OpenCell: proteome-scale endogenous tagging enables 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[1,4,5], Shalin B. Mehta[1], Jonathan S. Weissman[6,7], Rafael Gómez-Sjöberg[1], Daniel N. Itzhak[1], Loic A. Royer[1], Matthias Mann[2,8], Manuel D. Leonetti[1,*]

[1] Chan Zuckerberg Biohub, San Francisco, 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, USA; [5] Department of Pharmaceutical Chemistry, University of California, San Francisco USA; [5] Chan Zuckerberg Initiative, Redwood City, USA; [6] Whitehead Institute, Koch Institute and Department of Biology, Massachusetts Institute of Technology, and Howard Hughes Medical Institute, Cambridge, USA; [7] Department of Cellular and Molecular Pharmacology, , University of California, San Francisco, USA; [8] NNF Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

† equal contribution; * correspondence: manuel.leonetti@czbiohub.org

***Abstract.* Elucidating the wiring diagram of the human cell is one of the central goals of the post-genomic era. Here, we integrate genome engineering, confocal imaging, mass spectrometry and data science to systematically map protein localization in live cells and protein interactions under endogenous expression conditions. For this, we generated a library of 1,311 CRISPR-edited cell lines harboring fluorescent tags that also serve as handles for affinity capture, and applied a new machine learning framework to encode the interaction and localization profiles of each protein. Our approach provides a data-driven description of the molecular and spatial networks that organize the human proteome. We show that unsupervised clustering of these networks delineates functional groups and facilitates biological discovery, while hierarchical analyses uncover the core features that template cellular architecture. Furthermore, we discover that localization signatures are remarkably predictive of protein function, and often contain enough information to identify molecular interactions. Paired with a fully interactive website (opencell.czbiohub.org), OpenCell is a resource for the quantitative cartography of human cellular organization.**

The sequencing of 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, so that 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 (Y2H) (6) or epitope tagging coupled to immunoprecipitation-mass spectrometry (IP-MS) (7, 8), while protein localization has been charted using immuno-fluorescence in fixed samples (9). However, these approaches do not measure protein interactions under native expression conditions (which can be more precise than using ectopic methods (10)), nor define protein localization in live and unperturbed cells. Furthermore, protein interactions and spatial localization have so far mostly been addressed separately. Therefore, the description of human cellular organization remains incomplete. By contrast, seminal work in the budding yeast S. cerevisiae has demonstrated how libraries of endogenously tagged strains can enable the comprehensive mapping of localization and interactions in a eukaryotic proteome (11–13). These libraries were made possible by the relative simplicity of homologous recombination in yeast (14) and enable the functional characterization of proteins in their native cellular context. Excitingly, recent advances in CRISPR-mediated genome engineering now allow for similar strategies to be applied for the interrogation of the human cell (15, 16).

Here, we combine experimental and analytical innovations to create OpenCell, a systematic proteomic map of human cellular architecture. We generated a library of 1,311 CRISPR-edited HEK293T cell lines expressing fluorescent protein fusions from endogenous genomic loci, which we characterized by pairing confocal microscopy and mass spectrometry. Our dataset constitutes the most comprehensive image collection of live-cell protein localization to date, while integration with IP-MS using the fluorescent tags for capture enables measurement of localization and interactions from the same samples. For a quantitative description of cellular architecture, we introduce a data-driven framework to quantitatively represent protein features, supported by a new machine learning approach for image encoding. This analysis allows us to delineate communities of functionally related proteins by unsupervised clustering and provides mechanistic insights on specific proteins or pathways, including many proteins that had so far remained uncharacterized. This approach further enables a hierarchical description of the human proteome's organization, and highlights in particular that intrinsically disordered proteins – and notably RNA-binding proteins – exhibit very unique functional signatures that shape the proteome's network. Finally, a direct comparison of imaging and mass spectrometry data establishes that localization patterns measured by light microscopy often contain enough information to predict interactions at the molecular scale.

Results

The OpenCell library Fluorescent protein (FP) fusions are versatile tools that enable both the study of protein localization by microscopy and that of protein-protein interactions by acting as affinity handles for IP-MS (15, 17) (Fig. S1A). Here, we constructed a library of fluorescently tagged HEK293T cell lines by targeting human genes with the split-mNeonGreen2 system (18) (Fig. 1A). Split-FPs greatly simplify CRISPR-based genome engineering (15), which allowed us to generate fusions directly into endogenous genomic loci and to preserve native expression regulation (Fig. 1B). A full description of our pipeline is available in the Methods section (summarized in Figs. 1C-E). In brief, FP insertions sites (N- or C-terminus) were informed by literature curation or structural analysis. For each tagged target we isolated a polyclonal pool of CRISPR-edited cells, which was then characterized by live-cell 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 software for CRISPR-based integration experiments enabling guide RNA selection and homology donor sequence design (github.com/czbiohub/crispycrunch). We also fully automated microscopy acquisition in Python to enable on-the-fly computer vision and selection of desirable field of views imaged in 96-well plates (github.com/czbiohub/opencell-microscopy-automation). Furthermore, our mass-spectrometry protocols take advantage of the high sensitivity of timsTOF instruments (19) which allowed miniaturization of IP-MS down to 0.8x10^6 cells of starting material (Fig. S1B; 12-well plate culture, a >10-fold reduction compared to previous approaches (7, 8)).

In total, we targeted 1728 genes, of which 1311 (76%) could be successfully detected by fluorescence and form our current dataset (full library details in Suppl. Table 1). From these, we obtained paired IP/MS measurements for 1261 targets (96% of the publication set, Fig. 1D). Expression level is the main limitation to the successful detection of endogenous fluorescent fusions. Indeed, RNA-Seq analysis revealed that "unsuccessful" targets are expressed at significantly lower levels than successful ones (Fig. 1D, bottom left panel), and a correlation between transcript abundance and protein fluorescence identified the expression threshold corresponding to our fluorescence detection limit (log[RNA tpm] = 1.5, Fig. 1D, bottom right panel).

This threshold corresponds to the median expression level in the HEK293T line (Fig. S1C), meaning that the top ~50% of expressed genes are detectable at endogenous level using current FPs.

To maximize throughput, we used a polyclonal strategy to select genome-edited cells by fluorescent cell sorting. These polyclonal pools contain cells with different genotypes. On the one hand, HEK293T are pseudo-triploid (20) and a single edited allele is sufficient to confer fluorescence. On the other hand, different DNA repair mechanisms compete with homologous recombination for the resolution of CRISPR-induced genomic breaks (21) so that alleles containing non-functional mutations can be present in addition to the desired fusion alleles. However, such alleles do not support fluorescence and are therefore unlikely to impact downstream measurements, especially in the context of a polyclonal pool. In addition, we derived a stringent selection scheme to significantly enrich for fluorescent fusion alleles (Fig. S1D). Our final cell library has a median 61% of mNeonGreen-integrated alleles, 5% wild-type and 26% other non-functional alleles (Fig. S1E, full CRISPR design genotype information in Suppl. Table 1).

Finally, we verified that our engineering approach maintained the endogenous expression level of the tagged targets. For this, we quantified protein expression by Western blotting using antibodies specific to proteins targeted in 12 different cell pools. This analysis revealed that the median abundance of a target protein in engineered lines was 79% of its abundance in wt HEK293T cells (Fig. S1F). Thus, our gene-editing strategy preserves near-endogenous abundances and circumvents the limitations of ectopic overexpression (16, 22, 23), which include aberrant localization, changes in organellar morphology or masking effects (see the respective examples of SPTLC1, TOMM20 and MAP1LC3B in Fig. S1G). Therefore, OpenCell supports the functional profiling of tagged proteins in their native cellular context.

The OpenCell interactome. Affinity enrichment coupled to mass spectrometry is an efficient and sensitive method for the systematic mapping of protein interactions networks (24). We isolated tagged proteins ("baits") from cell lysates solubilized in digitonin, a mild non-ionic detergent known to preserve the native structure and properties of membrane proteins (25). Specific protein interactors ("preys") were identified from biological triplicate experiments using label-free bottom-up proteomics on a timsTOF instrument (19) (see Figure S2A-B for a detailed description of our statistical analysis, which builds upon established methods (7)). In total, the full interactome from our 1261 OpenCell baits includes 30,293 interactions between 5271 proteins (baits and preys, Fig. 2A, full interactome data in Suppl. Table 2).

To assess the quality of our interactome, we estimated its precision and recall using reference data (Fig. S2B). For recall analysis, we quantified the coverage in our data of interactions included in CORUM(26), 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 (27) (Fig. S2B). Benchmarking OpenCell against other large-scale interactomes, we compared its precision and recall to Bioplex (overexpression of HA-tagged baits (8, 28)), HuRI (Y2H (6)) and our own previous data (GFP fusions expressed from bacterial artificial chromosomes (7)) (Fig. S2C-E). We also calculated compression rates for each dataset as a measure of the overall inter-connectivity in the interaction networks (29) (Fig. S2F). Across all metrics, OpenCell outperformed previous approaches. Together, these findings establish the high quality of our interactome, which likely reflects both our preservation of near-endogenous protein expression and the sensitivity of our mass spectrometry analyses.

Analyzing the distribution of the number of interactors per tagged target allowed us to investigate the global properties of the human interaction network (Fig. 2B). While this distribution approximates a power-law for moderate interaction counts (i.e., a linear relationship in log-log scale, Fig. 2B), our data shows significantly more targets with a large number of interactors ($N_{interaction}$≥100 in Fig. 2B) than expected from a "scale-free" model of protein interaction networks(30), as has been noted in other analyses (31). We discovered that highly interacting proteins (i.e., $N_{interaction}$≥100) are not simply the most highly expressed (Fig. 2C), but rather exhibit specific biophysical signatures.

A sequence-base analysis showed that highly interacting proteins are significantly less a-helical and less hydrophobic than other proteins, but more likely to contain disordered domains (Fig. 2D, 2E). Ontology analysis also revealed that the group is enriched for RNA-binding proteins (Fig. 2F). The propensity of highly interacting proteins for high intrinsic disorder has been recognized in Y2H datasets (32). Intrinsic disorder is also a common property of proteins that form biomolecular condensates, which include a large number of RNA-binding proteins (33, 34). Interestingly, disorder has been proposed to be under positive selection in viral proteomes as a means for the relatively small number of proteins encoded in viral genomes to be able to interact pleiotropically with the host machinery (35).

Stoichiometry-driven clustering of interaction signatures. A powerful way to interpret interactomes is to identify communities of interactors. These communities highlight complexes or functional pathways and also facilitate the assignment of protein function via "guilt-by-association" (8, 12). To this end, we applied unsupervised Markov clustering (MCL) (36) to the graph of interactions defined by our data (5148 baits and preys). We first measured the stoichiometry of each interaction as a proxy for interaction strength, using a quantitative approach we previously established (7), and used it to weigh the edges in the interaction graph (Fig. 2G). A first round of stoichiometry-weighted Markov clustering delineated inter-connected protein communities and outperformed clustering on the basis of connectivity alone (Fig. S2G). To further refine annotations, we subjected each individual MCL community to another round of clustering in which low-stoichiometry interactions were removed. The resulting sub-clusters outline core interactions within existing communities (Fig. 2G). An illustrative example of how this unsupervised approach allows us to delineate functionally related proteins is shown in Figure 2H: all subunits of the machinery responsible for the translocation of newly translated proteins at the ER membrane (SEC61/62/63) and of the EMC (ER Membrane Complex) are grouped within respective core interaction clusters, but both are part of the same larger MCL community. This mirrors the recently appreciated co-translational role of EMC for insertion of transmembrane domains at the ER (37). Interestingly, additional proteins, which have only been recently shown to have co-translational role, are found clustering with translocon or EMC subunits. These including ERN1 (IRE1), a folding sensor (38), and CCDC47, a poorly characterized translocon interactor which, like EMC, regulates the biogenesis of membrane proteins at the ER (39, 40). This highlights how clustering can facilitate mechanistic exploration by grouping together proteins involved in related pathways. Overall, we identified 300 communities including a total of 2154 baits and preys (full details in Suppl. Table 2). A graph of interactions between communities reveals a richly inter-connected network (Fig. 2I), the structure of which outlines the global architecture of the human interactome (discussed further below).

The OpenCell image dataset. 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 fluorescence microscopy on a spinning-disk confocal microscope equipped with a 63x 1.47NA objective under environmental control (37°C, 5% CO_2), and imaged the full 3D distribution of proteins in consecutive z-slices. Microscopy acquisition was fully automated in Python to enable scalability (Fig. S3A-B). In particular, we trained a computer vision model to identify fields of view (FOVs) with homogeneous cell density on-the-fly, significantly reducing uncontrolled experimental variation between images. Our resulting dataset contains a collection of 5176 3D stacks (4-5 different FOVs for each target) and includes paired imaging of nuclear morphology with Hoechst 33342, a cell-permeable DNA dye compatible with live-cell measurements.

We manually annotated localization patterns by assigning each protein to one or more of 15 separate cellular compartments such as nucleolus, centrosome or Golgi apparatus (see Figure 3A for the full list and example images). Because proteins often populate multiple compartments at steady-state (9), we annotated localizations using a three-tier grade system: grade 3 identifies the most prominent localization compartment(s), grade 2 represents clearly detectable but minor localizations, and grade 1 annotates weak localization patterns nearing our limit of detection (see Figure S4A for two representative examples, full annotations in Suppl. Table 3). Ignoring grade 1 annotations which are inherently less precise, 55% of proteins in our library are multi-localizing and outline known functional relationships with, for example, clear connections between secretory compartments (ER, Golgi, vesicles, plasma membrane), or between cytoskeleton and plasma membrane (Fig. 3B). The most common pattern of multi-localization involves proteins found in both nucleus and cytoplasm (21% of our whole library), highlighting the importance of the nucleo-cytoplasmic import/export machinery in shaping global cellular function(41, 42). Importantly, because our split-FP system does not enable the detection of proteins in the lumen of organelles, multi-localization involving translocation across an organellar membrane (which is rare but does happen for mitochondrial or peroxisomal proteins) will not be detected in our data.

Quantitative localization encoding with self-supervised machine learning. Extracting functional insights directly from cellular images is a major goal of modern cell biology and data science (43). Because the function of a protein is tightly linked to its localization, we explored whether a quantitative comparison of localization signatures would allow us to delineate groups of co-functioning proteins. For this, we developed a deep learning model, which is fully described in a companion study (44). 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 “localization encoding” (Fig. 3C) that captures the complex set of features that define the localization of a protein across the full cell population. 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 (45)), our method extracts localization signatures from raw images without any a priori assumptions, manually assigned labels, or other additional information. This allows us to objectively compare proteins by measuring the similarity between their localization signatures.

A UMAP representation of the localization encodings for the entire OpenCell library is shown in Figure 3D. This map is organized in distinct territories that closely match manual annotations (Fig. 3D, highlighting mono-localizing proteins). This validates that our approach yields a quantitative representation of the biologically relevant information in our microscopy data. We then asked what degree of functional relationship could be inferred between proteins solely on the basis of their localization patterns. For this, we employed an unsupervised Leiden clustering strategy classically used to identify cell types in single-cell RNA sequencing datasets (46). Strikingly, applying this data-driven approach identified groups of proteins that are closely related mechanistically (182 clusters in total, full list in Suppl. Table 3). For example (Figure 3E), our analysis not only separated P-body proteins (cluster #83) from other forms of punctated cytoplasmic structures, but also unambiguously differentiated vesicular trafficking pathways despite very similar localization patterns: the endosomal machinery (cluster #40), plasma membrane endocytic pits (cluster #117) or COP-II vesicles (cluster #143) were all delineated with high precision. Proteins involved in closely inter-related cellular functions were also found to cluster together: for example, the ER translocon clusters with the SRP receptor, EMC subunits and the OST glycosylation complex, all responsible for co-translational operations (cluster #9). Clustering performance was also high for non-organellar proteins, as shown in Figure S4B (cytoplasmic clusters) and Fig S4C (nuclear clusters). Altogether, our results show that the localization pattern of a given protein reflects its function with high specificity, down to specific pathways, and that this information can be captured by self-supervised deep learning algorithms. While closely related proteins cluster tightly together, the existence of many separated groups also underlines the fascinating diversity of localization patterns across the full proteome. Images from nuclear proteins offer compelling illustrative examples of this diversity (Fig. S4D).

Interactive data sharing at opencell.czbiohub.org To enable widespread access to the OpenCell datasets, we built an interactive web app that provides side-by-side visualizations of the confocal images and of the interaction network for each tagged protein (Fig. 4A-B). The app is organized around a 'target profile' page that displays all of the metadata, images, and interactions for a selected mNG11-tagged target (Fig. 4B). Confocal fluorescent images can be visualized either in 2D as a scrollable stack of z-slices, or in 3D via an interactive volume rendering module we developed (Fig. 4C). Our interface also allows the user to toggle between fluorescence channels (tagged protein and DNA stain) and to adjust image brightness and contrast. The interaction network, which consists of the target, its direct interactors, and the interactions between them, is organized by the communities and core clusters identified by Markov clustering and is positioned directly adjacent to the image viewer (Fig. 4B). This side-by-side juxtaposition of images and interactions encourages the comparison between subcellular localizations and interaction signatures. The interactome visualization can be switched to reveal quantitative data for each pull-down in the form of volcano and stoichiometry plots (Fig. 4D). To enable the interactive navigation of the full interaction network, interactors in the network that are themselves OpenCell targets are hyperlinked (from any visualization mode) to their corresponding target pages. Interacting proteins in the network that are not OpenCell targets are likewise hyperlinked to a distinct 'interactor profile' page (Fig. 4A, middle panel) that displays information about all the pull-downs in which the interacting protein appears. Finally, to explore the dataset by localization pattern, a separate 'gallery' page displays a grid of thumbnail microscopy images for all targets in the library, filtered according to a user-defined set of subcellular localizations (Figure 4A, right panel). The app itself is supported by a relational database and a REST API that also allows for programmatic access to the underlying raw data.

Comparing interactions and spatial relationships. IP-MS and microscopy examine the architecture of the cellular proteome at very different scales (molecular for IP-MS, pan-cellular for microscopy). Localization and interactions are linked: proteins must co-localize to interact. But while the localization patterns of each interactor must overlap to some extent, depending on the nature of these interactions (stable or transient), they do not need to match completely. Therefore, the degree to which interacting proteins also share the same overall spatial signature represents an organizing feature of the cellular protein network. To quantify the similarity of localization of any two targets in OpenCell, we measured the Pearson correlation between their localization encodings; this gives a distance measure of similarity in "localization space" between two proteins (Fig. 5A, upper panel). In parallel, similarities can be measured in "interaction space" by comparing the interaction profiles between any two proteins based on stoichiometry information (Fig. 5A, lower panel). As expected, the distribution of localization vs. interaction similarities between all pairs of OpenCell targets that were found to interact shows a positive correlation between the two parameters (Fig. 5B); however, it also highlights that the vast majority of interacting proteins are not particularly similar by either measure (large cloud around the origin of the graph). Examining the relationship between localization similarity and interaction stoichiometry further reveals two separate groups of interacting pairs (Fig. 5C): 1) those that interact with low stoichiometry and whose spatial signatures do not specifically overlap (i.e., have low localization similarity, solid line in Fig. 5C), which represents the largest proportion and 2) a smaller but well-delineated group of stoichiometric interactors that share very similar localization patterns (dashed line). This second result makes intuitive sense: interactions that are stoichiometric should be stable in space and time, and a high overlap between localization patterns is expected. Indeed, different subunits of known stable complexes share extremely similar localization signatures and form tight clusters in our image UMAP (Fig. 5D). Overall, our analysis underscores the organization of the cell's proteome along two modes of interactions: small communities of high-stoichiometry protein groups, whose functions are intertwined to the point that their steady-state localization patterns are very similar, and a much larger set of low-stoichiometry, and presumably more dynamic interactions with lower spatial overlap.

This intuitive result also has an important correlate: that highly similar localization patterns between two proteins can be used to infer close molecular interaction. In fact, looking at the entire set of OpenCell target pairs (predicted to interact or not), proteins that share high localization similarities are also very likely to interact (Fig 5E). 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). Therefore, our analysis demonstrates that a quantitative comparison of localization patterns can also make predictions about the molecular-level architecture of the proteome. This also reveals that the localization pattern of a given protein contains highly specific information from which precise functional attributes can be extracted, in particular by modern machine learning algorithms.

Biological discovery using interactomes or images. As demonstrated above, unsupervised clustering of both localization and interaction signatures can be used to derive functional relationships between proteins (for example, unambiguously linking together different forms of co-translational processes). A direct application of this result is to help elucidate the cellular roles of the many human proteins that remain poorly characterized (47). We first mined our interactome for poorly studied proteins. We focused on the set of baits and preys found within MCL communities, reasoning that belonging to a community was a strong signal to map function via "guilt by association". As a simple metric for how well characterized a protein is, we quantified its occurrence in article titles and abstracts from PubMed. Empirically, we determined that proteins in the bottom 10th percentile of publication count (corresponding to less than 10 publications) were very poorly annotated (Fig. 6A). This set encompasses a total of 251 proteins for which our dataset offers specific mechanistic insights. For example, the poorly characterized NHSL1, NHSL2 and KIAA1522 are all found as part of an interaction community centered around SCAR/WAVE, a large multi-subunit complex nucleating actin polymerization (Fig. 6B). Interestingly, all three proteins share sequence homology and are all homologous to NHS (Fig. S5A), a protein mutated in patients with Nance-Horan syndrome and that interacts with SCAR/WAVE components to coordinate actin remodeling (48). This suggests that NHSL1, NHSL2 and KIAA1522 also act to regulate actin assembly. A recent mechanistic study made public after our initial prediction supports this hypothesis: NHSL1 was found to localize at the cell's leading edge and to directly bind SCAR/WAVE to negatively regulate its activity, reducing F-actin content in lamellipodia and inhibiting cell migration (49). Importantly, the authors identified NHSL1's SCAR/WAVE binding sites, and we find these sequences to be conserved in NSHL2 and KIA1522 (Fig. 6B). Therefore, we propose that both NHSL2 and KIAA1522 are also direct SCAR/WAVE binders and new modulators of the actin cytoskeleton.

Our data also uncovers a specific function for ROGDI, whose variants cause Kohlschuetter-Toenz syndrome (a recessive developmental disease characterized by epilepsy and psychomotor regression (50)). ROGDI appears in the literature because of its association with disease, but no study, to our knowledge, specifically determines its molecular function. To delineate the function of ROGDI, we first observed that its interaction pattern matched very closely that of three other proteins in our dataset: DMXL1, DMXL2 and WDR7 (Fig. 6C). This set exhibited a specific interaction signature related to v-ATPase, the lysosomal proton pump. All four proteins interact with soluble v-ATPase subunits (ATP6-V1), but not its intra-membrane machinery (ATP6-V0). Interestingly, DMXL1 and WDR7 have been shown to interact with V1 v-ATPase, and their knock-down compromises lysosomal re-acidification (51). In fact, sequence analysis showed that DMXL1/2, WDR7 and ROGDI are homologous to proteins from yeast or Drosophila that have been involved in the regulation of assembly of the soluble V1 subunits onto the V0 transmembrane ATPase core (52, 53) (Fig. S5). 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 (53). To further characterize the function of these proteins, we generated new tagged cell lines for DMXL1/2, WDR7 and ROGDI. Because of the low expression level of these proteins, imaging analysis proved difficult, and fluorescent fusions of DMXL2 and ROGDI did not lead to detectable fluorescence. However, pull-downs of DMXL1 and WDR7 confirmed a stoichiometric interaction between DMXL1/2, WDR7 and ROGDI (Fig. 6C, right panels). Interestingly, no direct interaction between DXML1 and DMXL2 was detected, suggesting that they might nucleate two separate sub-complexes. Therefore, our data uncovers a human RAVE complex comprising DMXL1/2, WDR7 and ROGDI, which likely acts as a chaperone for v-ATPase assembly. Altogether, NHSL1/2-KIAA1552 and DMXL1/2-WDR7-ROGDI illustrate how OpenCell catalyzes new biological insights by combining quantitative analysis, literature curation and new functional data – including a direct mechanistic role for ROGDI, shedding light on the biology of Kohlschuetter-Toenz syndrome.

These examples underscore the power of mining interactome data for mechanistic predictions. Having established that quantitative localization encoding on its own could help elucidate the molecular function of a given protein, we next asked to which extent the function of an orphan protein could be characterized by imaging alone. FAM241A (or C4orf32) is a human protein without any functional annotation. Our initial interactome dataset placed FAM241A as part of a community centered around the OST complex, the transferase responsible for co-translational glycosylation (Fig. 6D). We generated an endogenous FAM241A fluorescent fusion and separately used imaging and mass-spectrometry to elucidate its function. Importantly, the deep learning model we used to generate its localization encoding was not trained with images of this new target (“naïve” model). Strikingly, the quantitative distances between FAM241A and other OpenCell targets measured using either localization or interaction signatures both identified FAM241A as a new OST subunit (Fig. 6D). Moreover, separate unsupervised clustering analyses using the two signatures both placed FAM241A in well-defined clusters with other OST subunits (Fig. 6D, right panels). Thus, the function of FAM241A could have been predicted with the same degree of precision by using either its interaction signature or its localization encoding. This proof-of-concept example establishes the potential of live-cell imaging as a specific readout of protein function, including for the characterization of poorly studied human proteins.

Hierarchical structure(s) of proteome organization. Finally, we explored the global structure of our datasets by using hierarchical clustering to highlight the signatures patterning the proteome. Starting with the 300 interactome communities or the 182 localization clusters outlined above, we mapped the full hierarchical relationships underlying both our interactome and imaging sets. Specifically, we implemented an agglomerative clustering strategy based on node pair sampling (54), which we applied separately to the graph of interactions between interactome communities and to the graph connecting localization clusters derived from the corresponding UMAP adjacency matrix (55). This resulted in fully connected hierarchical trees, shown in Figure 7. Isolating groups of proteins at separate hierarchical layers reveals different levels of the proteome’s organization. At an intermediate layer, the proteome can be delineated into sets of 18-19 separate “modules”, each including a median of 99 proteins for the interactome (modules M1-M18, Fig. 7A, see composition in Suppl. Table 4A) and of 66 proteins for the imaging dataset (modules N1-N19, Fig, 7B, see composition in Suppl. Table 4B), respectively. At a higher layer, each dataset can be divided into three “branches”, which separately represent the core features that shape the proteome’s global architecture from a molecular or spatial perspective. Performing gene ontology analysis underlined that modules and branches are enriched for specific cellular functions or compartments, which define unique functional signatures (labeled in Fig 7A,7B – details in Figure S6 for branches; Suppl. Tables 4A, 4B contain the full gene ontology analysis for both modules and branches).

Overall, the hierarchy of localization encodings reveals that, as expected, localization patterns are organized along the three foundational compartments of the eukaryotic cell: nucleus, cytoplasm and membrane-bound organelles (Fig. 7B, Fig. S6G). Each localization branch is sub-divided into modules that correspond to discrete sub-cellular territories. For example, the organellar branch separates into the different components of the secretory pathway: ER, Golgi, endosome, lysosome or plasma membrane, mirroring the known spatial segregation between these compartments. The fact that this unsupervised clustering strategy broadly recapitulates known cellular compartments validates our approach. By contrast, the hierarchical analysis of the interactome graph reveals how the proteome is organized at the molecular scale (Fig. 7A). The 18 modules separate the interactome into clear cellular functions such as transcription, splicing or vesicular transport. This reflects that functional pathways are templated by groups of proteins that physically interact, a principle also underscored by the overlap between genetic interactions and protein complexes in eukaryotes (56, 57). More interestingly, analysis of the high-level branches reveals a separation between three groups of proteins that differ in their functional and biophysical properties. Ontology enrichment highlights that proteins related in membrane processes (branch B) and RNA-binding (branch C) clearly segregate from the rest of the proteome in term of their interaction profiles (Fig. S6A-E). This functional segregation is correlated with different biophysical properties: branch B proteins are significantly more structured, more hydrophobic and richer in aromatic residues than the rest of our dataset; conversely, branch C proteins have higher intrinsic disorder and higher isoelectric points (Fig. S6B-C). Overall, our data reveals that RNA-binding proteins form a specific molecular sub-group that shapes the global organization of the cell, similar to how association with membranes is a molecular feature that sets apart a large fraction of the proteome. Strikingly, RNA-binding proteins are known to form membrane-less organelles under stress conditions, especially through phase transition processes supported by intrinsic disorder (33, 34). This suggests that the biophysical properties underlying the formation of biomolecular condensates might also be a global driving force shaping the cellular proteome network under normal conditions.

Discussion

OpenCell combines innovations at four separate levels to augment the elucidation of human cellular architecture. First, we describe 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 pioneer new analytical strategies for the representation and comparison of interaction or localization signatures (including a fully self-supervised machine learning approach for image encoding). And fourth, we demonstrate how our dataset can be used both for fine-grained mechanistic exploration (by elucidating the function of multiple proteins that were previously uncharacterized), as well as for investigating the core organizational principles that wire the proteome. In particular, we uncover two global features that shape cellular architecture. First, we show that most proteins interact with low stoichiometry and distribute unequally within the cell, whereas high-stoichiometry interactors share very similar localization patterns. 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 together (7) but also connecting different cellular compartments. Second, we reveal that two separate groups of interacting proteins segregate from the global proteome: membrane-related and RNA-binding, both of which exhibit specific biophysical signatures (in particular hydrophobicity and high intrinsic disorder, respectively). That membrane-related proteins form a specific interaction group is perhaps not surprising as the two-dimensional membrane surface drives their sequestration within the three-dimensional cell. By contrast, the discovery of RNA-binding proteins as a separate sub-group is significant given the growing appreciation that RNA-binding proteins, together with RNAs themselves, can form condensates in the cytoplasm and nucleoplasm. Interestingly, intrinsic disorder is an important modulator of partition into biomolecular condensates, as are specific protein-protein interaction domains (33, 34, 58, 59). Therefore, RNA-binding proteins might have evolved to multiply molecular interactions between themselves to create compartments that can concentrate a large functional variety of proteins (for example RNA processing factors, translation machinery or silencing complexes (60, 61)) and enable the spatial specialization of cellular processes. In this context, it is interesting to consider a role for RNA itself as an organizer of the cellular proteome, as is for example the case for some non-coding RNAs whose function is to template molecular interactions with proteins to form nuclear bodies (62).

While OpenCell joins a growing number of large-scale datasets (6, 8, 9, 16, 27, 63–66) that contribute to the systematic and quantitative dissection of the human cell, it is unique in its analysis of both quantitative interactomes and live-cell localization of genome-edited proteins. However, the full description of human cellular architecture remains a formidable challenge, especially considering the vast diversity of cell types and cell states that shape human physiology. Mirroring the advances in genomics following the sequencing the human genome (2), open-source systematic datasets will likely play an important role in how the growth of cell biology measurements can be transformed into fundamental discoveries by an entire community (67). To date, OpenCell includes functional information for 1311 targets (~7% of the human proteome) and a total of 5148 proteins found in our interactome (baits and preys, ~26% of the proteome). Our approach that combines split-FP systems and HEK293T – a cell line that is heavily transformed but easily manipulatable – is mostly constrained by scalability considerations. But given the current pace of technological advances, the large scale of measurements required to match the full extent of cellular complexity might soon be within reach. In particular, advances in stem cell technologies enable the generation of libraries that can be differentiated in multiple cell types (16), while innovations in genome engineering (for example, by modulating DNA repair (68)) pave the way for the scalable insertion of gene-sized payload (e.g., fluorescent proteins, HaloTag, degrons), for the combination of multiple edits in the same cell (e.g., dual-tagged libraries for co-localization studies) or for increased homozygosity in polyclonal pools. In addition, our live-cell imaging approach also paves the way for the systematic description of 4D intracellular dynamics (64), which is being transformed by recent developments in high-throughput light-sheet microscopy (69).

Finally, OpenCell provides a large set of open-source, quantitative and curated data that can be used as a proving ground for data science and algorithm development. Our own innovation in machine learning to encode localization signatures was made possible by the availability of a critical mass of high-quality live-cell images taken under uniform experimental conditions – itself facilitated by our collection of genome-edited lines that can be characterized repeatedly and in a native cellular context. Our results also demonstrate the power of self-supervised deep learning models to identify complex but deterministic signatures from light microscopy images. In particular, we show that remarkably detailed functional relationships can be inferred on the sole basis of similarities between localization patterns, including the prediction of molecular interactions. This opens exciting avenues for the use of imaging as a high-throughput, information-rich method for deep phenotyping and functional genomics (70). Because light microscopy is easily scalable, can be performed in live, unperturbed samples and enables measurements at the single-cell level, this offers rich opportunities for the full quantitative description of cellular diversity in normal physiology and disease.

*

* *

Acknowledgements. 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. 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 pre-doctoral 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.

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.

Materials and Methods are available as a supplementary file.

References

1. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

2. Hood, L. & Rowen, L. The Human Genome Project: big science transforms biology and medicine. *Genome Med* **5**, 79 (2013).

3. Nurse, P. & Hayles, J. The Cell in an Era of Systems Biology. *Cell* **144**, 850–854 (2011).

4. Mast, F. D., Ratushny, A. V. & Aitchison, J. D. Systems cell biologySystems cell biology. *The Journal of Cell Biology* **206**, 695–706 (2014).

5. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology* **20**, 285–302 (2019).

6. Luck, K. *et al*. A reference map of the human binary protein interactome. *Nature* **580**, 1–7 (2020).

7. Hein, M. Y. *et al*. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).

8. Huttlin, E. L. *et al*. Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, (2017).

9. Thul, P. J. *et al*. A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

10. Mering, C. von *et al*. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).

11. Ghaemmaghami, S. *et al*. Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).

12. Collins, S. R. *et al*. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP* **6**, 439–450 (2007).

13. Weill, U. *et al*. Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nature Methods* **15**, (2018).

14. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic acids research* **21**, 3329–3330 (1993).

15. Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3501-8 (2016).

16. Roberts, B. *et al*. Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Molecular biology of the cell* mbc.E17-03-0209 (2017) doi:10.1091/mbc.e17-03-0209.

17. Hubner, N. C. *et al*. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *The Journal of cell biology* **189**, 739–754 (2010).

18. Feng, S. *et al*. Improved split fluorescent proteins for endogenous protein labeling. *Nature communications* **8**, 370 (2017).

19. Meier, F. *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).

20. Lin, Y.-C. *et al*. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nature communications* **5**, 4767 (2014).

21. Lin, S., Staahl, B., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, (2014).

22. Doyon, J. B. *et al*. Rapid and efficient clathrin-mediated endocytosis revealed in genome-edited mammalian cells. *Nature cell biology* **13**, 331–337 (2011).

23. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nat Methods* **10**, 715–721 (2013).

24. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol Cell Proteom Mcp* **14**, 120–35 (2014).

25. Thomas, J. A. & Tate, C. G. Quality Control in Eukaryotic Membrane Protein Overproduction. *J Mol Biol* **426**, 4139–4154 (2014).

26. Giurgiu, M. *et al*. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res* **47**, gky973- (2018).

27. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, 570 (2016).

28. Huttlin, E. L. *et al*. Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. *Biorxiv* 2020.01.19.905109 (2020) doi:10.1101/2020.01.19.905109.

29. Royer, L., Reimann, M., Stewart, A. F. & Schroeder, M. Network Compression as a Quality Measure for Protein Interaction Networks. *PLoS ONE* **7**, e35729 (2012).

30. Albert, R. Scale-free networks in cell biology. *Journal of Cell Science* **118**, 4947–4957 (2005).

31. Tanaka, R., Yi, T.-M. & Doyle, J. Some protein interaction data do not exhibit power law statistics. *Febs Lett* **579**, 5140–5144 (2005).

32. Haynes, C. *et al*. Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. *Plos Comput Biol* **2**, e100 (2006).

33. Alberti, S. & Dormann, D. Liquid–Liquid Phase Separation in Disease. *Annu Rev Genet* **53**, 1–24 (2019).

34. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).

35. Xue, B. *et al*. Structural Disorder in Viral Proteins. *Chem Rev* **114**, 6880–6911 (2014).

36. Enright, A. J., Dongen, S. V. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584 (2002).

37. Shurtleff, M. J. *et al*. The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *Elife* **7**, e37018 (2018).

38. Acosta-Alvear, D. *et al*. The unfolded protein response and endoplasmic reticulum protein targeting machineries converge on the stress sensor IRE1. *Elife* **7**, e43036 (2018).

39. McGilvray, P. T. *et al*. An ER translocon for multi-pass membrane protein biogenesis. *Elife* **9**, e56889 (2020).

40. Chitwood, P. J. & Hegde, R. S. An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature* **584**, 630–634 (2020).

41. Görlich, D. & Kutay, U. Transport between the cell nucleus and the cytoplasm. *Annu Rev Cell Dev Bi* **15**, 607–660 (1999).

42. Lusk, C. P. & King, M. C. The nucleus: keeping it together by keeping it apart. *Curr Opin Cell Biol* **44**, 44–50 (2017).

43. Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. & Olivo-Marin, J.-C. Imagining the future of bioimage analysis. *Nat Biotechnol* **34**, 1250–1255 (2016).

44. Kobayashi, H., Cheveralls, K. C., Leonetti, M. D. & Royer, L. A. Self-Supervised Deep-Learning Encodes High-Resolution Features of Protein Subcellular Localization. *bioRxiv*.

45. Ouyang, W. *et al*. Analysis of the Human Protein Atlas Image Classification competition. *Nat Methods* **16**, 1254–1261 (2019).

46. Traag, V. A., Waltman, L. & Eck, N. J. van. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep-uk* **9**, 5233 (2019).

47. Stoeger, T., Gerlach, M., Morimoto, R. I. & Amaral, L. A. N. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS biology* **16**, e2006643 (2018).

48. Brooks, S. P. *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).

49. Law, A.-L. *et al*. Nance-Horan Syndrome-like 1 protein negatively regulates Scar/WAVE-Arp2/3 activity and inhibits lamellipodia stability and cell migration. *Biorxiv* 2020.05.11.083030 (2020) doi:10.1101/2020.05.11.083030.

50. Schossig, A. *et al*. Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. *Am J Hum Genetics* **90**, 701–707 (2012).

51. Merkulova, M. *et al*. Mapping the H+ (V)-ATPase interactome: identification of proteins involved in trafficking, folding, assembly and phosphorylation. *Scientific Reports* **5**, (2015).

52. Yan, Y., Denef, N. & Schüpbach, T. The Vacuolar Proton Pump, V-ATPase, Is Required for Notch

Signaling and Endosomal Trafficking in Drosophila. *Dev Cell* **17**, 387–402 (2009).

53. Vasanthakumar, T. & Rubinstein, J. L. Structure and Roles of V-type ATPases. *Trends Biochem Sci* **45**, 295–307 (2020).

54. Bonald, T., Charpentier, B., Galland, A. & Hollocou, A. Hierarchical Graph Clustering using Node Pair Sampling. *Arxiv* (2018).

55. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv* (2018).

56. Costanzo, M. *et al*. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).

57. Horlbeck, M. A. *et al*. Mapping the Genetic Landscape of Human Cells. *Cell* **174**, 953-967.e22 (2018).

58. Sanders, D. W. *et al*. Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. *Cell* **181**, 306-324.e28 (2020).

59. Yang, P. *et al*. G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**, 325-345.e28 (2020).

60. Markmiller, S. *et al*. Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604.e13 (2018).

61. Youn, J.-Y. *et al*. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532.e11 (2018).

62. Chujo, T. & Hirose, T. Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function. *Mol Cells* (2017) doi:10.14348/molcells.2017.0263.

63. Go, C. D. *et al*. A proximity biotinylation map of a human cell. *Biorxiv* 796391 (2019) doi:10.1101/796391.

64. Cai, Y. *et al*. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature* **561**, 411–415 (2018).

65. Gut, G., Herrmann, M. D. & Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science* **361**, eaar7042 (2018).

66. Hutchins, J. R. A. *et al*. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science (New York, N.Y.)* **328**, 593–599 (2010).

67. Ellenberg, J. *et al*. A call for public archives for biological image data. *Nature Methods* **15**, 849–854 (2018).

68. Riesenberg, S. *et al*. Simultaneous precise editing of multiple genes in human cells. *Nucleic acids research* **2**, 163 (2019).

69. Yang, B. *et al*. Epi-illumination SPIM for volumetric imaging with high spatial-temporal resolution. *Nature methods* **16**, 501–504 (2019).

70. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat Rev Drug Discov* **20**, 145–159 (2021).

71. Drew, K. *et al*. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology* **13**, 932 (2017).

72. Young, C. L., Britton, Z. T. & Robinson, A. S. Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnol J* **7**, 620–634 (2012).

73. Kamiyama, D. *et al*. Versatile protein tagging in cells with split fluorescent protein. *Nature communications* **7**, 11046 (2016).

74. Dingle, G. CrispyCrunch: High-throughput Design and Analysis of CRISPR+HDR Experiments. *undefined* https://blog.addgene.org/crispycrunch-high-throughput-design-and-analysis-of-crisprhdr-experiments.

75. Jinek, M. *et al*. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816–821 (2012).

76. Bache, N. *et al*. A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics*. *Mol Cell Proteomics* **17**, 2284–2296 (2018).

77. Meier, F. *et al*. Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer*. *Mol Cell Proteomics* **17**, i–2545 (2018).

78. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range

mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–1372 (2008).

79. Prianichnikov, N. *et al*. MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics*. *Mol Cell Proteomics* **19**, 1058–1069 (2020).

80. Vizcaíno, J. A. *et al*. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**, 223–226 (2014).

81. Cock, P. J. A. *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

82. Mészáros, B., Erdős, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, gky384- (2018).

83. Razavi, A., Oord, A. van den & Vinyals, O. Generating Diverse High-Fidelity Images with VQ-VAE-2. *Arxiv* (2019).

84. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

85. Mi, H. *et al*. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res* **49**, gkaa1106- (2020).

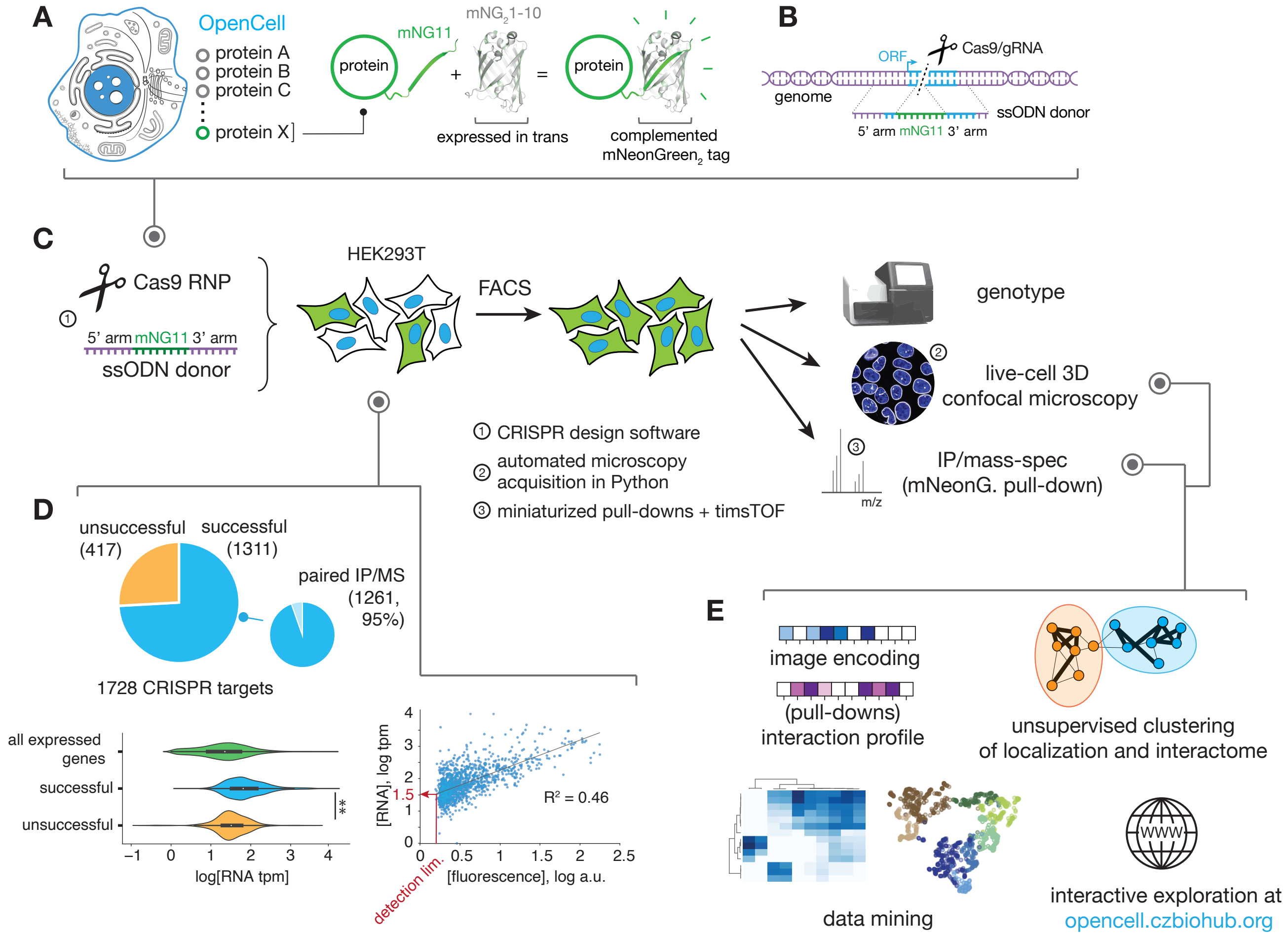


Figure 1: the OpenCell library. (A) Functional tagging with split-mNeonGreen2. **(B)** Endogenous tagging strategy: mNG11 fusion sequences are inserted directly within genomic open reading frames (ORFs) using CRISPR-Cas9 gene editing and homologous recombination with single-stranded oligonucleotides (ssODN) donors. **(C)** The OpenCell experimental pipeline. See text for details. **(D)** Successful detection of fluorescence in the OpenCell library. Out of 1728 genes that were originally targeted, fluorescent signal was successfully detected for 1311 (left panel). Low expression level is the main obstacle to successful detection (bottom left panel, showing the full distribution of RNA expression in transcripts per million reads – tpm – for all genes expressed in HEK293T vs. successfully or unsuccessfully detected OpenCell targets; **: p < 10-5, t-test). A correlation between expression level (RNASeq tpm) and fluorescence intensity (as read by flow cytometry) is shown for all successful OpenCell targets (bottom right panel). A linear regression (solid line, Pearson R2 = 0.49) allows to estimate the expression threshold required for successful detection. **(E)** The OpenCell data analysis pipeline, described in subsequent sections.

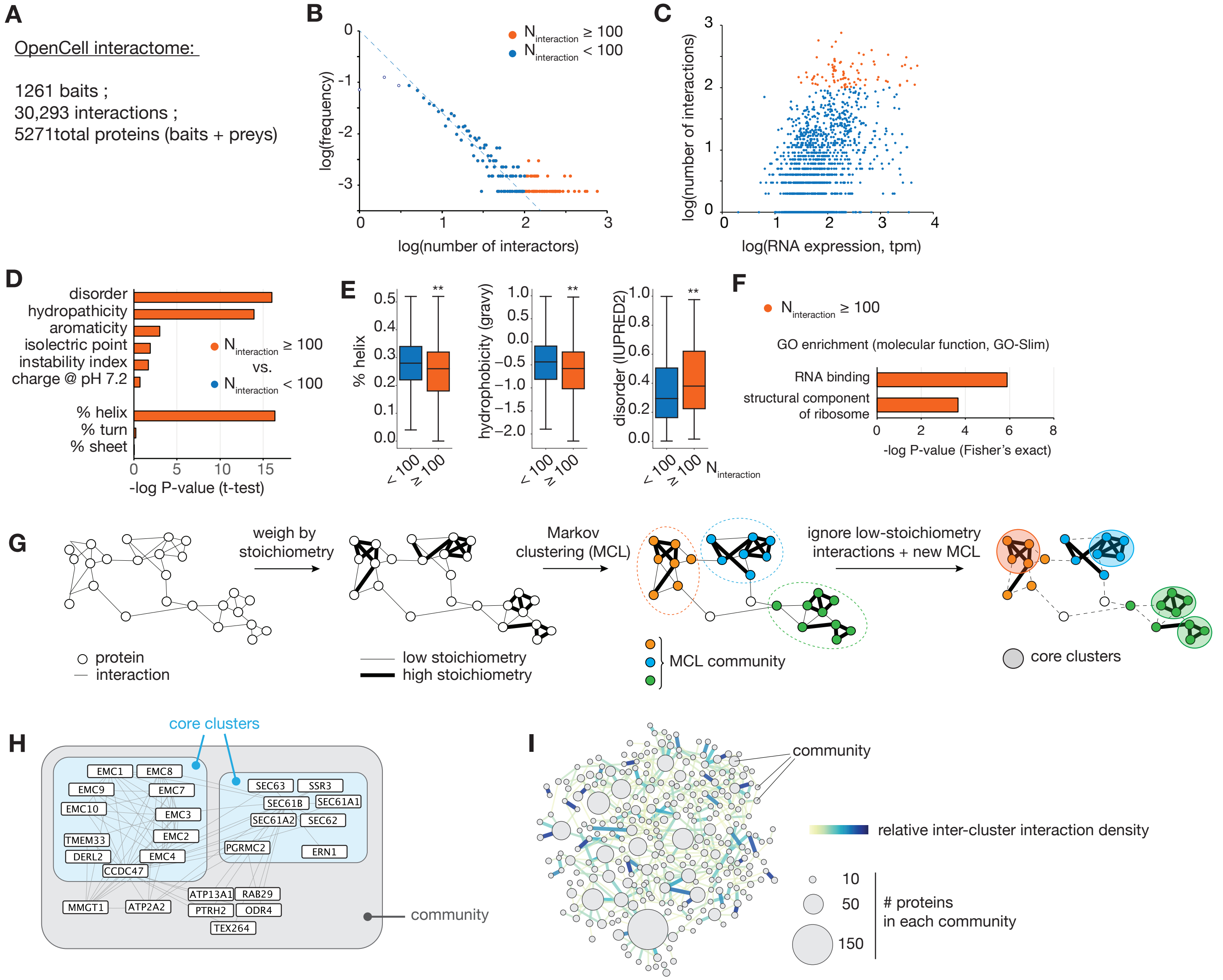


Figure 2: the OpenCell interactome. (A) Overall description of the interactome. **(B)** Frequency distribution of the number of interactors per targets on a log-log scale; the dotted line represents the linear fit for moderate interaction numbers ($4<N_{interaction}<100$). Highly interacting targets (>=100 interactors) are highlighted in orange. **(C)** Correlation between number of interactions and expression level (RNASeq tpm) for each target. High numbers of interactions are not restricted to highly expressed proteins. **(D)** Comparing biophysical properties of highly interacting (>=100 interactions) vs. other targets (<100 interactions). This analysis was performed by first breaking down the amino acid sequence of each target into 100 a.a. windows. Shown are t-tests comparing the set of 100-a.a. windows from both groups. **(E)** Full distribution of % helical 2^{ndary} structure, hydrophobicity or intrinsic disorder of the set of 100-a.a. windows from highly interacting vs. other targets. Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values. ** $p < 10^{-10}$ (t-test) **(F)** Gene ontology (GO, molecular function) enrichment analysis in highly interacting vs. other targets. **(G)** Unsupervised Markov clustering of the interactome graph. **(H)** Example of community and core cluster definition for the translocon/EMC community. **(I)** The complete graph of connections between interactome communities. The density of protein-protein interactions between communities is color-coded. The numbers of targets included in each community is represented by circles of increasing diameter.

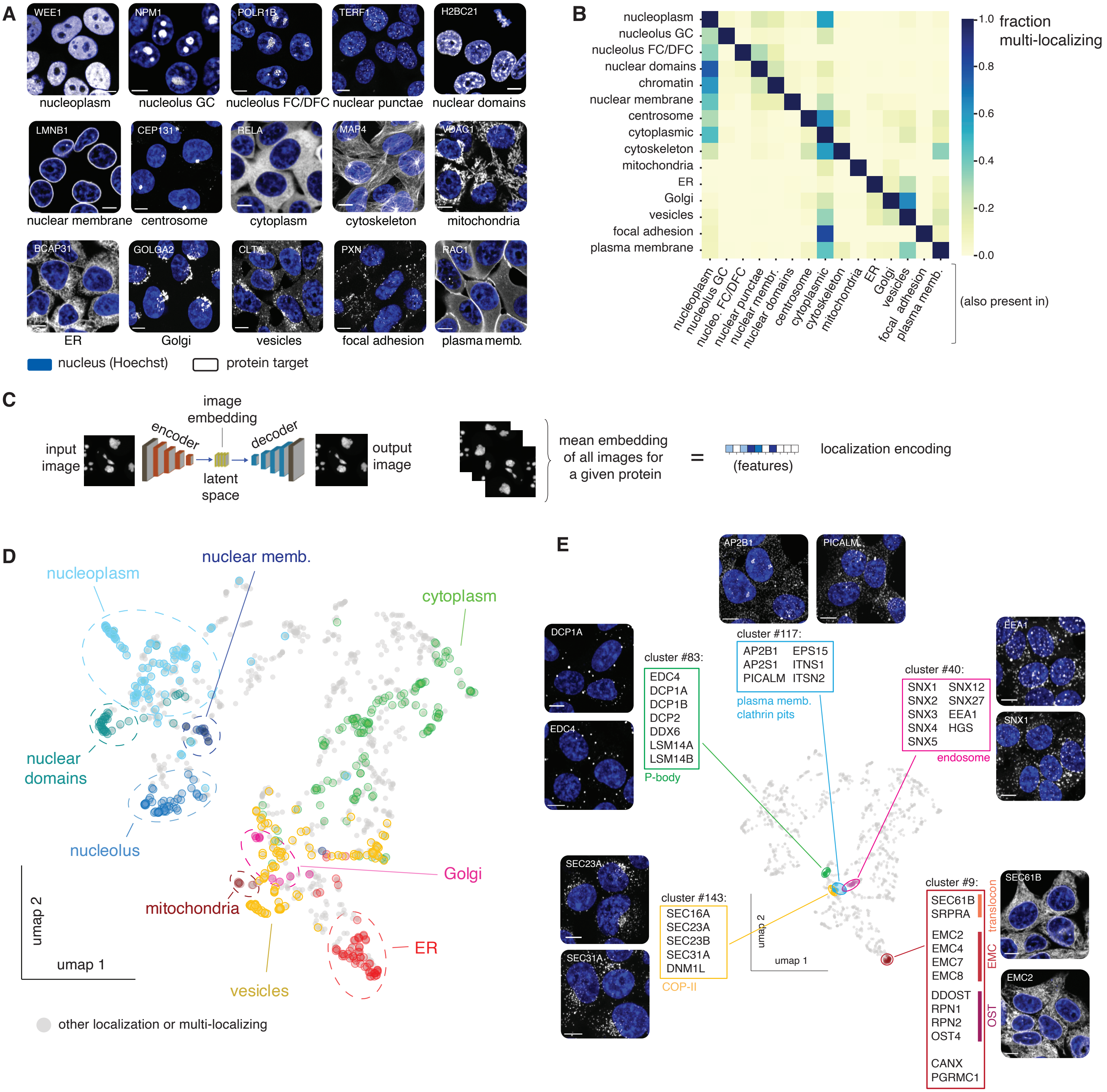
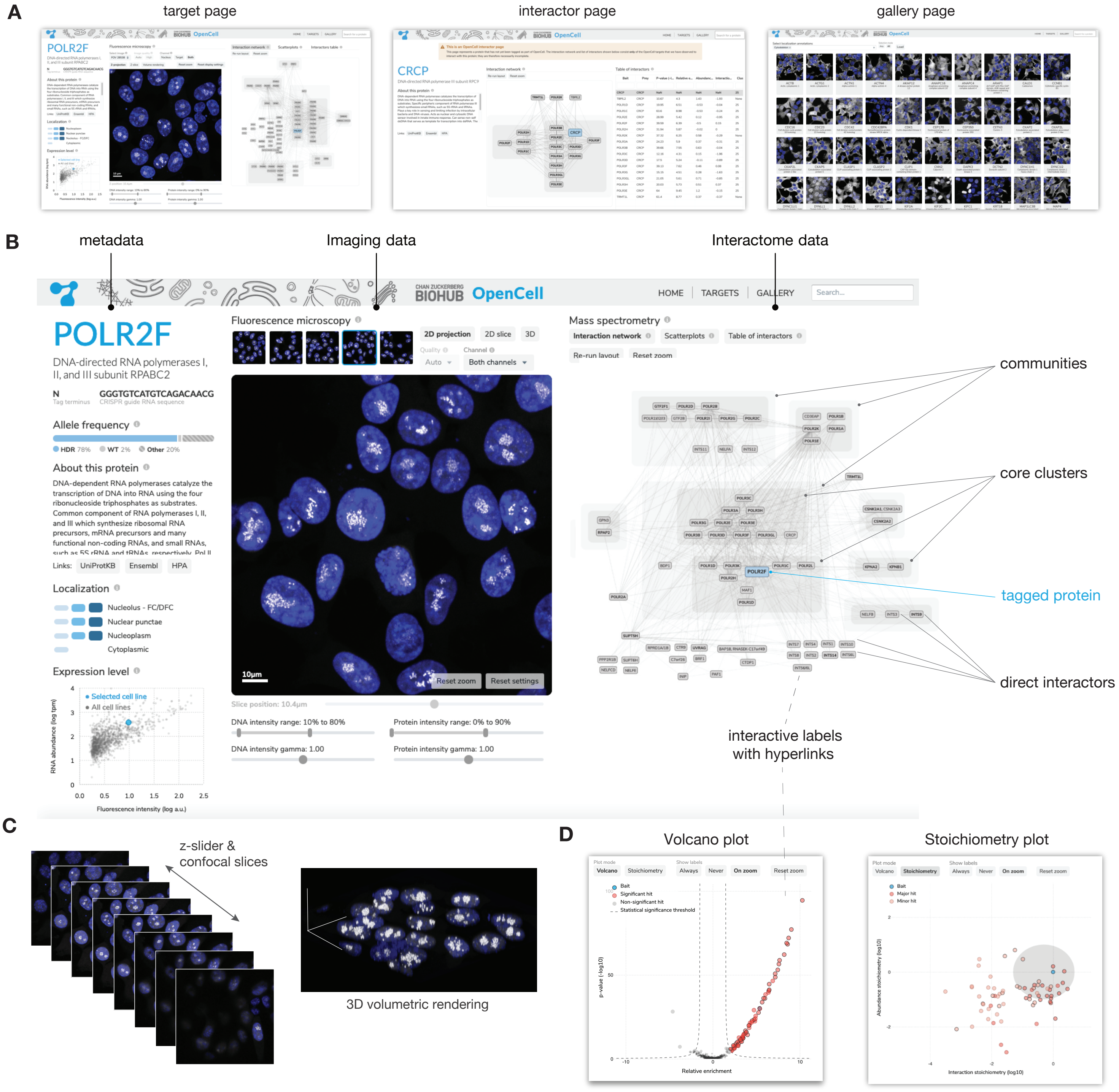


Figure 3: the OpenCell image collection. (A) The 15 cellular compartments segregated for annotating localization patterns. The localization of a representative protein belonging to each group is shown (greyscale, gene names in top left corners; scalebar: 10 µm). Nuclear stain (Hoechst) is shown in blue. **(B)** Fraction of multi-localization between cellular compartments. **(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)** Examples of clusters delineated by unsupervised Leiden clustering of the localization dataset, highlighted on the localization UMAP.



POLR2F

DNA-directed RNA polymerases I, II, and III subunit RPABC2

N GGGTGTCATGTCAGACAACG
Tag terminus CRISPR guide RNA sequence

Allele frequency

● HDR 78% ● WT 2% ● Other 20%

About this protein

DNA-dependent RNA polymerases catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. Common component of RNA polymerases I, II, and III which synthesize ribosomal RNA precursors, mRNA precursors and many functional non-coding RNAs, and small RNAs, such as 5S rRNA and tRNAs, respectively. Pol II

Links: UniProtKB Ensembl HPA

Localization

Nucleolus - FC/DFC
Nuclear punctae
Nucleoplasm
Cytoplasmic

Expression level

● Selected cell line
● All cell lines

RNA abundance (log tpm)
Fluorescence intensity (log a.u.)

Fluorescence microscopy

2D projection 2D slice 3D

Quality: Auto Channel: Both channels

10μm

Reset zoom Reset settings

Slice position: 10.4μm

DNA intensity range: 10% to 80%

Protein intensity range: 0% to 90%

DNA intensity gamma: 1.00

Protein intensity gamma: 1.00

Mass spectrometry

Interaction network Scatterplots Table of interactors

Re-run layout Reset zoom

Plot mode: Volcano Stoichiometry Show labels: Always Never On zoom Reset zoom

● Bait
● Significant hit
● Non-significant hit
Statistical significance threshold

p-value (-log10)
Relative enrichment

● Bait
● Major hit
● Minor hit

Abundance stoichiometry (log10)
Interaction stoichiometry (log10)

Figure 4: interactive data exploration at opencell.czbiohub.org. (A) The three principal pages of the OpenCell web app. From left to right: the target page, interactor page, and gallery page. **(B)** The target page consists of three columns. The leftmost column contains the functional annotation for the target from UniProt, links to other databases, our manually-assigned localization annotations, and measures of protein expression. The middle column contains the image viewer, and the rightmost column the interaction network. **(C)** The image viewer allows the user to scroll through the confocal z-slices using a slider or to visualize the z-stack in 3D as a volume rendering; in either mode, the user can pan and zoom by clicking, dragging, and scrolling. **(D)** The interaction network can be toggled with two alternative, complementary visualizations of the target's protein interactions: a volcano plot of relative enrichment vs. p-value and a scatterplot of interaction stoichiometry vs. abundance stoichiometry. In both the network view and the scatterplots, the user can click on an interactor to open the target or the interactor page for the corresponding protein.

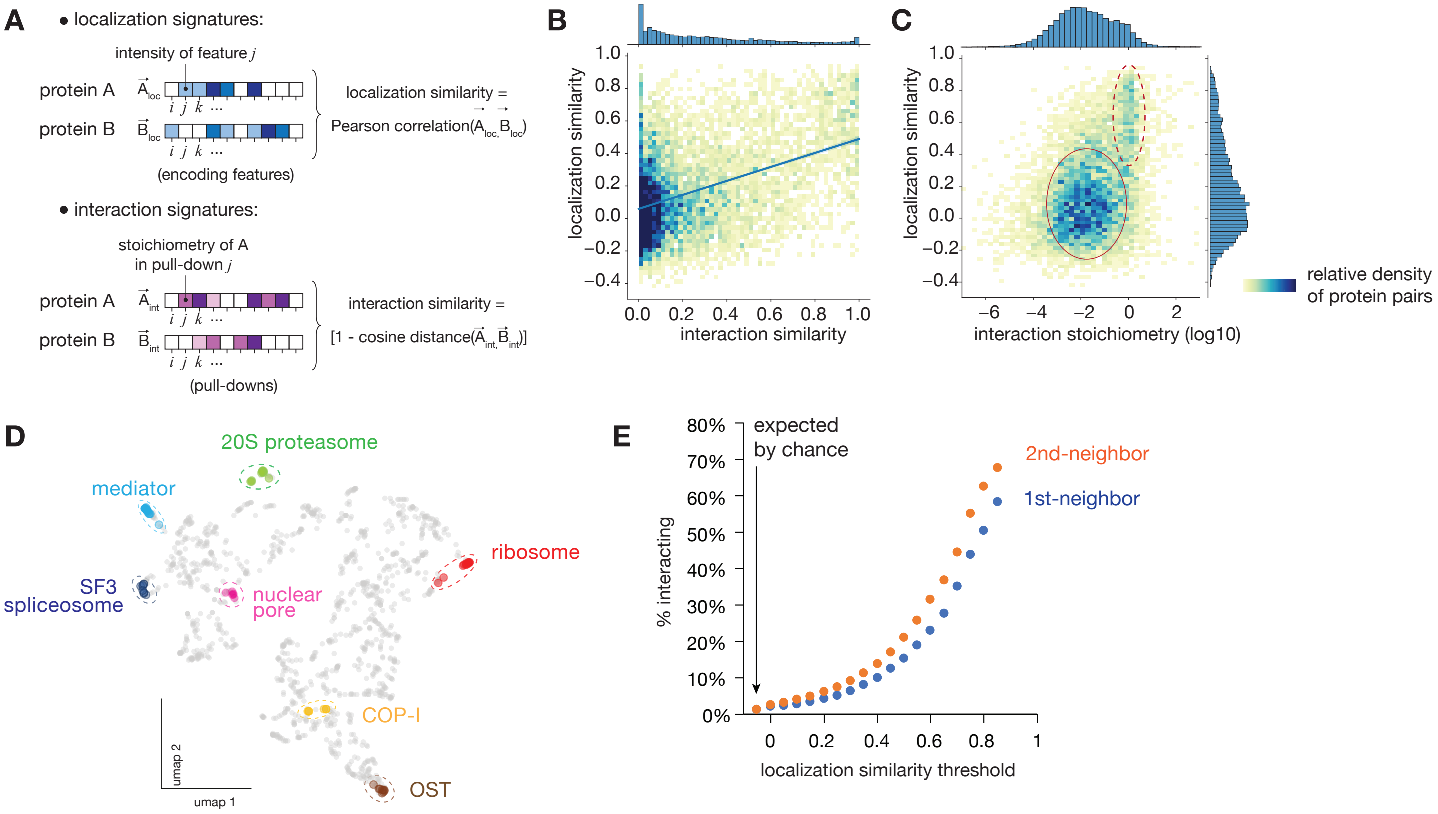


Figure 5: quantitative comparisons of localization and interaction signatures. (A) Measure of localization (top) and interaction (bottom) similarities between proteins. **(B)** Heatmap distribution of localization vs. interaction similarities between all interacting pairs of OpenCell targets. **(C)** Heatmap distribution of localization vs. interaction stoichiometry between all interacting pairs of OpenCell targets. The discrete sub-group of high-stoichiometry/high localization similarity pairs is outlined. **(D)** Localization patterns of different subunits from example stable protein complexes, represented on the localization UMAP (cf. Figure 3). **(E)** Frequency of direct (1st-neighbor) or once-removed (2nd-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).

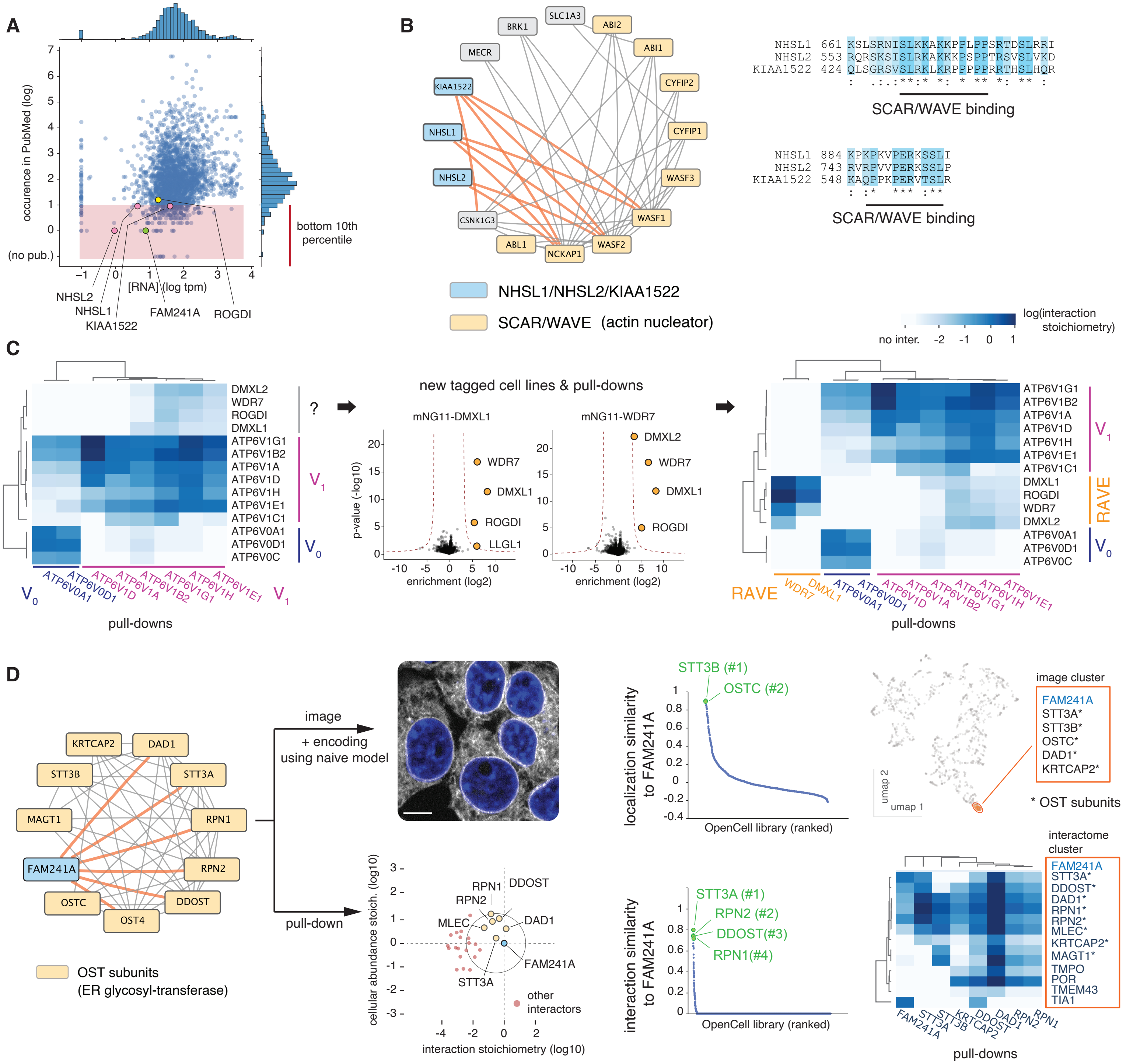


Figure 6: Biological discovery with OpenCell. (A) Distribution of occurrence in PubMed articles vs. RNA expression for all proteins found within interactome communities. The bottom 10th percentile of publication count (poorly characterized proteins) is highlighted. **(B)** NHSL1/NSHL2/KIAA1522 are part of the SCAR/WAVE community and share amino-acid sequence homology (right panel). **(C)** 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. **(D)** Parallel identification of FAM241A as a new OST subunit by imaging or mass-spectrometry. See text for details.

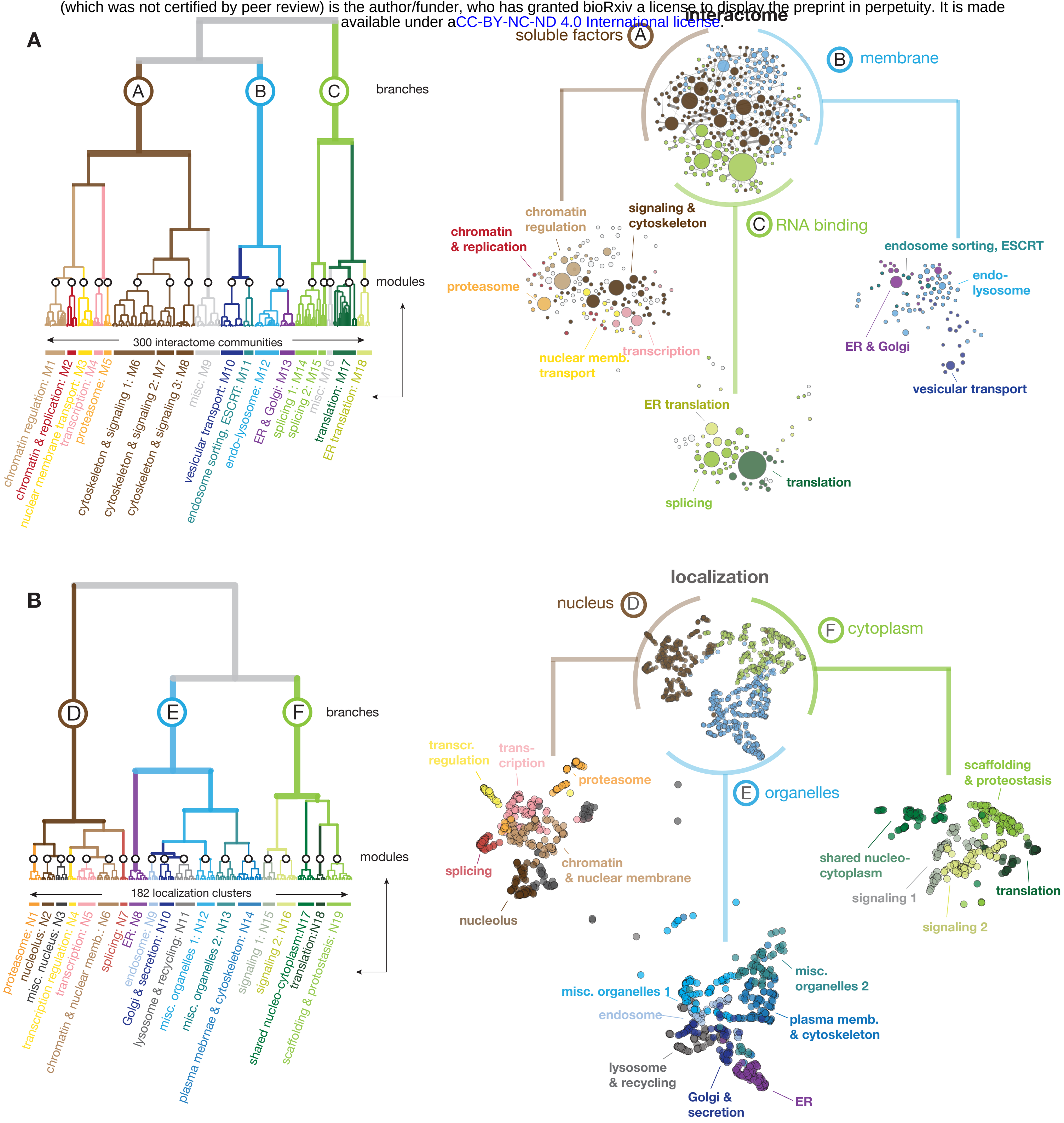


Figure 7: Full hierarchical structure of interactome and localization datasets. Dendrograms represent the hierarchical relationships connecting **(A)** the full set of protein communities identified in the OpenCell interactome (see Fig. 2) or **(B)** the full set of localization clusters identified in the OpenCell image collection (see Fig. 3). For each dataset, an intermediate layer of hierarchy separates 16-18 separate modules, while an upper hierarchical layer delineates three separate branches. Modules and branches are annotated on the basis of gene ontology enrichment analysis. Right-hand panels present the topological arrangement of branches (top) and modules (bottoms) in each dataset, highlighted from the the full graph of connections between interaction communities ("interactome", see Fig. 2I) or from the OpenCell localization UMAP ("localization", see Fig. 3D). The color codes between interactome and localization datasets are not directly comparable (i.e. same colors are not meant to represent the same set of proteins).

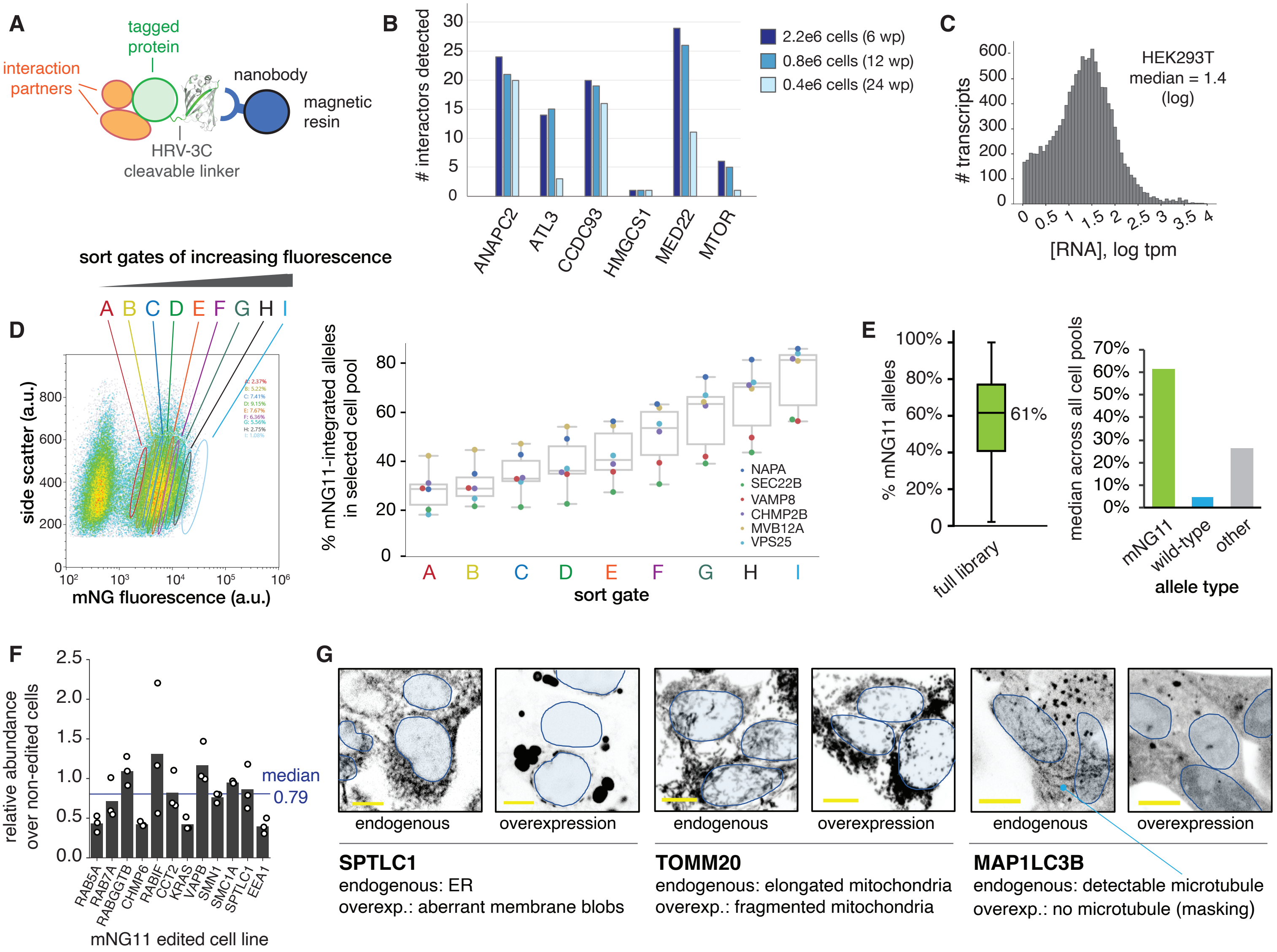
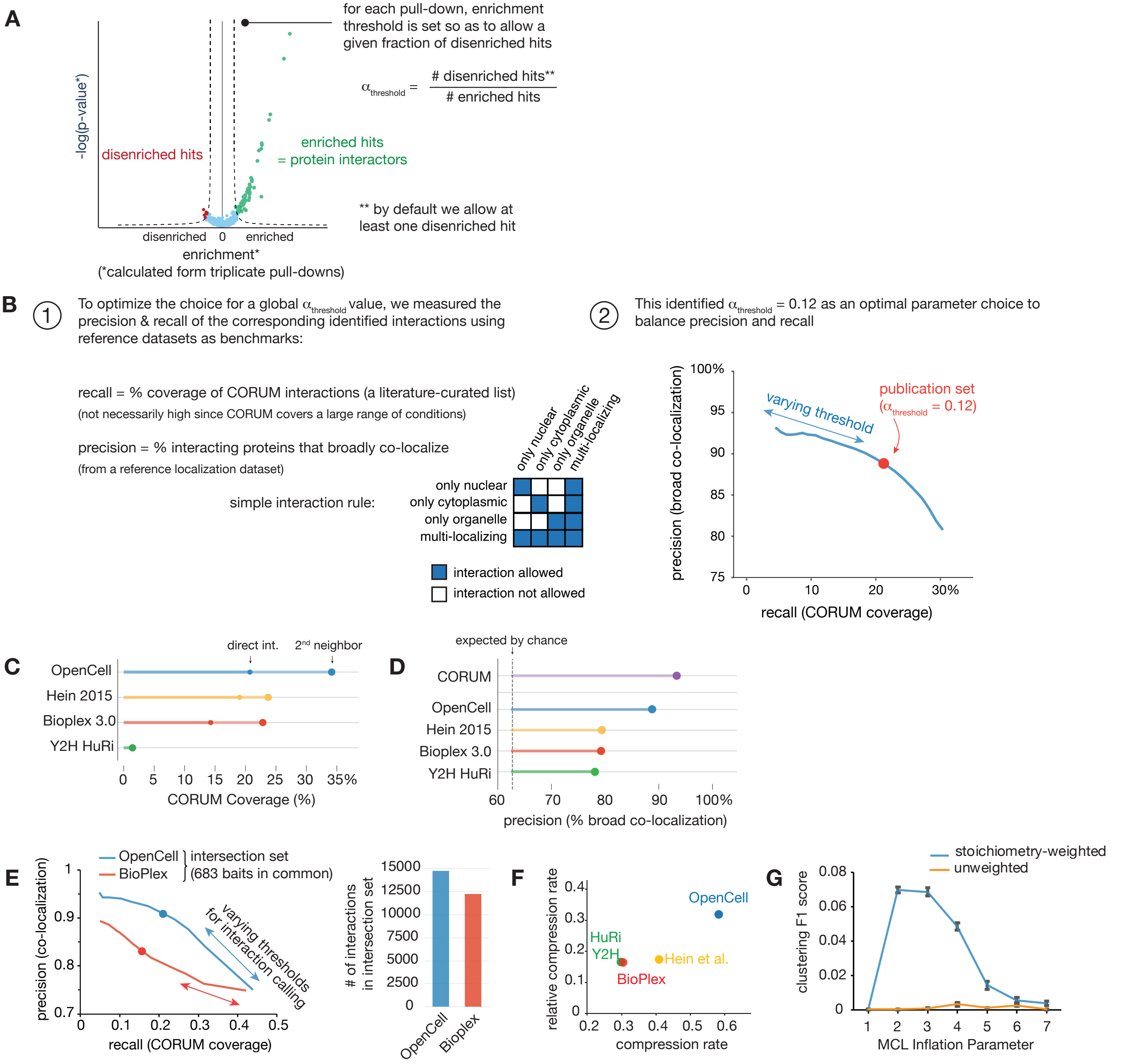


Figure S1: experimental pipeline (related to Figure 1). (A) IP/MS using FP capture. All mNG11 tagging constructs also include an HRV-3C cleavable linker for optional release from the capture resin. **(B)** Sensitivity of interaction proteomics detection on a timsTOF instrument. The number of interactors detected in pull-downs from 6 different targets is shown, varying the amount of input material. To balance sensitivity and scalability, 0.8e6 cells were used for high-throughput assays (12 well-plate, wp). **(C)** Distribution of transcript abundance in HEK293T. **(D)** Optimization of sorting strategy. Polyclonal cell pools were sorted using gates of increasing fluorescence (left panel) and genotyped to quantify the enrichment for mNG11-inserted alleles (right panel, showing data for 6 different target genes). This informed our final sorting strategy in which the top 1% of fluorescent cells (gate I) were selected. **(E)** Genotype analysis of the polyclonal OpenCell library. A single allele is required for fluorescence, but our cell collection is enriched for homozygous insertions. In total, mNG11 insertions account for 61% (median) of alleles in a given cell pool across the full library (Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values). The median values of mNG11 integrated alleles, wt alleles and other alleles are shown on the right. **(F)** Measurement of total target protein abundance by quantitative Western blotting in final selected cell pools vs. parental cell line. **(G)** Examples of overexpression artifacts. Single z-slice confocal images are shown (scale bar: 10 μm). Endogenous and overexpression cells were not imaged using the same laser power, so that signal intensities are not directly comparable. Nuclei are shown as blue outlines (nuclei can be located in a different z-plane than the one shown).



for each pull-down, enrichment threshold is set so as to allow a given fraction of disenriched hits

$$\alpha_{\text{threshold}} = \frac{\text{\# disenriched hits**}}{\text{\# enriched hits}}$$

** by default we allow at least one disenriched hit

① To optimize the choice for a global $\alpha_{\text{threshold}}$ value, we measured the precision & recall of the corresponding identified interactions using reference datasets as benchmarks:

recall = % coverage of CORUM interactions (a literature-curated list)
(not necessarily high since CORUM covers a large range of conditions)

precision = % interacting proteins that broadly co-localize
(from a reference localization dataset)

simple interaction rule:

	only nuclear	only cytoplasmic	only organelle	multi-localizing
only nuclear	■	□	□	■
only cytoplasmic	□	■	□	■
only organelle	□	□	■	■
multi-localizing	■	■	■	■

■ interaction allowed
□ interaction not allowed

② This identified $\alpha_{\text{threshold}} = 0.12$ as an optimal parameter choice to balance precision and recall

Figure S2: interactome analysis (related to Figure 2). (A) Strategy for defining enrichment threshold to define interactions. Our strategy builds upon methods described by Hein et al (7). Here we use a quantitative approach to define enrichment thresholds dynamically for each replicate set, globally constrained by the parameter $\alpha_{\text{threshold}}$. **(B)** To optimize parameter choice, we measured how precision (% co-localization) and recall (% CORUM coverage) of the corresponding interaction network varied with $\alpha_{\text{threshold}}$. This informed a final value of 0.12. **(C)** Comparing interaction recall (% CORUM coverage) of OpenCell vs. other large-scale interactomes, including direct or 2nd-neighbor interactions (i.e., sharing a direct interactor in common). **(D)** Comparing interaction precision (% co-localization) of OpenCell vs. other large-scale interactomes. CORUM interactions are shown as a reference. **(E)** Direct comparison of OpenCell vs. Bioplex 3.0. Both datasets use the same HEK293T cell line and share a large number (683) of baits in common. Precision and recall analysis by varying threshold for interaction detection is shown for the intersection set of 683 baits (dots represent values using thresholds used for final publication sets in both studies). For these set of overlapping baits, OpenCell also includes more measured interactions (right panel). **(F)** Compressibility analysis (29) of OpenCell vs. other large-scale interactomes. **(G)** MCL clustering performance (F1 score) using stoichiometry-weighted or unweighted interaction graphs, derived from CORUM interactions as described in Drew et al (71).

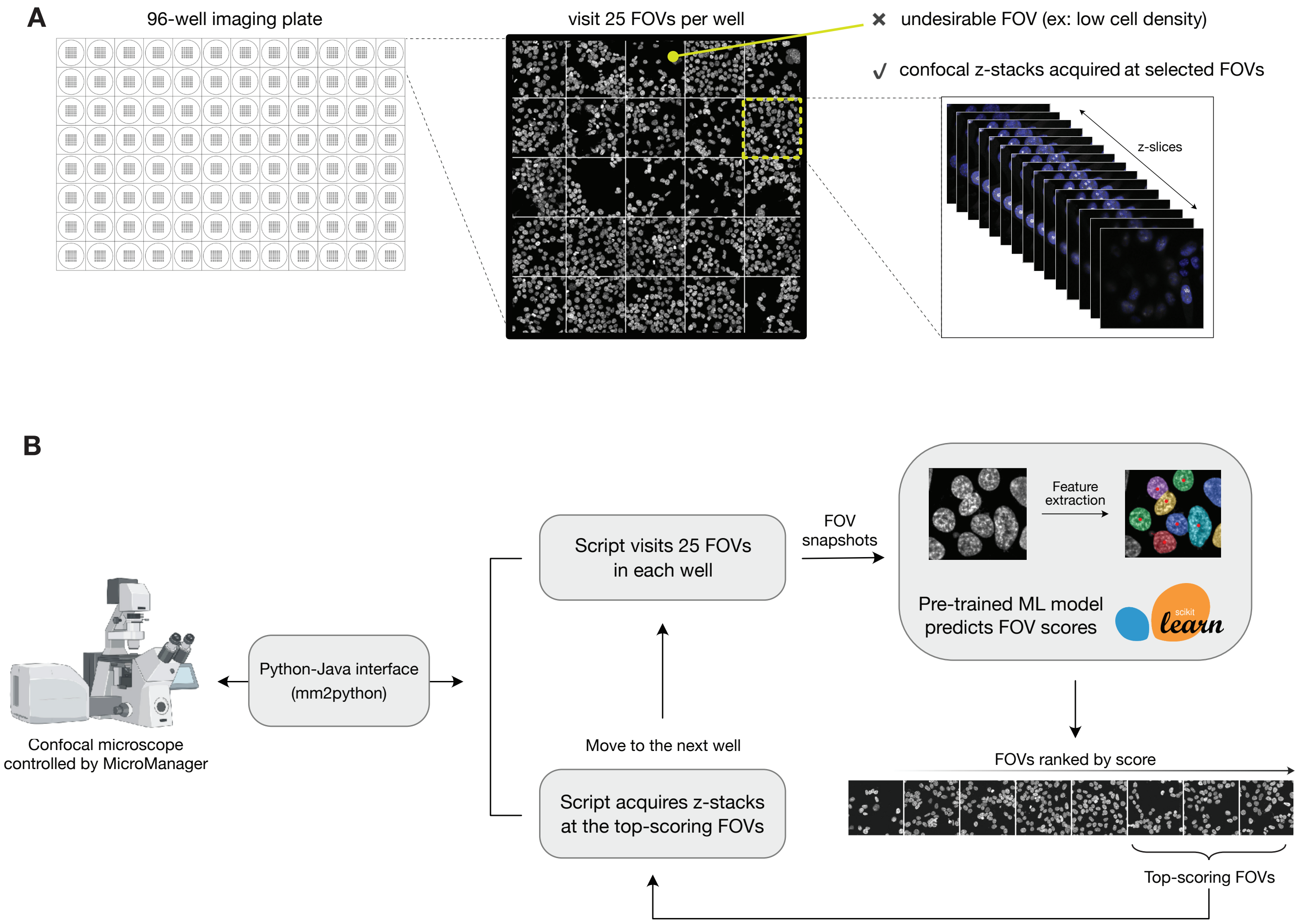


Figure S3: computer vision for automated microscopy acquisition (related to Figure 3). (A) To automate microscopy acquisition on 96-well plates and to limit experimental variability between imaging sessions (e.g., to limit variations in cell density) we paired an acquisition script, written in Python, with a pre-trained machine learning model to select fields of view (FOVs) on-the-fly during the acquisition. A total of 25 FOVs are sampled per well in a single z-plane, and desirable FOVs are selected for further 3D confocal acquisition on the basis of a score predicted by the pre-trained model. **(B)** Microscopy automation workflow. Microscope hardware is controlled by a Python-based acquisition script via an open-source MicroManager-Python bridge (mm2python; https://github.com/czbiohub/mm2python). This approach enables us to combine custom acquisition logic with the rich ecosystem of Python-based machine-learning packages. Here, we use the scikit-image package to extract features from each FOV snapshot, then use a pre-trained random-forest regression model (scikit-learn) to predict a quality score for the FOV. This process is not computationally expensive and requires less than a second; the FOV score can therefore be used immediately to determine whether the script should acquire a z-stack or else move on to the next position. To maximize the quality of our confocal z-stacks, however, we chose to visit and score all 25 FOVs in each well, then re-visit the top-scoring FOVs for confocal z-stack acquisition.

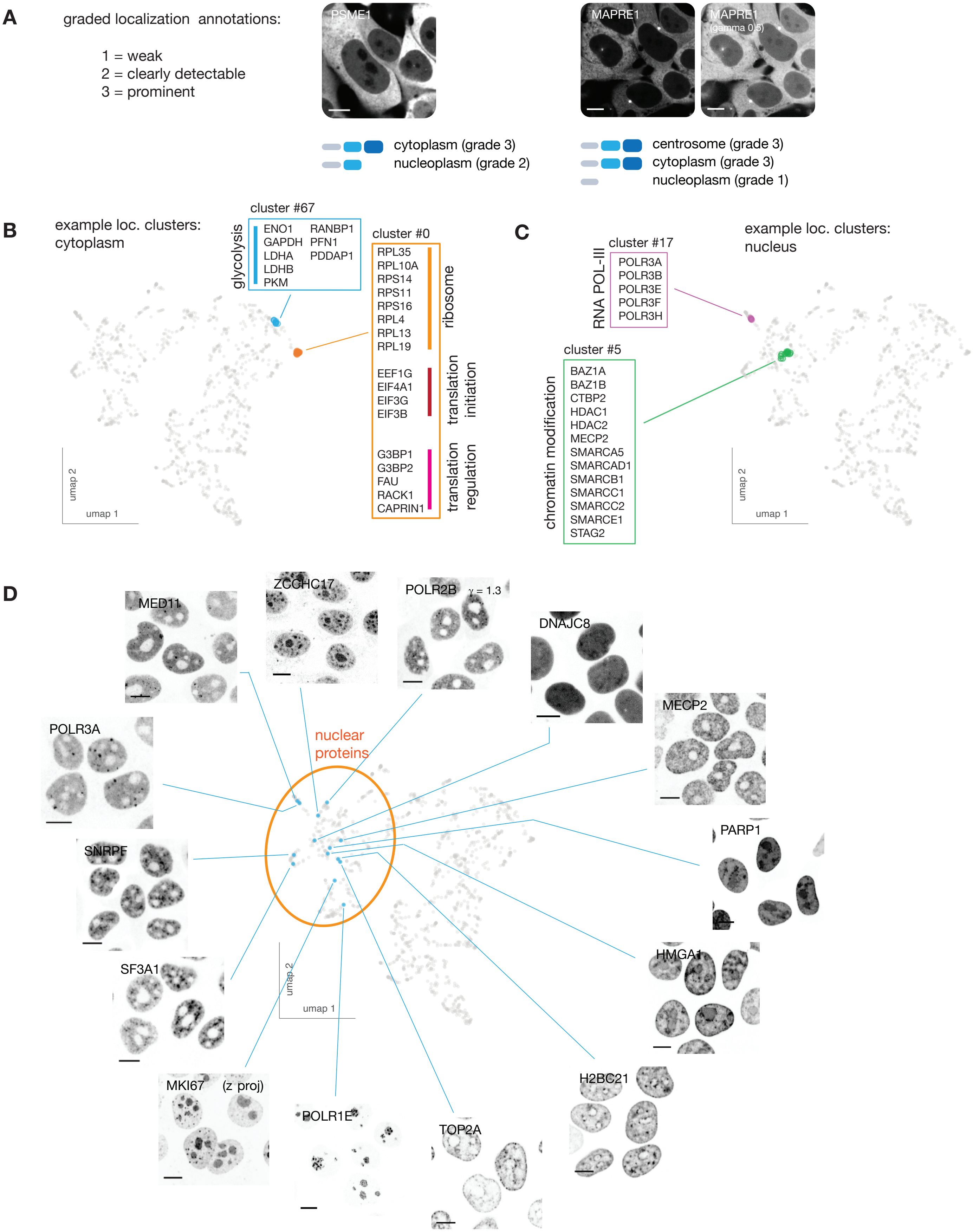


Figure S4: the OpenCell image dataset (related to Figure 3). (A) Principle of graded localization annotation (manual annotations). **(B)** Examples of cytoplasmic localization clusters. **(C)** Examples of nuclear localization clusters. **(D)** Representative images for 14 nuclear targets that exemplify the diversity of localization patterns across the proteome. Scale barsv: 10 μm.

A

```
   NHSL1    1 ------------------------------------------------------------ 0
   NHSL2    1 MPFYRRTVVPQRLCPRNPP----------------------------------------- 19
KIAA1522    1 ------------------------------------------------------------ 0
     NHS    1 MPFAKRIVEPQWLCRQRRPAPGPAVDASGGSAEPPPPLQPPGRRDLDEVEAPGPEEPARA 60

   NHSL1    1 ------------------------------MKKEGSSGSFRLQPNTGSL----------- 19
   NHSL2   20 ----------------------------------------------------------QQ 21
KIAA1522    1 ------------------------------------------------------------ 0
     NHS   61 VPAPSGLPPPPPPLPAPADQTQPPHGEASVAGEESTAGIPEAAPAAGEASSAAAAAAVLL 120

   NHSL1   20 ---SRAVSWINFSSLSRQT-------KRLFR-SDGELSVCGQQVEVDDENWIYR------ 62
   NHSL2   22 LAELRDVSHLAALSLLRQLADLCGHSLALLEDLEGHLLALG------RRTDSLYRRTVRL 75
KIAA1522    1 -----------------------------------------MVVFVGRRLPALLGLF--- 16
     NHS  121 MLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTH------RRVWALQGKLGGV 174
                                                             .

   NHSL1   63 --------------AQPRKAVS--NLDEESRWTVHYTAPWHQQENVFLPTTRPPCVEDLH 106
   NHSL2   76 RRRLPCRLLGPEEDEEELAAAN--SGRENATATAHSRSSWRQPVNVFLSSGRPPSVEELL 133
KIAA1522   17 ---------KKKG---SAKAENDKHLSVG-PGQGPGSAVDEHQDNVFFPSGRPPHLEELH 63
     NHS  175 QRVL--STLDPKQE--AVPVS---NLDIESKLSVYYRAPWHQQRNIFLPATRPPCVEELH 227
                                 .                 :  .:  *:*: : *** :*:*

   NHSL1  107 RQAKLNLKSVLRECDKLRHD-GYRSSQYYSQG-----------PTFAANASPFCDDYQDE 154
   NHSL2  134 REAQLNLQSLLQEEYEEQ----YSEARLVGQTFRSSDEATKPTPN-PRPQSARRL----- 183
KIAA1522   64 TQAQEGLRSLQHQEKQKLNKGGWDHGDTQSI--Q----------------S--SRTGPDE 103
     NHS  228 RHARQSLQALRREHRSRS----DRREQRAAAPLS---IAAPPLPAYPPAHSQRRREFKDR 280
               .*: .*::: ::  .             .                    *

   NHSL1  155 D---------EETDQKCSL---------SSSEEERFISIRRPKTPASSDFSDLNTQTNWT 196
   NHSL2  184 EFILMPTKR-------------------QLSEDETTTQGVRAPEASLSLSTTADKQTAWN 224
KIAA1522  104 DNISFCSQ----------TTSYV------------------------------------- 116
     NHS  281 HFLTFNSTRSPSPTECCHMTPWSRKSHPPEDEDTDVMLGQRPKNPIHNIPSTLDKQTNWS 340
              .

   NHSL1  197 KSLPLPTPEEKMRQQAQTVQADVVPINITASGTGQDDADGHSVYTPDHYS---------- 246
   NHSL2  225 SLFPLPILEEKRWPQLCSTQSDIVPINISGQQFDKHASLRHSLFNTETAVNPKSTLRRRR 284
KIAA1522  117 -------AESSTAEDALSIRSEM--IQRKGSTFRPHDSFP------KSGKSGRRRRERRS 161
     NHS  341 KALPLPTPEEKMKQDAQVISSCIIPINVTGVGFDREASIRCSLVHSQSVLQRRRKLRRRK 400
                      *..   :     : :  *: ..     . :        .

   NHSL1  247 ---------------------------TLGRFNSCRSAGQRSETRDSSCQTEDVKVVPPS 279
   NHSL2  285 TIIGFSNFSQRDQGHSNSPAGSVAHSTT-SDIRPSHSV---------------------- 321
KIAA1522  162 TVLGLPQHVQKELGLRNEREAPGTP--------------RAPGARDA-------VRIPTV 200
     NHS  401 TISGIPRRVQQEIDSDESPVARERNVIVHTNPDPSNTVNRISGTRDSECQTEDILIAAPS 460

   NHSL1  280 MRRIRAQKGQGIAAQMGHFSGSSGNMSVLSDSAGIVFPSRLDSDAGF-HSLPRSGARANI 338
   NHSL2  322 ------------------PEGVHG---RVAVGQDARFPSLTSPVLRTPSSEPDEPHQARS 360
KIAA1522  201 DGRPRGTSGMGARVSLQALEAEAE----AGAETEAMLQRHIDRV---------------- 240
     NHS  461 RRRIRAQRGQSIAASLSHSAGNIS---ALADKGDTMFTPAVSSRTRSR-SLPRE--GNRG 514
                                  .        .      :    .

   NHSL1  339 QSLEPRLGALGPAGDMNGTFLYQR-----------GHPQADENLGHLGGASGTGTLLRPK 387
   NHSL2  361 GPNPPGME-------------------------------------SMGMVY--------- 374
KIAA1522  241 --------------YRDDTFVGRSTGTRAPPL---TRPMS---LAVPGLTGG-------- 272
     NHS  515 GDAEPKVGAKPSAYEEGESFVGDHERTPNDFSEAPSSPSAQDHQPTLGLACS-QHLHSPQ 573
                                                             * .

   NHSL1  388 SQ----ELRHFESENIMSPACVVSPHATYSTSIIPNATLSSSSEVIAIPTAQSAGQRESK 443
   NHSL2  375 ---------------SVPSSCNGPTESTFSTSWK-------------------------- 393
KIAA1522  273 ----------AGPAEPLSPAMSISPQATYLSKLIPHAVLPPTVDVVALGRCSLRTLSR-- 320
     NHS  574 HKLSERGRSRLSRMAADSGSCDISSNS----DTF-------------------------- 603
                                 :     .:    .

   NHSL1  444 SSGSSHARIKSRDH-LISRHAVKGDPQSPGRHWNEGHATILSQDLDPHSPGE-P--ALLS 499
   NHSL2  394 --GDAFTYMTPSATSQSNQVNENGKNPSCGNSWVSLNKV--------------PPLVPK- 436
KIAA1522  321 ---CSLHSASPASVRSLGRFSSVSSPQP-------------------------------- 345
     NHS  604 --GSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSS 661
                       .       .:    ..

   NHSL1  500 LCDSAVPLNAPANRENGSQAMPYNCRNNLAFPAHPQDVDGKSESSYS---------GGGG 550
   NHSL2  437 ---------------EAATLLVA-----RD-------------------------NPAGC 451
KIAA1522  346 -------------------------------------------------------RSRHP 350
     NHS  662 HCDSELSLNTAPHANEDASVFVT-----EQYNDHLDKVRGHRANSFTSTVADLLDDPNNS 716

   NHSL1  551 HSSSEPWEYKSSGNGR-----ASPLKP---HLATPGYSTPTSNMSSCSLDQTSNKEDAGS 602
   NHSL2  452 S---GSAGYPE----RLIQQRHMPERPSKIGLLTS-------GTSRLETGPGGASRFRER 497
KIAA1522  351 SSSSDTWSHSQSSDT--IVSDGSTLSSKGGSEGQPES-----------------ST-ASN 390
     NHS  717 NTSDSEWNYLHHHHDASCRQDFSPERPKADSLGCPSFTSMATYDSFLEKSPSDKAD-TSS 775
                      :

   NHSL1  603 LYSEDHDGYCASVHTDSGHGSGNLCNSSDGFGNPRHS-VINVFVGRAQKNQGDRSNYQDK 661
   NHSL2  498 SLSVPTDSGTTDVDYDEEQKANEACALPFAST---SSEGSNSADNIASLSA-QQEAQHRR 553
KIAA1522  391 SVVPPPQGG-------S------------GRGSPSGGSTAEASDTLS-------IRSSGQ 424
     NHS  776 HFSVDTEGYYTSMHFDCGLKGNKSYVCHYAALGPENGQGVGASPGLPDCAW-QDYLDHKR 834
                    :.                     .      .                      :

   NHSL1  662 SLSRNISLKKAKKPPLPPSRTDSLRRIPKKSSQ--------CNGQVLNESLIATLQHSLQ 713
   NHSL2  554 QRSKSISLRKAKKKPSPPTRSVSLVKDEPGLL--------PEGGSALPKDQ---RPKSLC 602
KIAA1522  425 LSGRSVSLRKLKRPPPPPRRTHSLHQRGLAV-----------------PDGP-------- 459
     NHS  835 QGRPSISFRKPKAKPTPPKRSSSLRKSDGNADISEKKEPKISSGQHLPHSS---REMKLP 891
                  .:*::* *  * ** *: ** :                      ..

   NHSL1  714 LSLPGKSGSSPSQSPCSDLEEPWLPRS-RSQSTVSAGSSMTSATTPNVYSLCGATPSQSD 772
   NHSL2  603 LSLE------------------------HQGHHSS------------HPDAQGHPAIPNH 626
KIAA1522  460 LGLPPKPERKQQ---------PQLPRPPTTGGSE----------------GAGAAPC--- 491
     NHS  892 LDFANTPSRMENANLPTKQEPSWINQS-EQGIKEP------------QLDASDIPPF-KD 937
              *.:                                                 .
```

```
 773 TSSVKSEYTDPWGYYI-----DYTGMQ---EDPGNPAGGCSTSSGVPTGNGPVRHVQEGS 824
 627 KDPESTQFSHHWYLTDWKSGDTYQSLSSSSTATGTTVIECTQVQG----SSESLASPSTS 682
 492 ----PPNPANSWVPGL-------------------------------------------- 503
 938 EVAESTHYADLWLLNDLKTNDPYRSLSNSSTATGTTVIECIKSPE----SSESQTSQSES 993
           . :. *

 825 RATMPQVPGGSVK-PKIMSPEKSHRVISPSSGYSSQSNTPTALTPVPVFLKSVSPANGKG 883
 683 RATTPSQLSIEVEAREISSPGRPPGLMSPSSGYSSQSETPTPTVSMSLTLGHLPPPSSSV 742
 504 ------SPGG-----SRRPPRSPERTLSPSSGYSSQSGTPTLPP-----KGLAGPPASPG 547
 994 RATTPSLPSVDNE-FKLASPEKLAGLASPSSGYSSQSETPTSSFPTAFFSGPLSPGG--S 1050
             .      .   *       ********** ***             *

 884 KPKPKVPERKSSLISSVSISSS-STSLSSSTS------TEGSGTMKKLDPA--------V 928
 743 RVRPVVPERKSSLPPTSPMEKF-PK-SR-----------------LSFD----------- 772
 548 KAQPPKPERVTSLRSPGASVSSSLTSLCSSSSDPAPSDRSGPQILTPLGDRFVIPPHPKV 607
1051 KRKPKVPERKSSLQQPSLKDGT-IS-LS-----------------KDLE----------- 1080
     : :*  *** :**           .                      :

 929 GSPPAPPPPPVPSPPFPCPADRSPFLPPPP-PVTDCSQGSPLPHSPVFPPPPPEALIPFC 987
 773 -LPLTS-SPNLDLSGMS-ISIRS------KTKVSRHHSETNFGVKLAQKTNPNQPIMPMV 823
 608 PAPFSPPPSKPRSPNPAAPALAAPAVVPGPVSTTDA--------SPQSPPTPQTTLTPLQ 659
1081 -LPIIP-PTHLDLSALH-NVLNKPFHHRHPLHVFTHNKQNTVGETLRSNPPPSLA----I 1133
       *                             .           .      *

 988 SPPDWCLSP-------------PRPALS----------PILPDSPVS----LPLPPPLLP 1020
 824 TQSDLRSVR-------------LRS--------------VSKSEPEDDIES----PEYAE 852
 660 ESPVISKDQSPPPSPPPSYHPPPPPTKKPEVVVEAPSASETAEEPLQDPNWPPPPPPAPE 719
1134 TPTILKSVN-------------LRS--------------INKSEEVKQKEENNTDLPYLE 1166
                                               ..  .

1021 SSEPPP--APPLDPKFMKDTRPPFTN-SGQPE--SS---------RGSLRPPSTKE---- 1062
 853 EPRAEE--VFTLPE---RKTKPPVAE-KPPVA--RR-------PPSLVHKPPSVPE---- 893
 720 EQDLSMADFPPPEEAFFSVASPEPAGPSGSPELVSSPAASSSSATALQIQPPGSPDPPPA 779
1167 ESTLTT--A-ALSP---SKIRPHTAN-------------------------KSVSR---- 1191
     .                    *  :                           .

1063 -----------------------------E--TSRPPMPLITTEALQMVQLR-PVRKNSG 1090
 894 -----------------------------EYALTSPTLAMPPRSSIQHAR---PL-PQDS 920
 780 PPAPAPASSAPGHVAKLPQKEPVGCSKGGGPPREDVGAPLVTPSLLQMVRLRSVGAP-GG 838
1192 -----------------------------QYSTEDTILSFLDSSAVEMGPDKLHLEKNST 1222
                                            :   . ::           .

1091 AEAAQLSERTAQEQRTPVAPQYHLKPSAFLKSRNSTNEMESESQPASV------------ 1138
 921 YTV---VRK-----------------------------------PKPS------------ 930
 839 APTPALG---------PSAPQKPL------R-----RALSGRASPVPAPSSGLHAAVRLK 878
1223 FDV---KNRCDPET-------ITSAGSSLLDSNVTKDQVRTETEPIPE-------NTPTK 1265
       .                                         *

1139 TSSLPTPAKSSS-----QGDH----------GSAAE------RGG-PVSRSPGAPSAGEA 1176
 931 --SFPDGRSPGES-TAPSSL------------VFTPF--ASSSDA-FFSGTQQPPQGSVE 972
 879 ACSLAAS--EGLSSAQPNGPPEAEPRPPQSPASTASFIFSKGSRKLQLER-PVSPETQ-A 934
1266 NCAFPTEGFQRVSAARPNDLD----------GKIIQY--GPGPDE-TLEQVQKAPSAGLE 1312
       ::             ..                            ..     *.

1177 EARPSPSTTP---------LPDSSPSRKPPPISKKPKLFLVV-PPPQKDFAVEPAENVSE 1226
 973 DEGPKVRVLPERISLQSQEEAEKKKGKIPPPVPKKPSVLYLPLTSPTAQ--------M-E 1023
 935 DLQRNLVAELRSISEQRPPQAPKKSPKAPPPVARKPSVGVPPPASPSYPRA--------E 986
1313 EV-----AQPESVD-----------------------VITSQSDSPTRATDV--SNQFKH 1342
     :      .                             :       *             .

1227 ALRAVPSPTTGEEGSVHSREAKESSAAQAGSHATHPGTSVLEGGAAGSMS-PSRV---EA 1282
1024 AY-------------V-----------AEPR-LPLSPIITLEE--------DTKCPATGD 1050
 987 PLTAP--PTN---GLPHTQDRTKREL------AENGG----------------------- 1012
1343 QF-------------V-----------MSRHHDKVPGTISYESEITSVNSFPEKCSKQEN 1378

1283 NVPMVQPDVSPAPKQEEPAENSADTGGD-GESCLSQQDGAAGVPETNAAGSSSEACDFLK 1341
1051 DLQSLGQRVTSTPQADS--EREASPLGSSV-----E----PG--TEEKSLISDKTAEWIA 1097
1013 VLQLVGPE---------------------------EKMGLPG---SD------------- 1029
1379 I----ASGISAKSASDN--SKAEETQGNVDEASLKE----SS--PSDDSIISPLSED-SQ 1425
                                        :     .    :

1342 EDGNDEVMTPSRPRTTEDLFAAIHRSKRKVLGRRDSDDDHSRNHSPS--PPV-------- 1391
1098 EDDDDVFV---ASRTTEDLFTVIHRSKRKLLGWKEPGEAFVGGRTSSHSPIKNTAE---- 1150
1030 ------------------------------------------------------------ 1029
1426 AEAEGVFVSPNKPRTTEDLFAVIHRSKRKVLGRKDSGDMSVRSKSRA--PLSSSSSSASS 1483

1392 --TPTGAAPSLASPKQVGSIQRSIRKSSTSSDNFKALLLKKGSRSDTSARMSAAEMLKNT 1449
1151 --SPISESTATA---GS---GSSANLDAGRNDDFKALLQKKGSKATPRSRPSAAELLKTT 1202
1030 --------------------------------SQKELA---------------------- 1035
1484 ITSPSSNVTTPNSQRSPGLIYRNAKKSNTSNEEFKLLLLKKGSRSDSSYRMSATEILKSP 1543
                                     . * *

1450 DPRFQRSRSEPSPDAPESP----------------SSCSPSKNRRAQEEWAKNEGLMPRS 1493
1203 NPLARRIIAQFSKDY-ETTDNPST------------------------------------ 1225
1036 ------------------------------------------------------------ 1035
1544 IL--PKPPGELTAESPQSTDDAHQGSQGAEALSPLSPCSPRV---------NAEGFSSKS 1592

1494 L-SFSGPRYGRSRTPPSAASSRYSMRNRIQSSPMTVISEGEGEAVEPVDSIARGALGAAE 1552
1226 ------------------------------------------------------------ 1225
1036 ------------------------------------------------------------ 1035
1593 FATSASARVGRSRAPPAASSSRYSVRCRLYNTPMQAISEGETENSDGSPHDDRSSQSST- 1651

1553 GCSLDGLAREEMDEGGLLCGEGPAASLQPQAPGPVDGTASAEGREPSPQCGGSLSEES 1610
1226 ---------------------------------------------------------- 1225
1036 ---------------------------------------------------------- 1035
1652 ---------------------------------------------------------- 1651
```

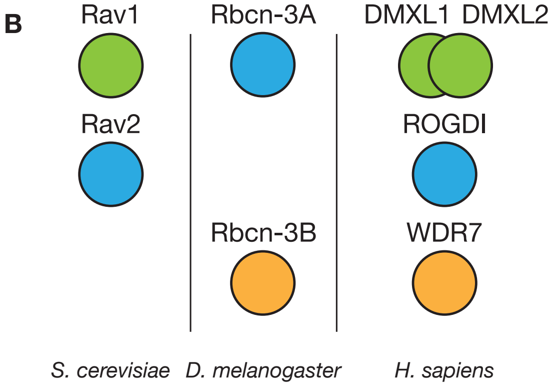


Figure S5: sequence analysis of orphan proteins (related to Figure 6). (A) Amino-acid sequence alignment between human NHSL1, NSHL2, KIAA1522 and NHS. **(B)** Correspondence of RAVE complex members in *S. cerevisiae, D. melanogaster* and *H. sapiens*. Note that in *S. cerevisiae* RAVE also includes Skp1, not depicted here.

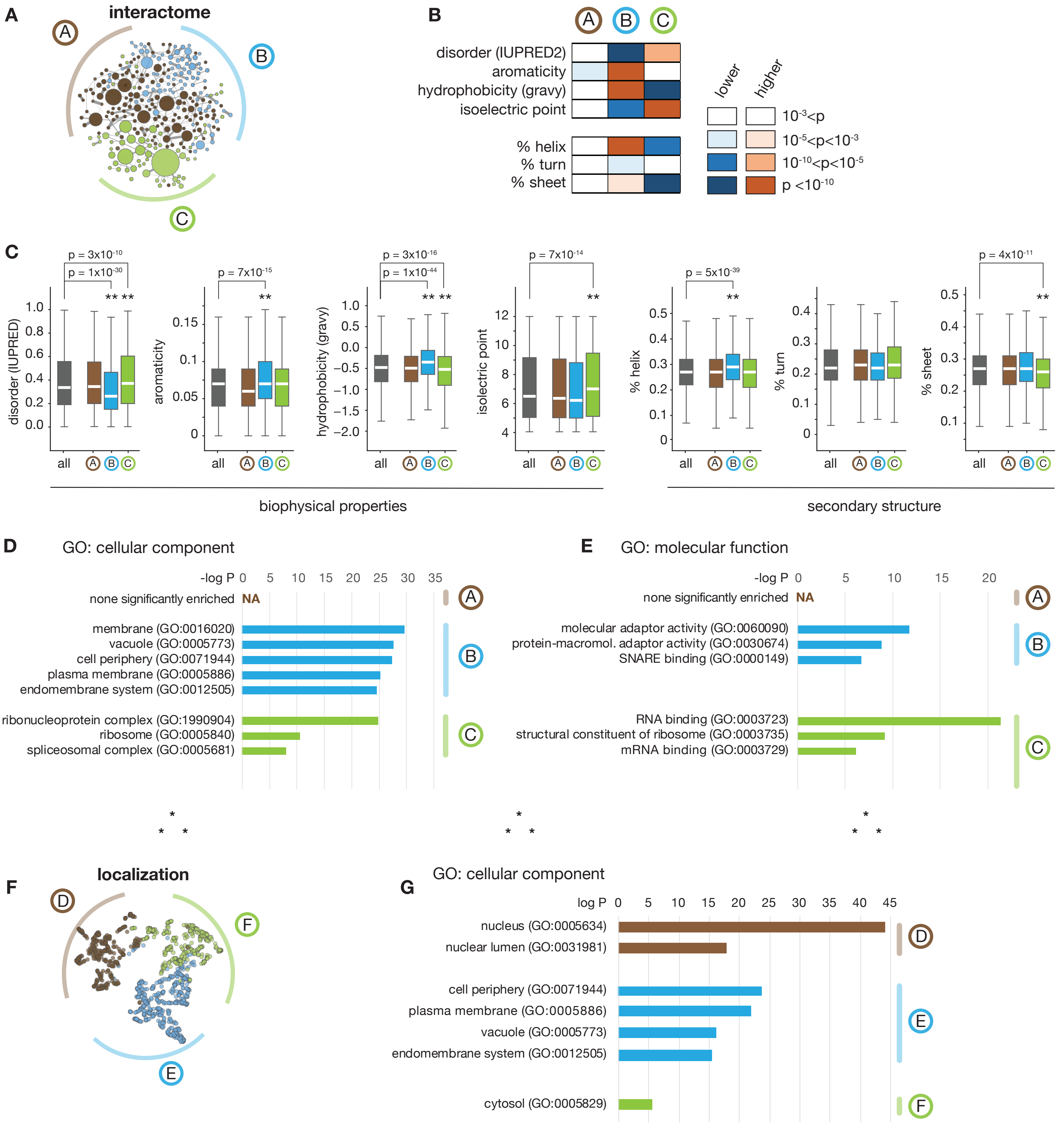


Figure S6: biophysical & ontology analysis of the main branches from interactome and localization hierarchies (related to Figure 7). **(A)** The three branches derived from the interactome hierarchy (see Figure 7A). **(B)** Heat-map representing significance testing of biophysical properties of protein sequences in the 3 branches. P-values were obtained using Student's t-test comparing proteins belonging to a specific hierarchical branch against all proteins in the three branches. **(C)** Box plot representing the significance testing of biophysical properties described in (B). Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent minimum and maximum values. Median is represented by a white line. ** p < 10-9 (Student's t-test), exact p-values are shown. **(D), (E)** Enrichment analysis of GO annotations in the hierarchical branches, testing GO term enrichment of proteins in each branch against all proteins in the interactome (Fisher's exact test, showing annotations enriched at $p < 10^{-10}$ and excluding near-synonymous annotations). **(F), (G)**: same as (A) and (D) but for localization-based hierarchical branches (see Figure 7B).