image data. *Nat. Methods* **15**, 849–854 (2018). doi: 10.1038/s41592-018-0195-8; pmid: 30377375
82. M. W. Hentze, A. Castello, T. Schwarzl, T. Preiss, A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018). doi: 10.1038/nrm.2017.130; pmid: 29339797
83. T. Chujo, T. Hirose, Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function. *Mol. Cells* **40**, 889–896 (2017). doi: 10.14348/molcells.2017.0263; pmid: 29276943
84. Z. Peng *et al*., Exceptionally abundant exceptions: Comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **72**, 137–151 (2015). doi: 10.1007/s00018-014-1661-9; pmid: 24939692
85. W. Basile, M. Salvatore, C. Bassot, A. Elofsson, Why do eukaryotic proteins contain more intrinsically disordered regions? *PLOS Comput. Biol.* **15**, e1007186 (2019). doi: 10.1371/journal.pcbi.1007186; pmid: 31329574
86. Y. Shin, C. P. Brangwynne, Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017). doi: 10.1126/science.aaf4382; pmid: 28935776
87. Y. Qin *et al*., A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021). doi: 10.1038/s41586-021-04115-9; pmid: 34819669
88. L. Hubert, P. Arabie, Comparing partitions. *J. Classif.* **2**, 193–218 (1985). doi: 10.1007/BF01908075
89. B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018). doi: 10.1007/BF01908075
90. R. J. Emenecker, D. Griffith, A. S. Holehouse, Metapredict: A fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120**, 4312–4319 (2021). doi: 10.1016/j.bpj.2021.08.039; pmid: 34480923

ACKNOWLEDGMENTS

We thank N. Neff, M. Tan, R. Sit, and A. Detweiler for help with high-throughput sequencing; G. Margulis, A. Sellas, E. Ho, and J. Mann for operational support; A. McGeever for help with web application architecture and deployment; and S. Schmid for critical feedback. M.D.L. thanks C. L. Tan for continuous discussions. **Funding:** H.K. was supported by an International Research Fellowship from the Japan Society for the Promotion of Science. R.M.B. was supported by a NIH predoctoral fellowship (F31 HL143882). B.H. was supported by NIH (R01GM131641) and is a Chan Zuckerberg Biohub Investigator. J.S.W. was supported by NIH (1RM1HG009490) and is an investigator with the Howard Hughes Medical Institute. M.M. was supported by a Max Planck Society for the Advancement of Science award. **Competing interests:** J.S.W. declares outside interest in Chroma Therapeutics, KSQ Therapeutics, Maze Therapeutics, Amgen, Tessera Therapeutics and 5 AM Ventures. M.M. is an indirect shareholder in EvoSep Biosystems. **Author contributions:** Conceptualization: M.D.L., M.M., J.S.W., B.H., M.Y.H. Methodology: M.D.L., M.M., L.A.R., D.N.I., R.G.-S., J.S.W., S.B.M., B.H., M.Y.H., K.K., K.C.C., N.H.C. Investigation: N.H.C., K.C.C., A.-D.B., K.K., A.C.M., P.R., H.K., L.S., J.Y.L., H.C., J.Y.S.K., E.M.S., C.G., F.M., J.P.C., R.M.B., B.B.C., G.D., M.Y.H., D.N.I., M.D.L. Visualization: M.D.L., K.C.C., K.K., M.Y.H. Funding acquisition: M.D.L., M.M., L.A.R., D.I.N., R.G.-S., J.S.W., S.B.M., B.H. Project administration: M.D.L. Supervision: M.D.L., M.M., L.A.R., D.N.I., R.G.-S., J.S.W., S.B.M., B.H. Writing–original draft: M.D.L., K.C.C., J.S.W., M.M., N.H.C., A.-D.B., K.K., A.C.M., P.R., M.Y.H. Writing–review and editing: M.D.L., K.C.C., K.K., M.Y.H. **Data and materials availability:** Mass spectrometry raw data and associated MaxQuant output tables are deposited to the ProteomeXchange Consortium via the PRIDEpartner repository (accession PXD024909 for interactome data, and accession PXD029191 for whole-cell abundance data). Bulk RNA-seq raw data and associated kallisto transcript abundance tables are available on GEO (accession GSE186192). Raw microscopy images are hosted by AWS's Open Datasets Program at https://registry.opendata.aws/czb-opencell/.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abi6983
Materials and Methods
Figs. S1 to S12
Tables S1 to S9
References (*91*–*106*)

View/request a protocol for this paper from *Bio-protocol*.

26 March 2021; accepted 18 January 2022
10.1126/science.abi6983

OpenCell: Endogenous tagging for the cartography of human cellular organization

Nathan H. ChoKeith C. CheverallsAndreas-David BrunnerKibeom KimAndré C. MichaelisPreethi RaghavanHirofumi KobayashiLaura SavyJason Y. LiHera CanajJames Y. S. KimEdna M. StewartChristian GnannFrank McCarthyJoana P. CabreraRachel M. BrunettiBryant B. ChhunGreg DingleMarco Y. HeinBo HuangShalin B. MehtaJonathan S. WeissmanRafael Gómez-SjöbergDaniel N. ItzhakLoïc A. RoyerMatthias MannManuel D. Leonetti

Science, 375 (6585), eabi6983. • DOI: 10.1126/science.abi6983

Tracking proteins

Improved understanding of how proteins are organized within human cells should enhance our systems-level understanding of how cells function. Cho *et al*. used CRISPR technology to express more than 1000 different proteins at near endogenous amounts with labels that allowed both fluorescent imaging of their location and immunoprecipitation and mass spectrometry analysis of interacting protein partners (see the Perspective by Michnick and Levy). The large-scale data are made available on an interactive website, with clustering and analysis performed by machine learning. The studies emphasize the unusual properties of RNA-binding proteins and indicate that protein localization is very specific and may allow predictions of function. —LBR