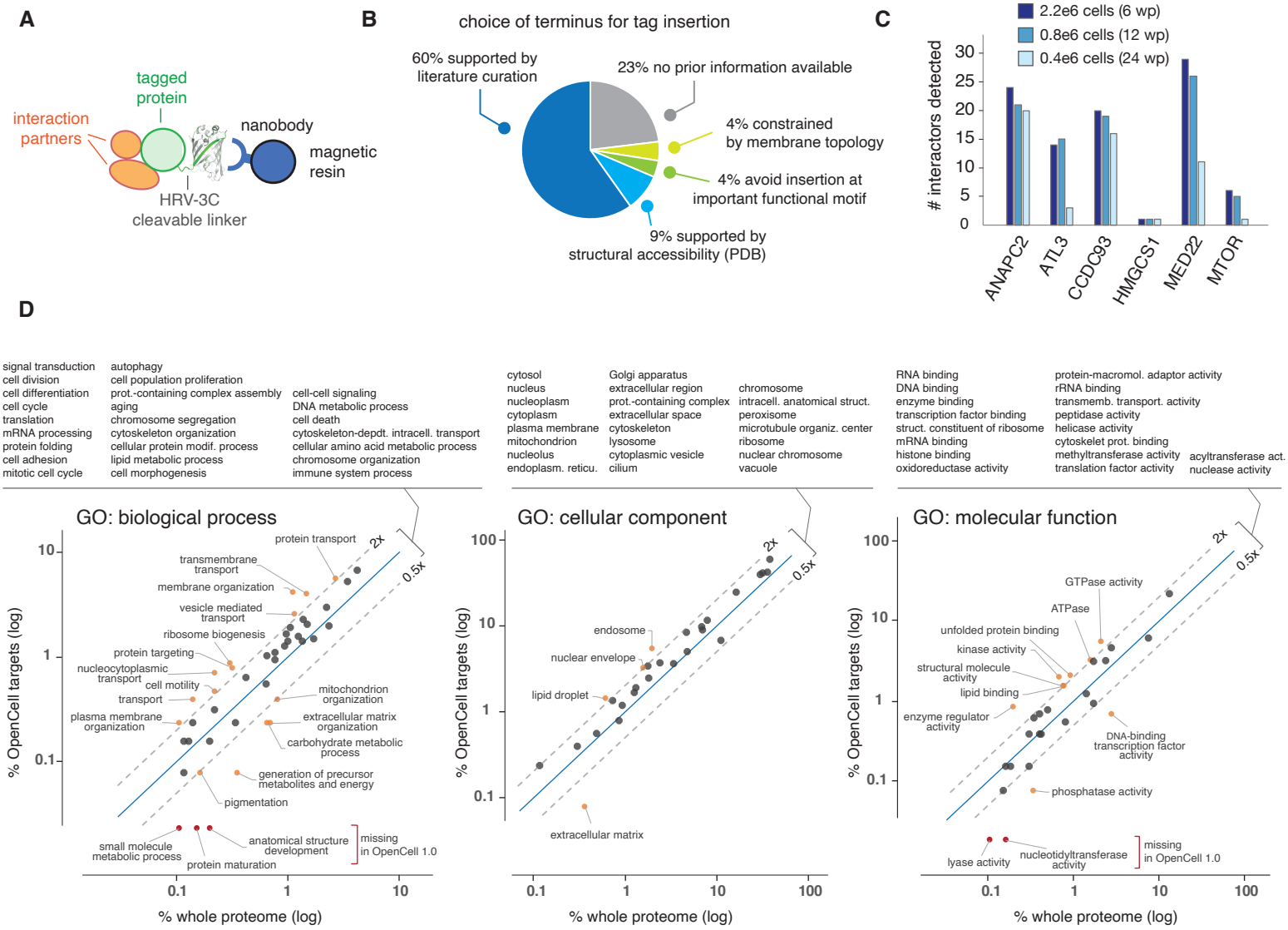

Figure S1: experimental pipeline (related to Figure 1). (A) IP/MS using FP capture. All mNG11 tagging constructs also include an HRV-3C cleavable linker for optional release from the capture resin. **(B)** Justifying the choice of tag insertion in engineered cell lines. To inform tag insertion sites, we used a combination of existing data from the literature suggesting preservation of properties, 3D structures of protein complexes from the PDB and sequence analysis to avoid important functional motifs. 4% of insertion sites were constrained by the topology of transmembrane protein targets (fusion to cytosolic termini), and for 23% of targets no prior data was available. See details in Suppl. Table 1. **(C)** 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). **(D)** Distribution of gene ontology annotations in the OpenCell library (successful targets only) compared the whole proteome. Over- and under-represented terms are outlined. Because organellar organization and transport between organelles are foundational to human cellular architecture, proteins in these groups are slightly enriched in our library. Under-represented groups are mostly comprised of proteins in compartments that are not accessible to our tagging strategy (mitochondrial functions, extracellular matrix) or proteins that are typically present at low copy numbers and therefore difficult to detect at endogenous levels (transcription factors).

Figure S1

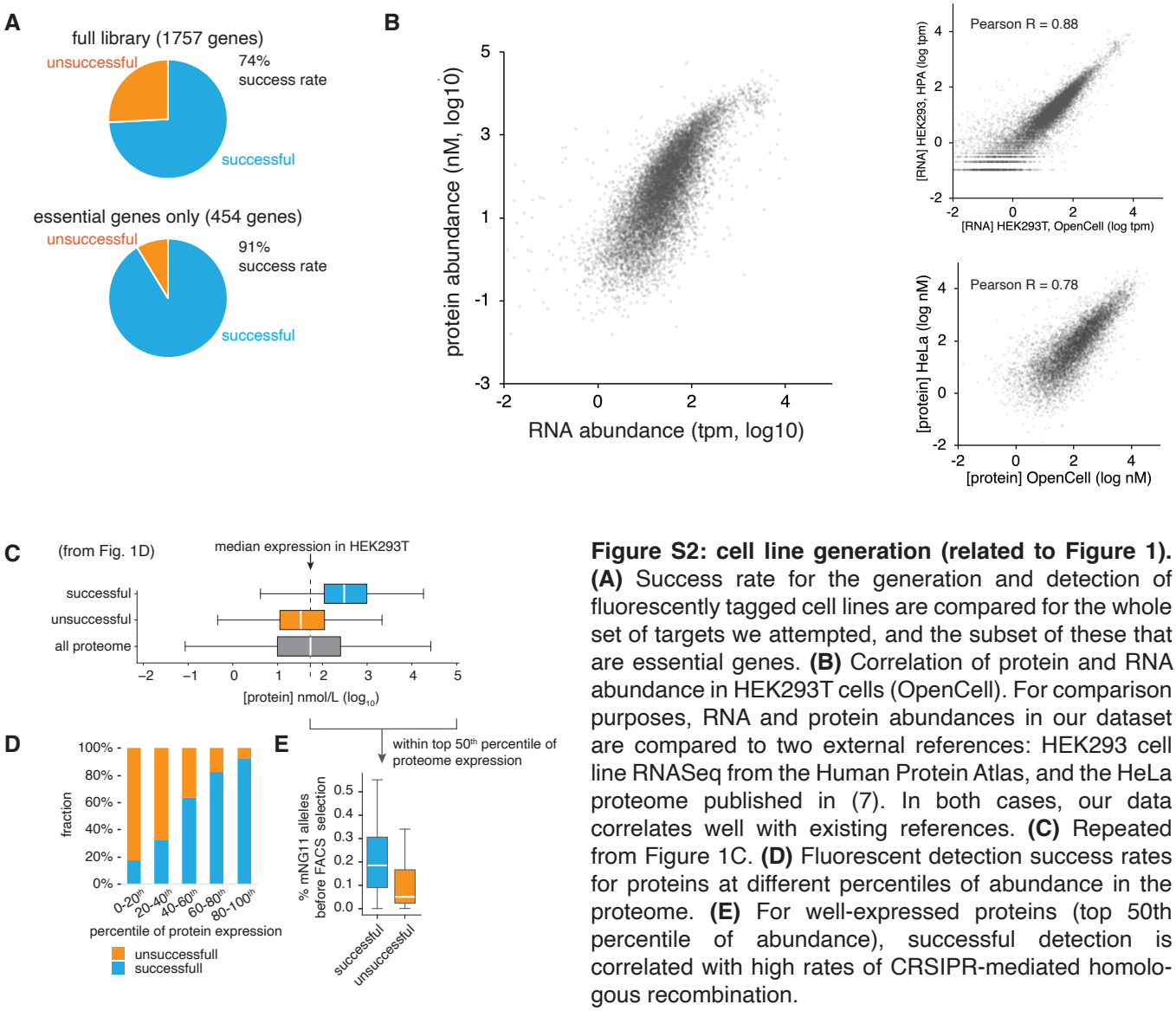

Figure S2: cell line generation (related to Figure 1). **(A)** Success rate for the generation and detection of fluorescently tagged cell lines are compared for the whole set of targets we attempted, and the subset of these that are essential genes. **(B)** Correlation of protein and RNA abundance in HEK293T cells (OpenCell). For comparison purposes, RNA and protein abundances in our dataset are compared to two external references: HEK293 cell line RNASeq from the Human Protein Atlas, and the HeLa proteome published in (7). In both cases, our data correlates well with existing references. **(C)** Repeated from Figure 1C. **(D)** Fluorescent detection success rates for proteins at different percentiles of abundance in the proteome. **(E)** For well-expressed proteins (top 50th percentile of abundance), successful detection is correlated with high rates of CRSIPR-mediated homologous recombination.

Figure S2

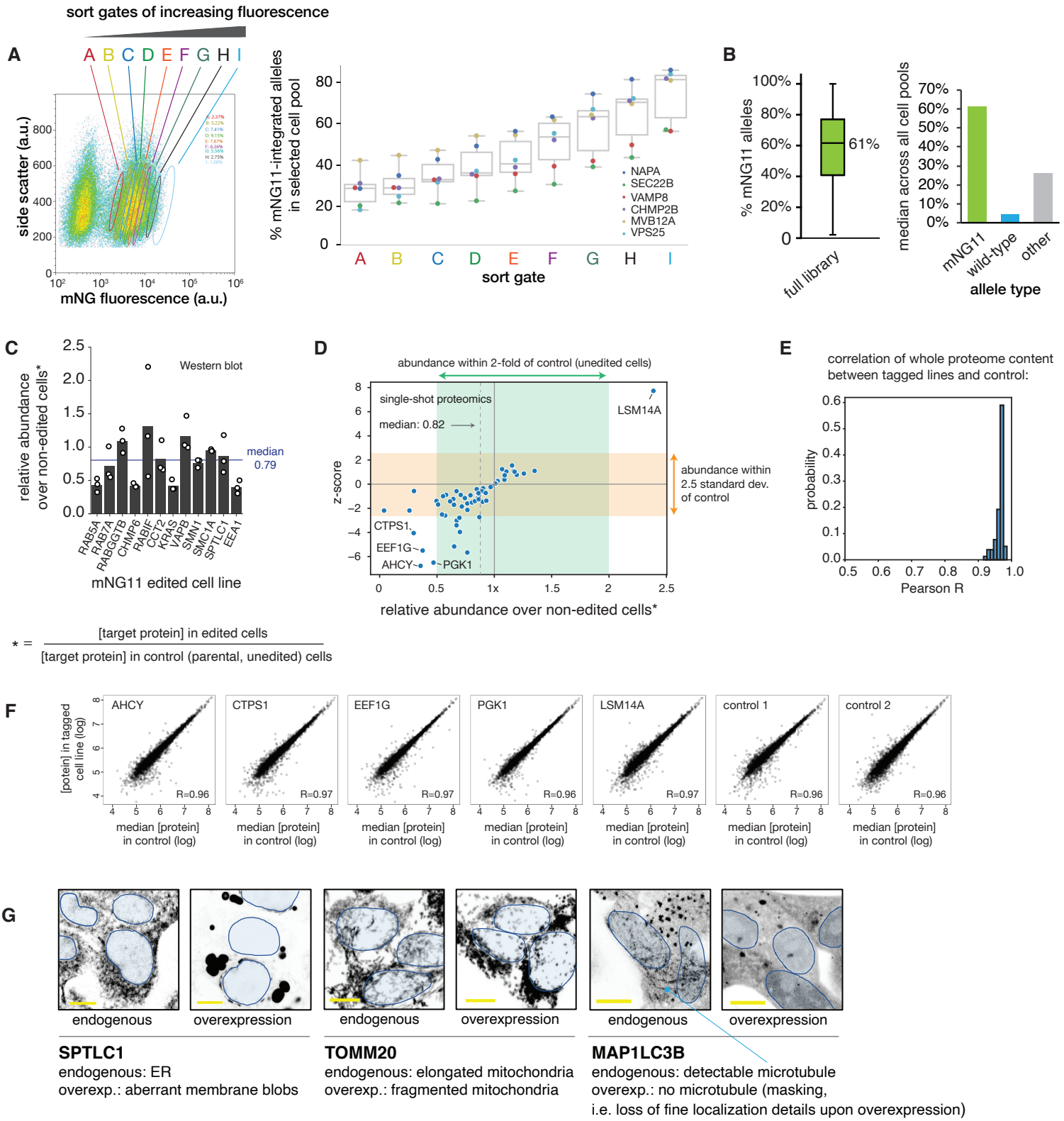

Figure S3

(legend on next page)

Figure S3: cell library characterization and quality control (related to Figure 1). (A) 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. **(B)** 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 1.5x interquartile range). The median values of mNG11 integrated alleles, wt alleles and other alleles are shown on the right. **(C)** Measurement of target protein abundance in final selected cell pools vs. parental cell line, by quantitative Western blotting. **(D)** Measurement of target protein abundance in final selected cell pools vs. parental cell line, by single-shot mass spectrometry. In these experiments, tagged lines are measured in a single replicate and compared to 6 replicates of non-edited control cell lines. Outliers targets are defined by an abundance that deviates by more than 2.5 standard deviations and by more than 2-fold of their abundance in the controls. The 5 outlier lines are outlined. **(E)** Distribution of Pearson correlation values measuring the overall correlation of abundances for all cellular proteins in each tagged cell line vs. median control. **(F)** For the outliers outlined in (D), correlation of abundances for all cellular proteins in the tagged cell line vs. median control. The abundance correlations for two individual control repeats are shown for reference. **(G)** Examples of overexpression artifacts. Single z-slice confocal images are shown (scale bar: 10 μm). Endogenously tagged lines and their equivalent overexpression constructs 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). “Masking effects” are defined as the loss of fine localization details upon overexpression.

Figure S3 (legend)

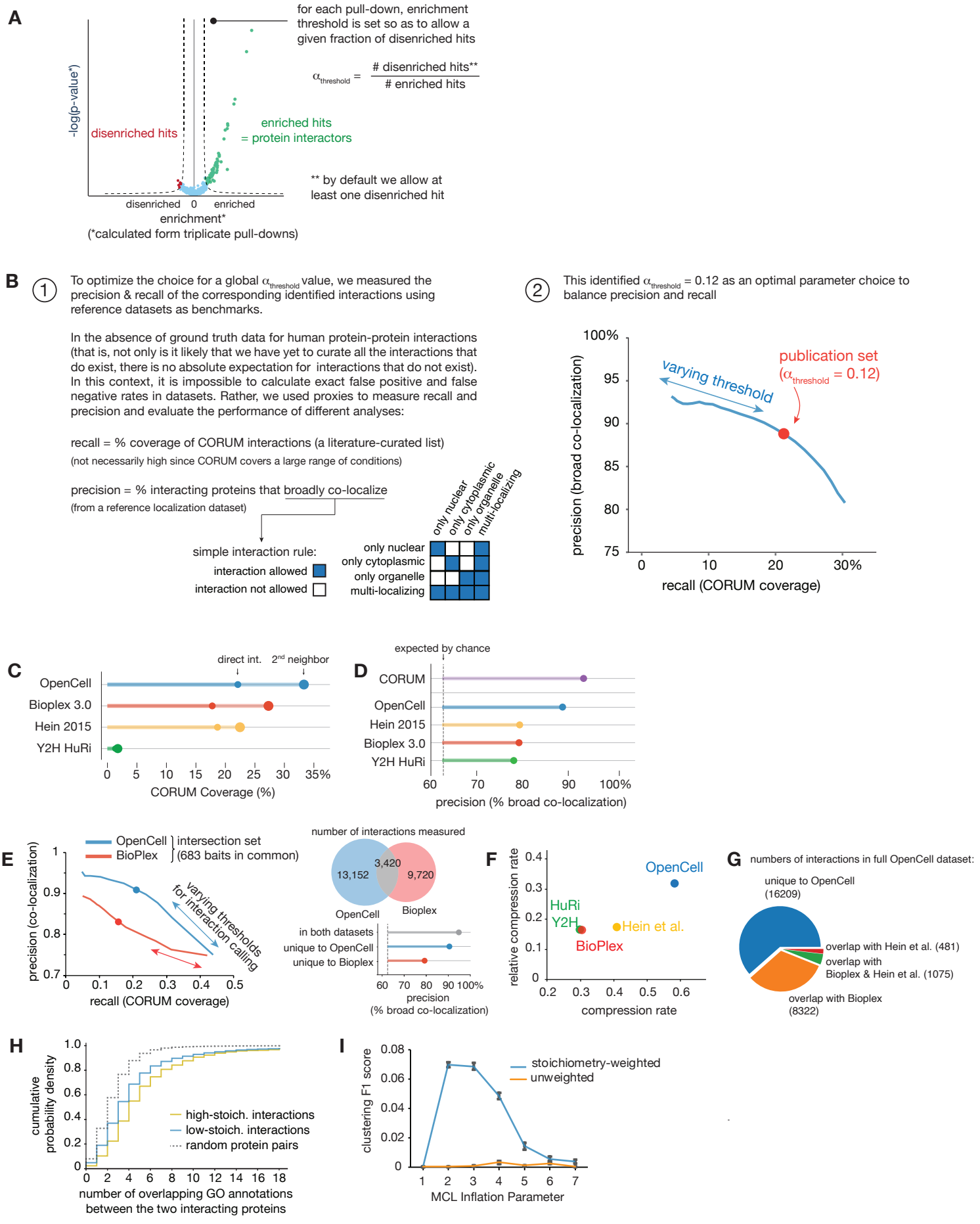

Figure S4

(legend on next page)

F**igure S4: 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 $\alpha_{threshold}$. **(B)** To optimize parameter choice, we measured how precision (% co-localization) and recall (% CORUM coverage) of the corresponding interaction network varied with $\alpha_{threshold}$. 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 on identical bait set. Both datasets use the same HEK-293T cell line and share a large number (683) of baits in common. Precision and recall analysis by varying threshold for interaction detection ($\alpha_{threshold}$ in OpenCell and *pInt* in Bioplex) 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 many new measured interactions for that intersection set of baits (right panel, top). Interestingly, the interactions unique to OpenCell have high precision values (right panel, bottom). **(F)** Compressibility analysis (31) of OpenCell vs. other large-scale interactomes. **(G)** Number of interactions measured in OpenCell (in the full dataset) that were also measured in Hein et al. (7) or BioPlex 3.0. **(H)** Distribution of GO annotation overlap between protein pairs identified in low-stoichiometry and high-stoichiometry interactions. **(I)** MCL clustering performance (F1 score) using stoichiometry-weighted or unweighted interaction graphs, derived from CORUM interactions as described in Drew et al (89).

Figure S4 (legend)

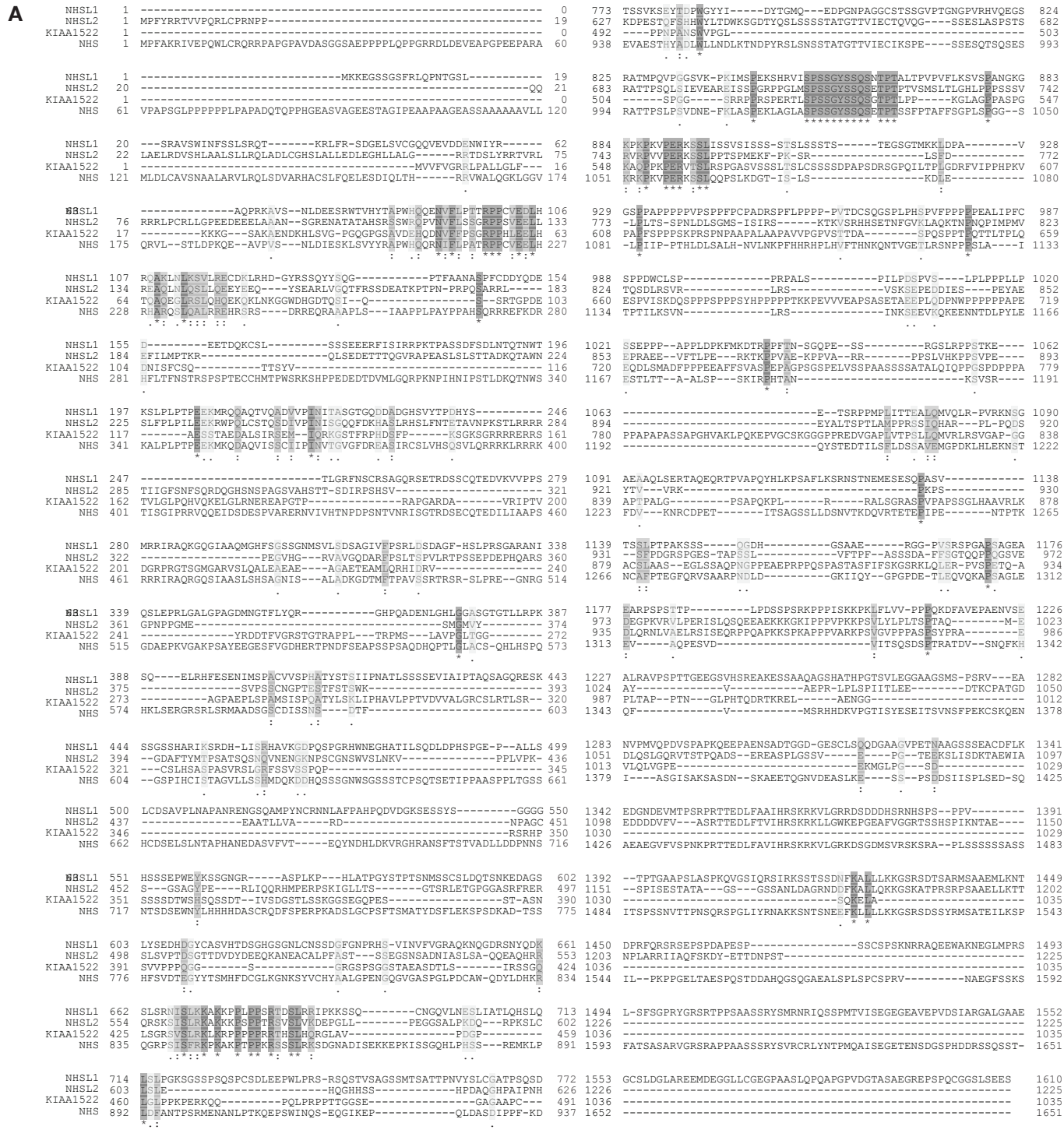

Figure S5: sequence analysis of orphan proteins (related to Figure 2). (A) Amino-acid sequence alignment between human NHSL1, NSHL2, KIAA1522 and NHS. **(B)** Correspondence of RAVE complex members in S. cerevisiae, D. melanogaster and H. sapiens. Note that in S. cerevisiae RAVE also includes Skp1, not depicted here.

Figure S5

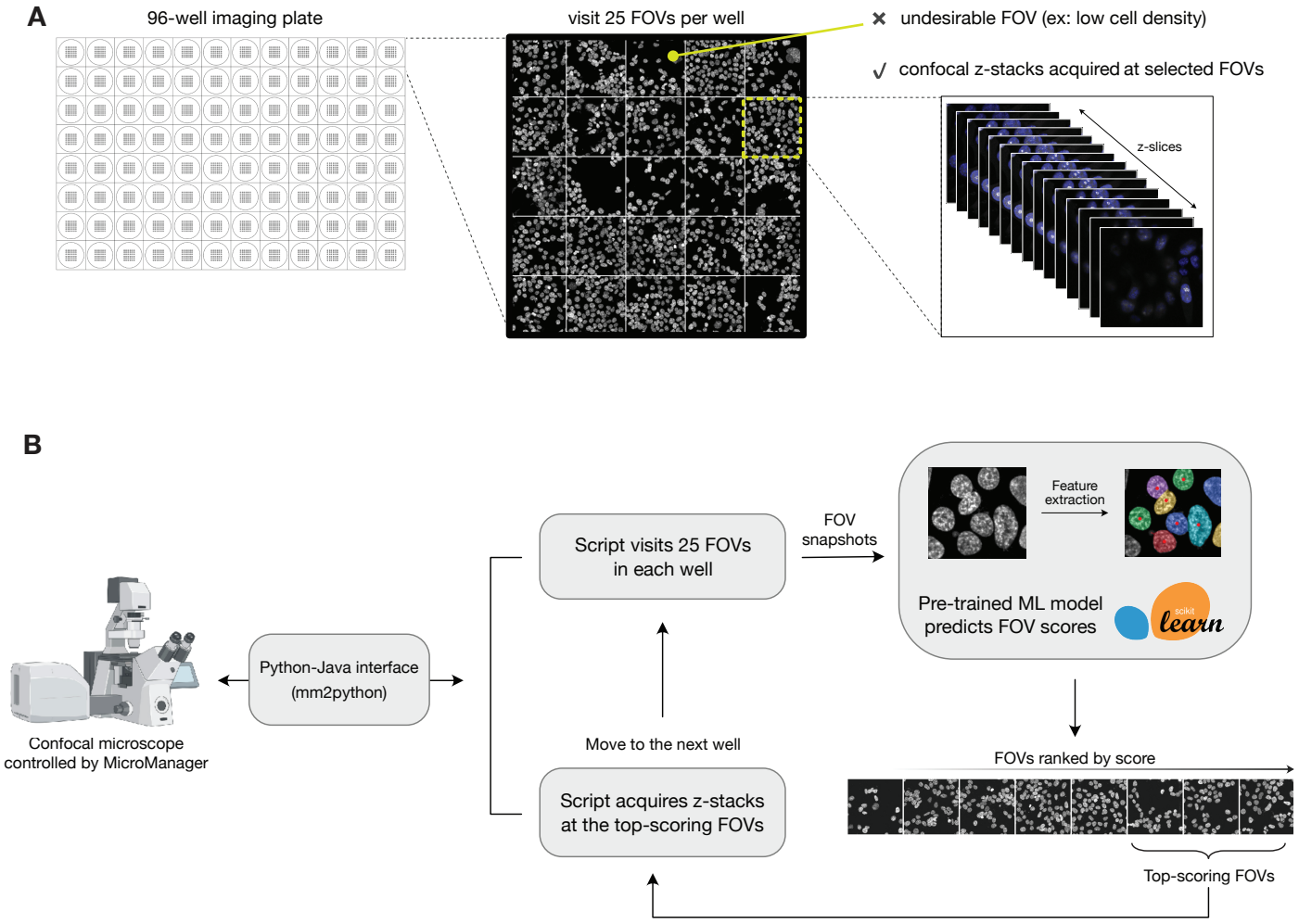

Figure S6: computer vision for automated microscopy acquisition (related to Figure 3). (A) To automate microscopy acquisition on 96-well plates and to limit experimental variability between imaging sessions (e.g., to limit variations in cell density) we paired an acquisition script, written in Python, with a pre-trained machine learning model to select field of views (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 S6

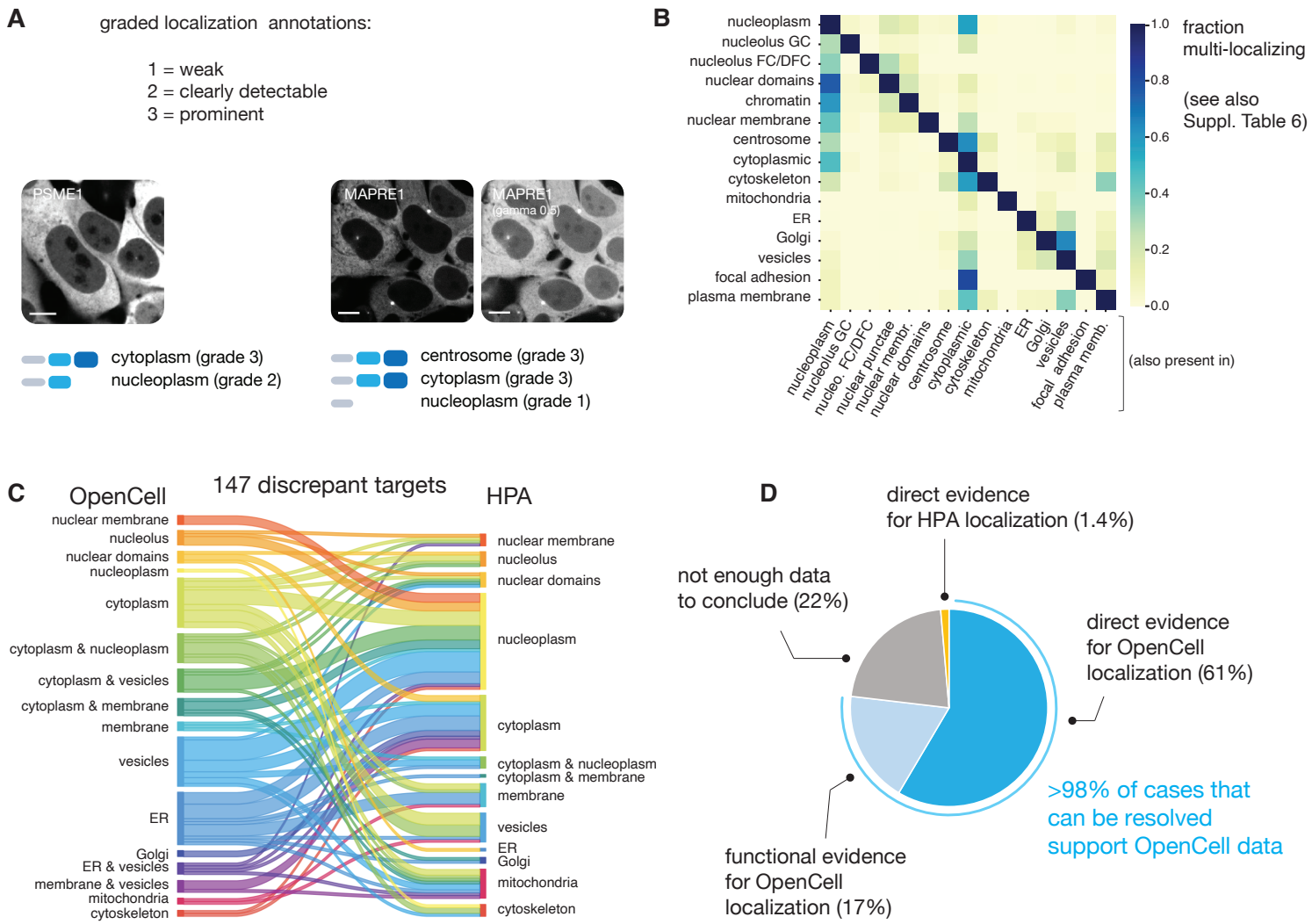

Figure S7: the OpenCell image dataset (related to Figure 3). (A) Principle of graded localization annotation (manual annotations). **(B)** Fraction of multi-localization between cellular compartments. Complete localization annotations can be found in Suppl. Table 6. **(C)** Comparison of annotated localization for proteins in OpenCell and Human Protein Atlas (HPA, version v20) datasets for which annotations are inconsistent. **(D)** Extensive literature curation allows to resolve 77% of OpenCell/HPA discrepancies (full details in Suppl. Table 8). Here "direct evidence" refers to proteins for which localization has been directly measured in published studies, while "functional evidence" refers to proteins for which localization might not have been directly measured, but for which literature establishes a function that is predictive of a specific localization. For example, SCFD1 is a protein whose main known function is to regulate transport between ER and Golgi. This qualifies as "functional evidence". It is annotated as localized in the ER and Golgi in OpenCell, and in the nucleoplasm (main) and cytosol (additional) in HPA. **(E)** Comparison of annotated localization for 350 orthologous proteins in OpenCell and *S. cerevisiae* yeast (from LoQate (46)). Note that in yeast Golgi and vesicles are difficult to distinguish.

Figure S7

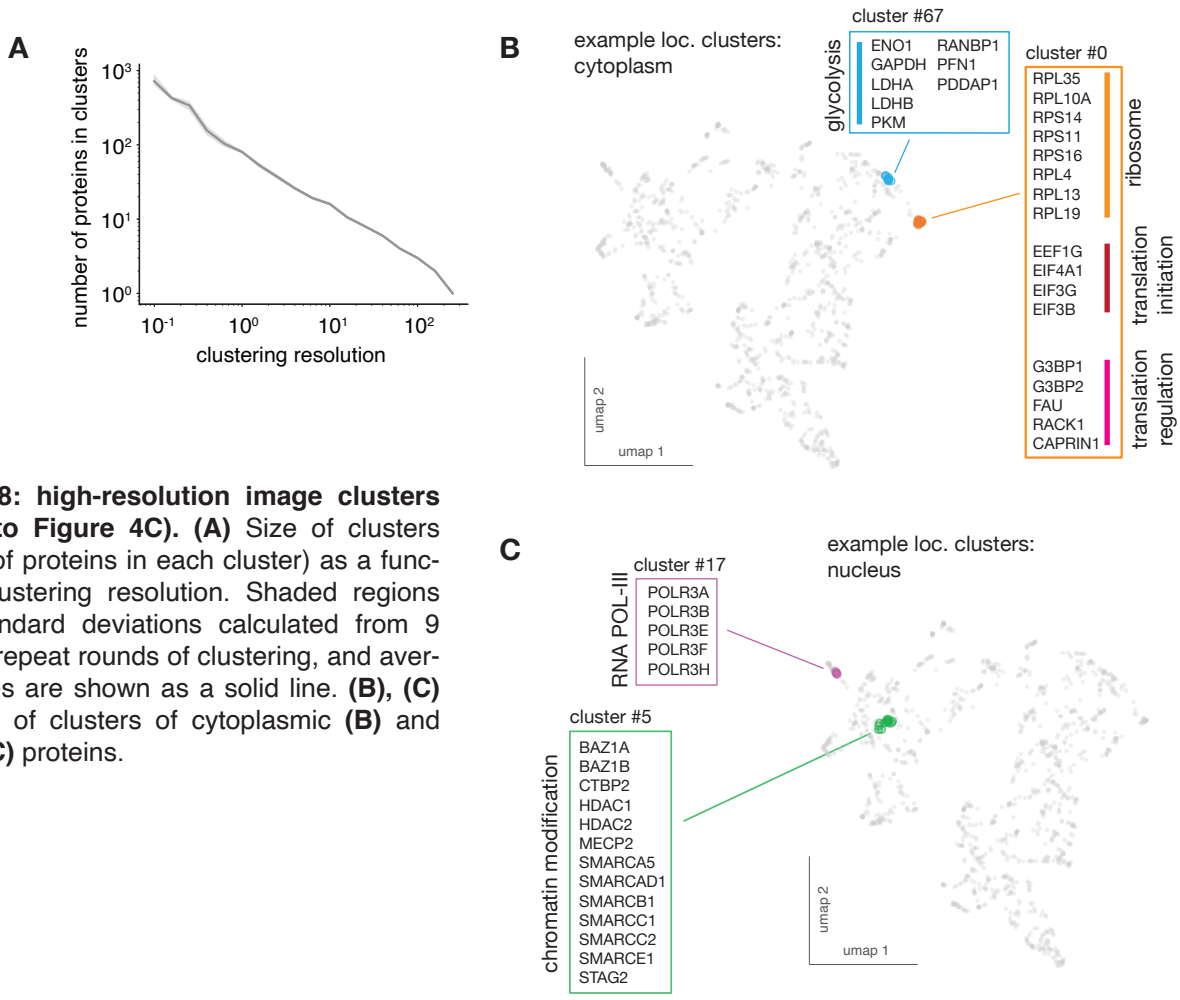

Figure S8: high-resolution image clusters (related to Figure 4C). (A) Size of clusters (number of proteins in each cluster) as a function of clustering resolution. Shaded regions show standard deviations calculated from 9 separate repeat rounds of clustering, and average values are shown as a solid line. **(B), (C)** Examples of clusters of cytoplasmic **(B)** and nuclear **(C)** proteins.

Figure S8

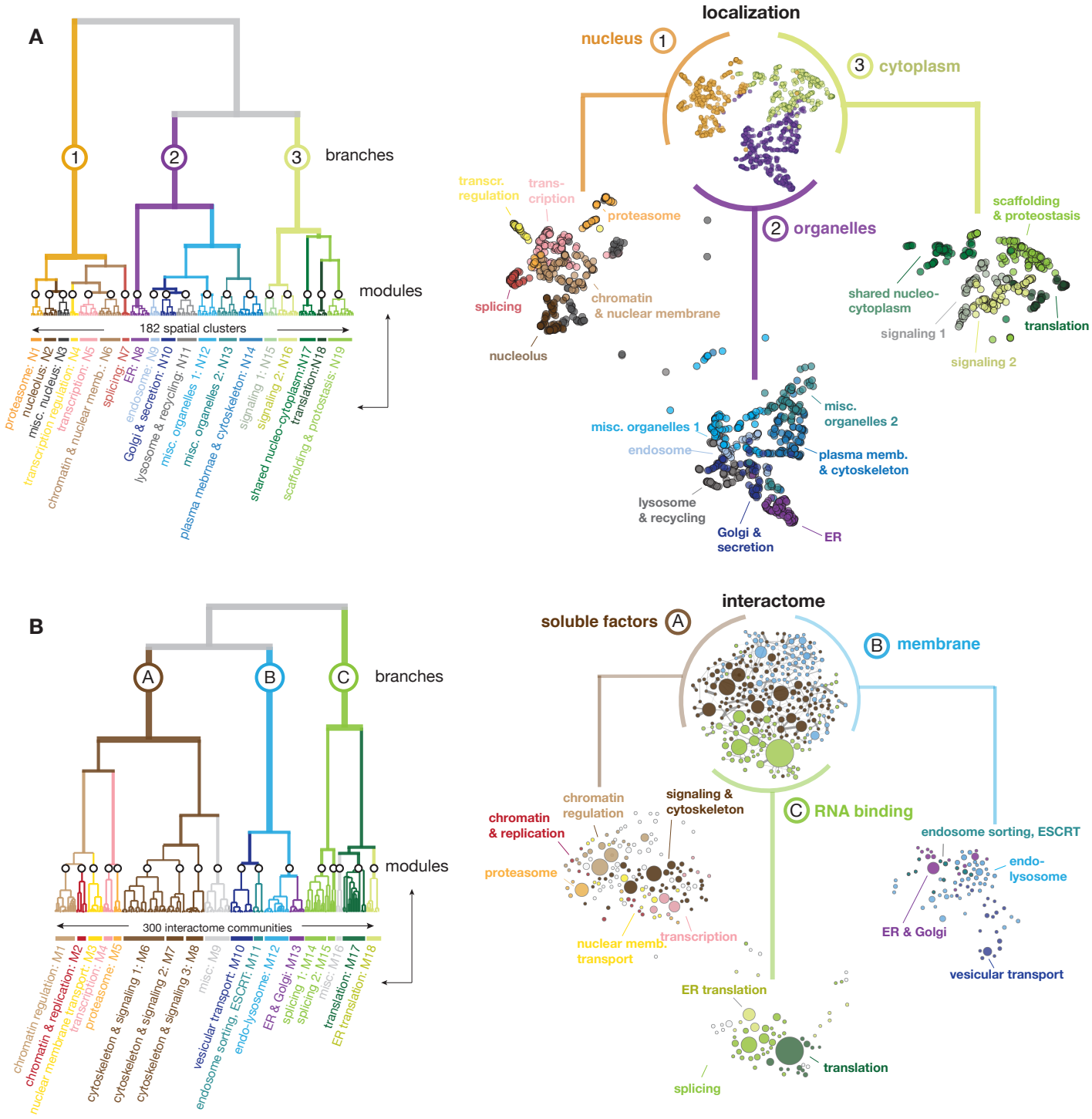

Figure S9: full hierarchical structure of interactome and localization datasets (related to Figure 5). Dendrograms represent the hierarchical relationships connecting **(A)** the full set of protein communities identified in the interactome (see Fig. 2) or **(B)** the full set of high-resolution clusters identified in the image collection (see Fig. 4C). For each dataset, an intermediate layer of hierarchy separates 18-19 modules, while an upper hierarchical layer delineates three separate branches. Modules and branches are annotated on the basis of gene ontology enrichment analysis (see Suppl. Tables 5 & 9). Right-hand panels present the topological arrangement of branches (top) and modules (bottoms) in each dataset, highlighted from the full graph of connections between interaction communities ("interactome", see Fig. 2D) or from the localization UMAP ("localization", see Fig. 4C). The color codes between interactome and localization datasets are not directly comparable (i.e. same colors are not meant to represent the same exact set of proteins). **(C)** The hierarchical structures derived from interactome (left) and localization (right) datasets are compared to the hierarchical structures derived from "scrambled" controls – that is, to the hierarchical structure that is expected by chance given the proteins present in our dataset. Controls are generated by randomly shuffling the membership of each protein between spatial clusters or interaction communities. The number of proteins in each cluster or community was preserved from the original data.

Figure S9

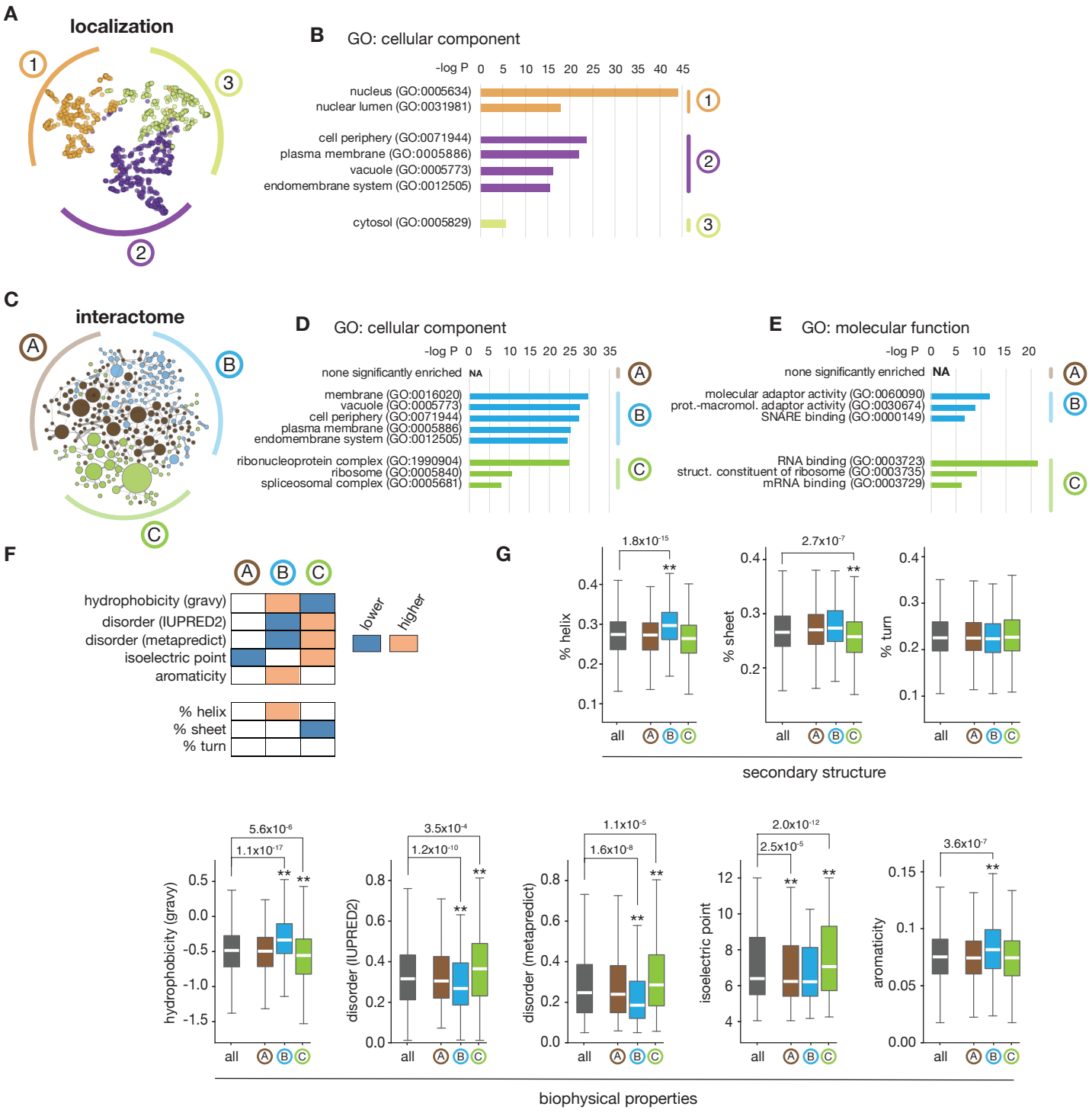

Figure S10: biophysical & ontology analysis of the main branches from interactome and localization hierarchies (related to Figures 5 and S9). (A) The three branches derived from the image-based hierarchy (see Figure S9A). **(B)** Enrichment analysis of GO annotations in the hierarchical branches, testing 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). **(C)** The three branches derived from the interactome hierarchy (see Figure S9B). **(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)** 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. **(G)** Box plots representing the significance testing of biophysical properties described in (F). Boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x inter-quartile ranges. Median is represented by a white line. ** $p < 10^{-3}$ (Student's t-test), exact p-values are shown.

Figure S10

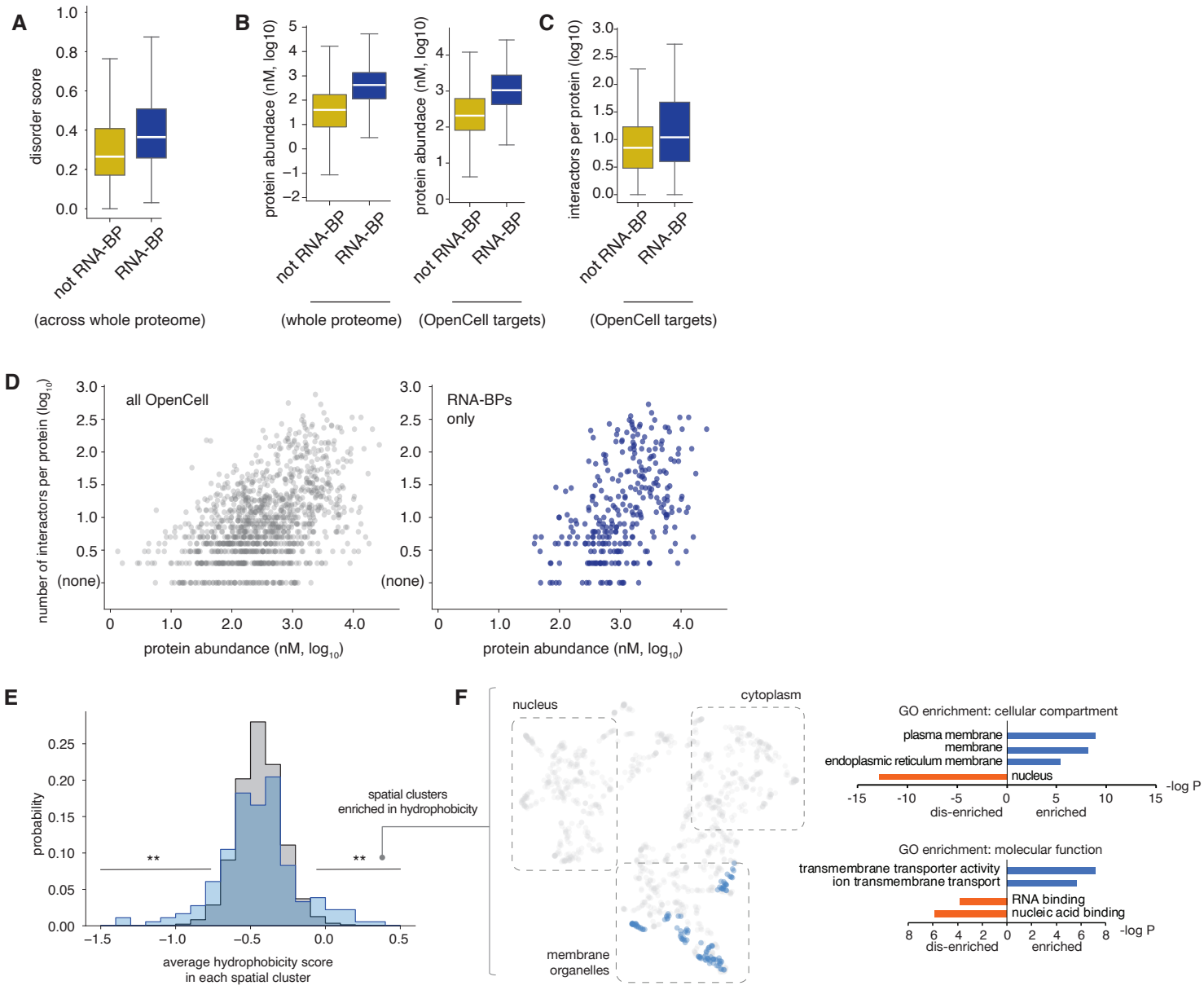

Figure S11: unique properties of RNA-binding proteins (RNA-BPs, related to Figure 5). (A) Distribution of disorder score (IUPRED2) for RNA-BPs vs. non-RNA-BPs across the whole proteome. **(B)** Distribution of protein abundance for RNA-BPs vs non-RNA-BPs across the whole proteome (left) and across OpenCell targets only (right). **(C)** Distribution of number of interactors for RNA-BPs vs non-RNA-BPs across OpenCell targets. **(D)** For each OpenCell target, the number of interactors is plotted as a function of protein abundance. The subset of targets that are RNA-BPs is highlighted on the right-hand panel. Note: for boxplots in (A), (B), (C) and (D), boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5x interquartile range. Median is represented by a white line. **(E)** Distribution of hydrophobicity score (gravy) across spatial clusters, comparing our data to a control in which the membership of proteins across clusters was randomized 1,000 times. Lines indicate parts of the distribution over-represented in our data vs control (**: $p < 2x10^{-3}$, Fisher's exact t-test). **(F)** Distribution of high-hydrophobicity spatial clusters (average hydrophobicity score > -0.1) in the UMAP embedding from Fig. 3D (left), and ontology enrichment analysis of proteins contained in these clusters (right). Enrichment compares to the whole set of OpenCell targets (p-value: Fisher's exact test).

Figure S11

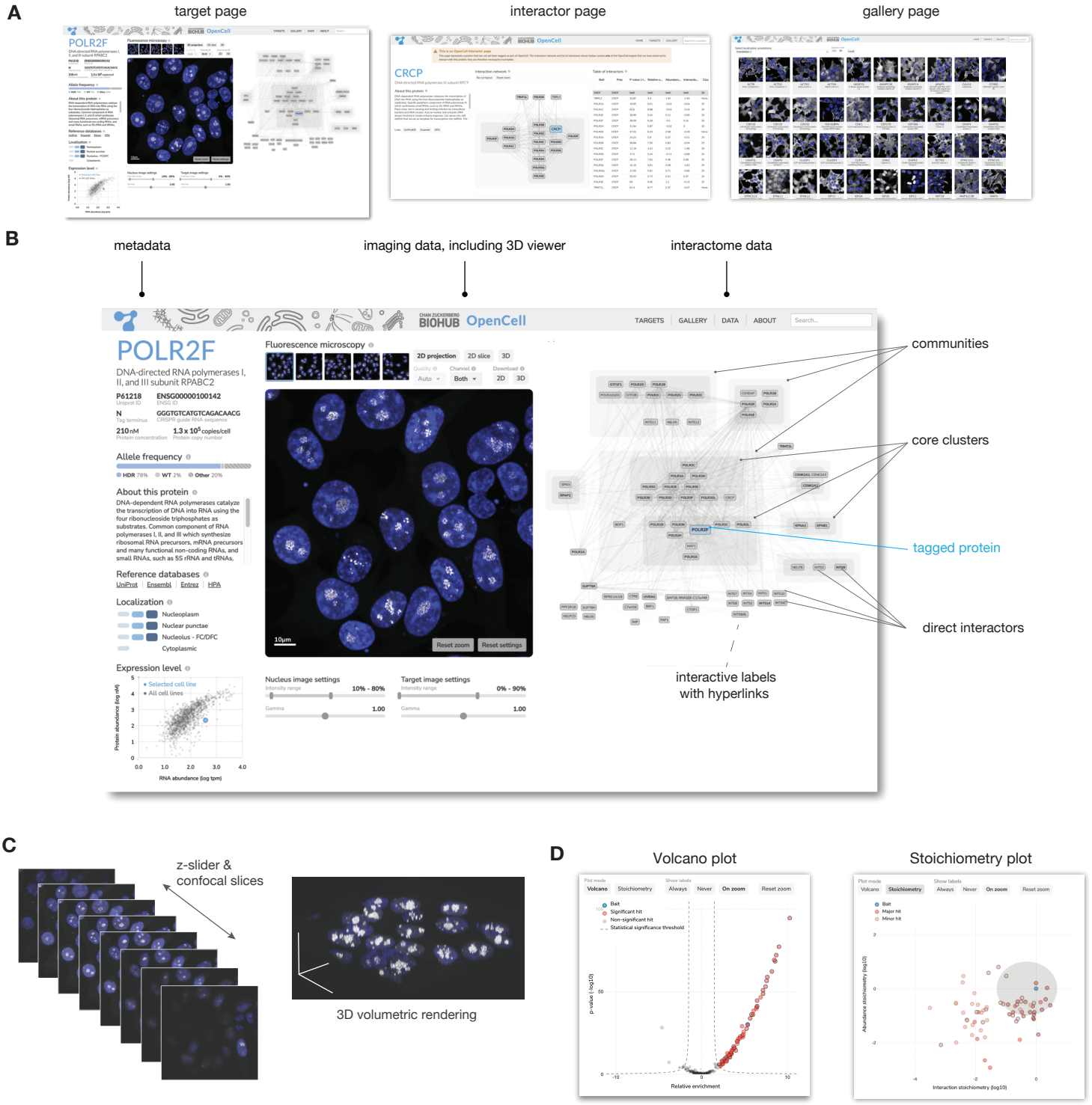

Figure S12: interactive data exploration at opencell.czbiohub.org. (A) The three principal pages of the OpenCell web app. From left to right: the target page, interactor page, and gallery page. **(B)** The target page consists of three columns. The leftmost column contains the functional annotation for 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 S12