UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards an *in silico* Cell

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Yue Qin

Committee in charge:

Professor Trey Ideker, Chair
Professor Jill P. Mesirov, Co-Chair
Professor Garrison W. Cottrell
Professor Prashant Mali
Professor Graham McVicker

2023

The Dissertation of Yue Qin is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

This dissertation is dedicated to my parents, Sujuan Jiang and Xiaohua Qin, for their endless love and support. I am incredibly blessed to be their daughter.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AP-MS	Affinity Purification-Mass Spectrometry
DsiRNA	Dicer-Substrate Small Interfering RNA
eCLIP	Enhanced Ultraviolet Cross-Linking and Immunoprecipitation
GO	Gene Ontology
HPA	Human Protein Atlas
IF	Immunofluorescence
MuSIC	Multi-Scale Integrated Cell
PPI	Protein-Protein Interaction
PRPPA	Pre-Ribosomal RNA Processing Assembly
rRNA	Ribosomal RNA
siRNA	Small Interfering RNA

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Trey Ideker. My training with Trey shaped who I am today. Trey not only taught me how to truly think deep into problems, but also inspired me to look at biology from unique perspectives. The advice I received from him will continue to benefit and influence my entire scientific career. In addition, the opportunities to present my work in various conferences have greatly advanced my ability to communicate science to both my peers and the general public, as well as building my confidence in handling all kinds of situations. Trey's unyielding support is also reflected in various fellowships and scholarships that I am really honored to receive, all of which would not have been possible without Trey. I feel extremely fortunate to have Trey as my mentor.

Next, I would like to thank my dissertation committee for their valuable inputs in my PhD training. During my rotation with Jill Mesirov, her scientific rigor and critical thinking have deeply influenced me and continues to benefit all of my research. I took Gary Cottrell's courses on machine learning during both my undergraduate and graduate studies at UC San Diego. Gary's courses laid the foundation for my dissertation work and enabled me to conduct research at the intersection of machine learning and biology. Prashant Mali has provided valuable feedbacks for my project design. In addition, I really enjoyed and learnt a lot from writing grants with him, including the Cancer Cell Map Initiative and Bridge2AI. Graham McVicker has kindly served on both my qualifying and thesis committees. My research has benefited a lot from his insightful feedbacks.

I would like to also thank all past and current members of Ideker Lab for their unyielding support and friendship. In particular, I would like to thank Jisoo Park, Samson Fong, Jason Kreisberg, Brenton Munson, Tongqiu Jia, Fan Zheng, Jianzhu Ma, Marcus Kelly,

Willy Markuske, Erica Silva, Michael Yu, Justin Huang, Adriana Pitea, Tina Wang, John Lee, Michael Chen, Katherine Licon, Charlotte Marquez, Anton Kratz, Karen Mei, Leah Schaffer, Sophie Liu, Dexter Pratt, Chris Churas, Jing Chen, and Ximena Gonzalez. Their unreserved help and support both in and out of lab are indispensable parts of my PhD. I would also like to thank members of Hannah Carter's lab. In particular, thank you to Michelle Dow for sharing my joys and sorrows during PhD. Furthermore, I would like to thank all my collaborators. It was a great pleasure to work with every one of them. Their valuable advises and unreserved help are no doubt essential for the completion of my dissertation work. In particular, I would like to thank Emma Lundberg. Emma co-advised my dissertation work and has provided me countless support both in research and for my scientific career.

Lastly, I would like to thank my parents Sujuan Jiang and Xiaohua Qin. Even though they do not understand what I am doing, they always offer me their endless love and support. My parents give me the courage to chase my dream, regardless of success or failure.

Chapters 1, 2, 3, and Epilogue, in full, are a refirantted reprint of the material as it appears in *Nature* 2021. Yue Qin, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, Fan Zheng, Michael Chen, Leah V. Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J. Lee, Wei Ouyang, Sophie N. Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F. Kreisberg, Steven P. Gygi, Jianzhu Ma, J. Wade Harper, Gene W. Yeo, Denis L. J. Lafontaine, Emma Lundberg & Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 600, 536–542 (2021). The dissertation author was the primary investigator and author of this paper.

VITA

2017 Bachelor of Science in Biology with Specialization in Bioinformatics, University of California San Diego

2023 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Qin, Y., Huttlin, E. L., Winsnes, C. F., Gosztyla, M. L., Wacheul, L., Kelly, M. R., ... & Ideker, T. (2021). A multi-scale map of cell structure fusing protein images and interactions. *Nature*, 600(7889), 536-542.

Kratz, A.*, Kim, M.*, Kelly, M. R., Zheng, F., Koczor, C. A., Li, J., Ono, K., **Qin, Y.**, Churas, C., Chen, J., Pillich, R. T., Park, J., Modak, M., Collier, R., Licon, K., Pratt, D., Sobol, R. W.†, Krogan N.† & Ideker, T.† A multi-scale map of protein assemblies in the DNA damage response. *In submission*.

CSBC/PS-ON Image Analysis Working Group, Vizcarra, J. C., Burlingame, E. A., Hug, C. B., Goltsev, Y., White, B. S., Tyson, D. R., & Sokolov, A. (2022). A community-based approach to image analysis of cells, tissues and tumors. *Computerized Medical Imaging and Graphics*, 95, 102013.

Dang, J., Tiwari, S. K., Agrawal, K., Hui, H., **Qin, Y.**, & Rana, T. M. (2021). Glial cell diversity and methamphetamine-induced neuroinflammation in human cerebral organoids. *Molecular psychiatry*, 26(4), 1194-1207.

Tiwari, S. K., Dang, J. W., Lin, N., **Qin, Y.**, Wang, S., & Rana, T. M. (2020). Zika virus depletes neural stem cells and evades selective autophagy by suppressing the Fanconi anemia protein FANCC. *EMBO reports*, 21(12), e49183.

Carlin, D. E., Fong, S. H., **Qin, Y.**, Jia, T., Huang, J. K., Bao, B., ... & Ideker, T. (2019). A fast and flexible framework for network-assisted genomic association. *iScience*, 16, 155-161.

Fong, S. H., Carlin, D. E., Ozturk, K., …, **Qin, Y.**, ... & Ideker, T. (2019). Strategies for network GWAS evaluated using classroom crowd science. *Cell Systems*, 8(4), 275-280.

Chao, T.-C.*, Zhang, Q.*, Li, Z., Tiwari, S. K., **Qin, Y.**, Yau, E., ... & Rana, T. M. (2019). The long noncoding RNA HEAL regulates HIV-1 replication through epigenetic regulation of the HIV-1 promoter. *mBio*, 10(5), e02016-19.

Dang, J. W., Tiwari, S. K., **Qin, Y.**, & Rana, T. M. (2019). Genome-wide Integrative Analysis of Zika-Virus-Infected Neuronal Stem Cells Reveals Roles for MicroRNAs in Cell Cycle and Stemness. *Cell Reports*, 27(12), 3618–3628.e5.

Zhang, Q.*, Chao, T.-C.*, Patil, V. S.*, **Qin, Y.**, Tiwari, S. K., Chiou, J., ... & Rana, T. M. (2019). The long noncoding RNA ROCKI regulates inflammatory gene expression. *The EMBO journal*, 38(8), e100041.

Tiwari, S. K.*, Dang, J.*, **Qin, Y.**, Lichinchi, G., Bansal, V., & Rana, T. M. (2017). Zika virus infection reprograms global transcription of host cells to allow sustained infection. *Emerging microbes & infections*, 6(1), 1-10.

Lichinchi, G.*, Zhao, B. S.*, Wu, Y., Lu, Z., **Qin, Y.**, He, C., & Rana, T. M. (2016). Dynamics of human and viral RNA methylation during Zika virus infection. *Cell host & microbe*, 20(5), 666-673.

Dang, J.*, Tiwari, S. K.*, Lichinchi, G., **Qin, Y.**, Patil, V. S., Eroshkin, A. M., & Rana, T. M. (2016). Zika Virus Depletes Neural Progenitors in Human Cerebral Organoids through Activation of the Innate Immune Receptor TLR3. *Cell Stem Cell*, 19(2), 258–265.

* These authors contributed equally to this work. † Co-corresponding authors.

ABSTRACT OF THE DISSERTATION

Towards an *in silico* Cell

by

Yue Qin

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Trey Ideker, Chair
Professor Jill P. Mesirov, Co-Chair

The cell is a multi-scale structure with modular organization across at least four orders of magnitude. Two central approaches for mapping this structure – protein fluorescent imaging and protein biophysical association – each generate extensive datasets, but of distinct qualities and resolutions that are typically treated separately. Here, we integrate immunofluorescence images in the Human Protein Atlas with affinity purification experiments from the BioPlex

resource to create a unified hierarchical map of eukaryotic cell architecture. Integration is achieved by configuring each approach to produce a general measure of protein distance, then calibrating the two measures using machine learning. The map, called the Multi-Scale Integrated Cell (MuSIC 1.0), currently resolves 69 subcellular systems of which approximately half are undocumented. Based on these findings we perform 134 additional affinity purifications, validating close subunit associations for the majority of systems. The map reveals ribosome biogenesis components, including a novel pre-ribosomal RNA processing assembly and numerous accessory factors which we show govern rRNA maturation. The map also elucidates roles for SRRM1 and FAM120C in chromatin and for RPS3A in splicing. By integration across scales, MuSIC substantially increases the mapping resolution obtained from imaging while giving protein interactions a spatial dimension, paving the way to incorporate diverse types of molecular data to create proteome-wide cell maps.

INTRODUCTION

Eukaryotic cells consist of a collection of large components, such as organelles, which recursively factor into ever smaller components, such as condensates and protein complexes, forming an intricate multi-scale structure (Schaffer & Ideker, 2021). Deciphering this multi-scale structure and its relation to function is one of the ultimate goals of cell biology (Alberts, 1998).

Two fundamental techniques for mapping subcellular structure are protein imaging and biophysical association, each of which has been extensively automated in recent years (Aebersold & Mann, 2016; Lundberg & Borner, 2019; Qin et al., 2021; Specht et al., 2017). In particular, advances in confocal microscopy and immunofluorescence (IF) imaging have made it possible to rapidly scan the spatial distribution of proteins *in situ* within single cells (Chen et al., 2019; Mori & Cardiff, 2016; Stadler et al., 2013). By combining these techniques with a library of antibodies, the Human Protein Atlas (HPA) has embarked on a major effort to systematically position human proteins into 33 different organelles and subcellular structures (Colwill et al., 2011; Thul et al., 2017). As a parallel means to map cellular architecture, mass spectrometry (MS) has been powerfully combined with methods such as affinity purification (AP-MS) (Lee et al., 2017) and proximity-dependent labeling (Gingras et al., 2019; Kalocsay, 2019; Rhee et al., 2013; Varnaitė & MacNeill, 2016; Youn et al., 2018) to enable rapid measurement of physical protein-protein associations. Using AP-MS, the BioPlex project is generating comprehensive maps of physical interactions for most human proteins across a variety of cell types (Huttlin et al., 2017).

Given these growing resources, a key question is how protein imaging and biophysical association should be properly combined to inform cell structure. In this regard, we reasoned that the two platforms provide complementary measures of protein location, albeit of vastly different characters. Imaging provides protein locations relative to cellular landmarks such as the nucleus,

while biophysical association positions proteins relative to other nearby proteins. In both cases, such positioning has become increasingly quantitative and precise in recent years, due in part to the ability of machine learning systems to recognize complex patterns in data (Cho et al., 2018; Grover & Leskovec, 2016; Ounkomol et al., 2018; Ouyang et al., 2019; Sullivan et al., 2018; Weigert et al., 2018).

Here, we demonstrate a machine learning approach by which protein imaging and biophysical association are integrated to create a unified map of human subcellular components, which we call the Multi-Scale Integrated Cell (MuSIC). First, we use neural networks to project human proteins into a small number of dimensions based on imaging or biophysical association data. Once protein coordinates have been determined for each platform, pairwise distances among proteins are calibrated and combined to reveal a collection of protein assemblies at different scales, from the very small (<50 nm) to the very large (>1 μm). MuSIC paves the way for building reference maps of cell biology by bridging individual technology platforms, connecting molecules to molecular assemblies to organelles to cells.

REFERENCES

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620), 347–355.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3), 291–294.

Chen, X., Zhang, D., Su, N., Bao, B., Xie, X., Zuo, F., Yang, L., Wang, H., Jiang, L., Lin, Q., Fang, M., Li, N., Hua, X., Chen, Z., Bao, C., Xu, J., Du, W., Zhang, L., Zhao, Y., … Yang, Y. (2019). Visualizing RNA dynamics in live cells with bright and stable fluorescent RNAs. *Nature Biotechnology*, 37(11), 1287–1293.

Cho, H., Berger, B., & Peng, J. (2018). Generalizable and Scalable Visualization of Single-Cell Data Using Neural Networks. *Cell Systems*, 7(2), 185–191.e4.

Colwill, K., Renewable Protein Binder Working Group, & Gräslund, S. (2011). A roadmap to generate renewable protein binders to the human proteome. *Nature Methods*, 8(7), 551–558.

Gingras, A.-C., Abe, K. T., & Raught, B. (2019). Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Current Opinion in Chemical Biology*, 48, 44–54.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. *International Conference on Knowledge Discovery & Data Mining*, 2016, 855–864.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655), 505–509.

Kalocsay, M. (2019). APEX Peroxidase-Catalyzed Proximity Labeling and Multiplexed Quantitative Proteomics. *Methods in Molecular Biology*, 2008, 41–55.

Lee, C.-M., Adamchek, C., Feke, A., Nusinow, D. A., & Gendron, J. M. (2017). Mapping Protein-Protein Interactions Using Affinity Purification and Mass Spectrometry. *Methods in Molecular Biology*, 1610, 231–249.

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

Mori, H., & Cardiff, R. D. (2016). Methods of Immunohistochemistry and Immunofluorescence: Converting Invisible to Visible. *Methods in Molecular Biology*, 1458, 1–12.

Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F., & Johnson, G. R. (2018). Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11), 917–920.

Ouyang, W., Winsnes, C. F., Hjelmare, M., Cesnik, A. J., Åkesson, L., Xu, H., Sullivan, D. P., Dai, S., Lan, J., Jinmo, P., Galib, S. M., Henkel, C., Hwang, K., Poplavskiy, D., Tunguz, B., Wolfinger, R. D., Gu, Y., Li, C., Xie, J., … Lundberg, E. (2019). Analysis of the Human Protein Atlas Image Classification competition. *Nature Methods*, 16(12), 1254–1261.

Qin, W., Cho, K. F., Cavanagh, P. E., & Ting, A. Y. (2021). Deciphering molecular interactions by proximity labeling. *Nature Methods*, 18(2), 133–143.

Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., & Ting, A. Y. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, 339(6125), 1328–1331.

Schaffer, L. V., & Ideker, T. (2021). Mapping the multiscale structure of biological systems. *Cell Systems* (Vol. 12, Issue 6, pp. 622–635).

Specht, E. A., Braselmann, E., & Palmer, A. E. (2017). A Critical and Comparative Review of Fluorescent Tools for Live-Cell Imaging. *Annual Review of Physiology*, 79, 93–117.

Stadler, C., Rexhepaj, E., Singan, V. R., Murphy, R. F., Pepperkok, R., Uhlén, M., Simpson, J. C., & Lundberg, E. (2013). Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nature Methods*, 10(4), 315–323.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., & Lundberg, E. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, 36(9), 820–828.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, 356(6340).

Varnaitė, R., & MacNeill, S. A. (2016). Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics*, 16(19), 2503–2518.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., Rocha-Martins, M., Segovia-Miranda, F., Norden, C., Henriques, R., Zerial, M., Solimena, M., Rink, J., Tomancak, P., Royer, L., … Myers, E. W. (2018). Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature Methods*, 15(12), 1090–1097.

Youn, J.-Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., Bagci, H., Rathod, B., MacLeod, G., Eng, S. W. M., Angers, S., Morris, Q., Fabian, M., Côté, J.-F., & Gingras, A.-C. (2018). High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell*, 69(3), 517–532.e11.

CHAPTER 1
Computational Pipeline for Cell Map Construction

1.1 Results

Protein position and distance, two ways

We assembled a matched dataset of IF images from HPA (Thul et al., 2017) and AP-MS data from BioPlex (**Figure 1.1a**) (Huttlin et al., 2017). Both resources are partially based on human embryonic kidney (HEK293-derived) cells, leading us to 661 proteins with compatible imaging (1,451 images including replicates, **Figure S1.1a-c**) and biophysical association data (291 proteins affinity-tagged as 'baits', the remaining 370 arising as interacting 'preys', **Table S1.1**). These proteins covered a wide distribution of subcellular locations, as previously annotated by HPA, which was very similar to the distribution seen for all human proteins (**Figure S1.1d**). Other proteins in HPA and BioPlex were measured in differing cell types that did not align across projects; thus, we focused on the 661 proteins in the common HEK293-derived context for prototyping the new approach.

We next used deep neural networks to embed each protein based on its IF and AP-MS data (**Figure 1.1b**). An embedding is a low-dimensional representation of a complex input, in which each data point (here a protein) is assigned a coordinate in the newly reduced dimensions. Much machine learning research has been concerned with creating a good embedding, in which similar inputs (in this case proteins with similar subcellular distributions or interaction neighborhoods) are placed close together in the embedded space (Goodfellow et al., 2016). For image embedding we used DenseNet (Ouyang et al., 2019), a deep convolutional neural network shown to have superior performance in capturing protein subcellular locations relative to counter-stained cellular landmarks (**Figure S1.2a-c**). Similarly, the node2vec deep neural network (Grover & Leskovec,

2016) was used to embed each protein based on its extended interaction neighborhood in the AP-MS data (**Figure 1.1b**, **Figure S1.2d-g**).

We then computed protein-protein distances (cosine distance) for all pairs of proteins, separately in the IF and AP-MS embeddings. We found that the closest protein pairs measured by one technique were significantly enriched for those measured as close by the other, demonstrating that despite their differences, the two measurement types share significant information (**Figure 1.1c, d**). As a means of calibrating distance in the embeddings to physical distance in cells, we sampled the literature to assemble a reference set of ten subcellular components with known physical sizes, from protein complexes of <20 nm to organelles >1 μm in diameter (**Table S1.2**). The size of each of these ten components strongly correlated with its number of protein species documented in the Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium, 2019), suggesting a general conversion from the number of protein species to diameter, in nanometers, of a subcellular component (**Figure 1.1e, Calibration Function**). We used this function to label pairs of proteins with a curated physical distance, based on the size of the smallest GO component to which that pair was assigned (**Methods**). With these curated distances as training labels, we taught a supervised machine learning model (random forest regression) to estimate the pairwise distance of any protein pair directly from its coordinates in the IF and AP-MS embeddings (**Figure 1.1f, g**).

A multi-scale map of subcellular systems

We analyzed all estimated distances among the 661 proteins to identify communities of proteins in close mutual proximity, suggesting distinct subcellular components. Protein communities were identified at multiple resolutions, starting with those that form at the smallest protein-protein distances then progressively relaxing the distance threshold (multi-scale

community detection (Fortunato, 2009; Kramer et al., 2014), (**Figure S1.3, Methods**). Communities at smaller distances were contained, in full or in part, inside larger communities as the threshold was relaxed, yielding a structural hierarchy. The sensitivity of community detection was tuned for best concordance with two independent datasets not used elsewhere in our study: a separate collection of protein interactions reported in the Human Cell Map (Go et al., 2021) using proximity biotinylation, also in HEK293 cells, and patterns of gene co-essentiality observed in the Cancer Cell Dependency Map (Hart et al., 2015; Meyers et al., 2017) (**Methods**). The agreement with the independent datasets was significant over a wide range of community detection parameters and was seen for both small and large communities (**Figure S1.4**). The final hierarchy, which we call the Multi-Scale Integrated Cell (MuSIC 1.0), contained 69 protein communities representing putative subcellular systems organized by 87 hierarchical containment relationships (**Figure 1.2**, **Table S1.3**). Sixteen systems were contained within two or more larger ones, suggesting pleiotropic roles or activity at multiple subcellular locations. To elucidate the biological roles of each system, we aligned the MuSIC hierarchy to the equivalent literature-curated hierarchy of cellular components provided by GO (**Methods**). Approximately 46% of systems had significant overlap with GO; the remaining 54% were annotated as putatively novel (**Figure 1.2**).

The physical sizes of MuSIC systems were estimated from their pairwise protein distances (**Figure 1.2, Methods**) and compared to the known diameters of nine well-characterized cellular components, independent from those used for earlier calibration (**Figure 1.3a, Table S1.4**). One of these was the pre-catalytic spliceosome, for which support from both IF and AP-MS modalities (**Figure 1.3b-e**) had induced a protein community at a resolution of around 48 nm (95% prediction interval [26, 90]), in agreement with its published size of 42 nm (Charenton et al., 2019; Deckert et al., 2006) (**Figure 1.3f, g**). Within this community, the analysis further resolved two smaller

systems representing the U1 and U2 subunits (U1: 8 nm, 95% prediction interval [4, 15]; U2: 33 nm, 95% prediction interval [17, 61]), again in agreement with the subunit arrangement and distances measured by cryo-electron microscopy (Protein Data Bank code 6QX9, **Figure 1.3g**) (Charenton et al., 2019). For all nine components, the estimated diameters were very close to the actual measurements from literature, validating that MuSIC captures and accurately sizes biological systems across a wide range of physical scales (**Figure 1.3a**).

1.2 Methods

Data sources

IF confocal images (63x oil immersion, NA 1.4) interrogating protein locations in the HEK293 cell line were downloaded from the HPA Cell Atlas (Ouyang et al., 2019; Sullivan et al., 2018; Thul et al., 2017). Physical protein interactions detected by AP-MS in the HEK293T cell line were downloaded from the BioPlex 2.0 protein interaction database (Huttlin et al., 2017). We focused our study on the intersection of these IF and AP-MS datasets, by selecting images of immunofluorescent proteins that had been affinity-tagged as baits or detected as preys in BioPlex. IF data: Each image had four channels, one for the protein of interest and three for reference markers of the nucleus, microtubules and endoplasmic reticulum. Images involving antibodies that targeted more than one protein, or that lacked annotated cellular localizations, were removed. The final imaging dataset contained 1,451 images covering 661 proteins and 726 antibodies, corresponding to a range of 2-6 images per protein (**Figure S1.1a-c**). We observed that the majority (27/33) of subcellular localizations tracked by the HPA were covered by the selected images (**Figure S1.1d**), suggesting that the data used in this study are representative. AP-MS data: The AP-MS dataset covered this same set of 661 proteins with 281 physical protein interactions observed among the 661 proteins; we also retained the entire BioPlex 2.0 network of 10,961

proteins and 56,553 protein interactions, which provided significant information about the extended network neighborhoods of the 661 proteins (used in **Data embeddings** below).

Data embeddings

IF data: Three different image embedding methods (DenseNet, Subcellular Location Features, Paired Cell Inpainting) were compared for their ability to enrich for the 281 BioPlex protein-protein interactions among the 661 proteins. Based on this performance comparison (**Figure S1.2c**), DenseNet was selected for all subsequent analyses. DenseNet (Densely Connected Convolutional Networks) is a recently introduced convolutional neural network for general object recognition in digital images, which has shown to achieve very high performance at a range of tasks and requires few parameters (Huang et al., 2016). We used the DenseNet-121 model optimized for analysis of HPA images as previously described (Ouyang et al., 2019). The embedding for each HPA input image was taken as the penultimate layer of resulting neuron values, yielding a 1024-dimension feature vector henceforth called $\mathbf{e_{IF}}$ (**Figure S1.2a, b**). AP-MS data: Likewise, three different AP-MS protein network embedding methods (node2vec, node properties, random walk with restart) were compared for their ability to enrich for the 281 protein pairs with highest cosine similarity calculated in the IF embedding (the number 281 was selected to match the number of protein-protein interactions used above for IF data analysis). Based on this performance comparison (**Figure S1.2g**), node2vec was selected for all subsequent analyses. Node2vec inputs an interaction network (protein-protein interactions) and uses a deep neural network model to learn feature representations of the interaction neighborhood surrounding each node (Grover & Leskovec, 2016). Here, we ran node2vec on all available BioPlex AP-MS data (n = 10,961 nodes, m = 56,553 edges, see **Data sources** above) with parameter settings $p = 2$, $q = 1$

to generate a 1024-dimension feature embedding for each protein, henceforth called $\mathbf{e}_{\text{AP-MS}}$ (**Figure S1.2d, e**).

Converting number of proteins in a component to physical size

Of the 1602 human cellular components documented in the Gene Ontology (The Gene Ontology Consortium, 2019) (GO, 25.9.2018 release), only a small fraction have been experimentally well characterized in terms of physical size measurements. However, all have been annotated with a set of associated proteins. To explore the correspondence between the number of proteins assigned to a component C and its physical diameter D, we curated a list of ten components for which the approximate physical sizes are known (see **Table S1.2** for sources of experimental data). Examination of their C and D values indicated a linear relationship between $log_{10}C$ and $log_{10}D$ that was well-modeled ($R^2 = 0.89$, **Figure 1.1e**) by the function:

$$log_{10}D = 1.05 \times log_{10}C - 0.14 \quad (1.1)$$

This function was subsequently used to convert from C to D for any cellular component.

Random forest prediction of protein distances

Among the 661 proteins under study, 602 had specific GO annotations (i.e., other than the root term; GO gene-to-term annotations based on HPA images were removed to avoid circularity). For each pair of these proteins p_1 and p_2, we measured $C(p_1, p_2)$, the number of proteins in the smallest GO cellular component to which both are annotated. This quantity was converted to a pairwise diameter D according to eqn. (1.1), as well as its opposite, pairwise proximity P:

$$P(p_1, p_2) = -log_{10}D(p_1, p_2) \quad (1.2)$$

Random forest regression models were then constructed (Python scikit-learn package, **Figure 1.1f**) (Pedregosa et al., 2011) to predict P from a set of input features derived from the IF and AP-MS

data embeddings, as follows. Input features: Each protein pair was associated with the following features from the IF embeddings (see **Data embeddings**):

Element-wise absolute difference ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

$= < |e_1(p_1) - e_1(p_2)|, |e_2(p_1) - e_2(p_2)|, \ldots , |e_{1024}(p_1) - e_{1024}(p_2)| >$

Manhattan distance ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

Euclidean distance ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

Cosine similarity ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

Pearson correlation ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

Spearman correlation ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

Kendall correlation ($\mathbf{e_{IF}}(p_1)$, $\mathbf{e_{IF}}(p_2)$)

A parallel set of features was constructed from the AP-MS embeddings of the two proteins. Because every protein had multiple images (ranging from two to six, see **Data sources** above), six different training sets were generated, each randomly selecting one of the available images per protein while requiring each image to be used at least once. For each training set, we trained random forest regressors using five-fold cross validation (**Figure 1.1g**). The final predicted proximity of each protein pair, $\hat{P}(p_1, p_2)$, was obtained by averaging the six random forest predictions. The set of $\hat{P}$ for all protein pairs is henceforth called the integrated *protein proximity network*. As a negative control, 1024-dimension random vectors sampled from a normal distribution were generated and used in place of $\mathbf{e_{IF}}$ and $\mathbf{e_{AP\text{-}MS}}$ (**Figure 1.1g**).

Pan-resolution community detection

The integrated protein proximity network was analyzed to detect distinct communities of proteins using the Clique eXtracted Ontology algorithm (Kramer et al., 2014) (CliXO v1.0, https://github.com/fanzheng10/CliXO-1.0). CliXO finds the maximal cliques in a weighted

network while progressively decreasing the threshold on edge weights. Lower thresholds yield cliques that may contain, in full or in part, cliques identified at higher thresholds, resulting in a hierarchy of communities interrelated by community containment relations (**Figure S1.3, Figure 1.3f**). CliXO has four parameters that control the depth (α), width (β), modularity (m) and modularity significance (z) of the community hierarchy (**Figure S1.4a**). We swept through 500 different combinations of these four parameters to obtain a pool of hierarchies. All communities, representing putative cellular systems, were required to have at least four proteins to further ensure quality and validity.

To select an optimal model, each hierarchy in the pool was evaluated based on its concordance with two independent datasets, the Human Cell Map (Go et al., 2021) and the Cancer Cell Dependency Map (Meyers et al., 2017) v18Q2, which were not used elsewhere in this study (**Figure S1.4b, c**). From the Human Cell Map, 178 protein-protein interactions detected in HEK293 cells with high-confidence were obtained (FDR $\leq$ 0.01, covering 293 proteins in MuSIC map). From the DepMap, we selected 14,588 "co-essential" gene pairs, for which the CRISPR gene disruptions of the two proteins led to highly similar (cosine similarity) patterns in fitness profiles across the panel of 730 DepMap cell lines (covering 651 proteins in MuSIC map). For each hierarchy, we recorded the number of Human Cell Map protein-protein interactions (x) or DepMap co-essential protein pairs (y) that were covered by systems that were significantly enriched (FDR $\leq$ 0.1) for those interactions (**Figure S1.4b, c**). The hierarchy having the highest number ($x \times y$) was selected for further study; among the several ties, we selected the hierarchy with the least number of systems (i.e., guided by the principle of parsimony).

The hierarchical structure was further matured by assigning additional hierarchical parent-child containment relations between pairs of systems having a containment index $\geq$0.75 and

removing redundant systems having Jaccard index ≥0.9 with parent systems. The containment and Jaccard indexes between two systems s_1 and s_2 were calculated based on the following formulae:

$$Containment(s_1, s_2) = \frac{|s_1 \cap s_2|}{\min(|s_1|, |s_2|)} \quad (1.3)$$

$$Jaccard(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \quad (1.4)$$

with $|s|$ representing the number of proteins in a system. Redundant systems having only one protein difference from a parent system were also removed. While removing a system, the integrity of the hierarchical structure was maintained by adding containment relations from all children of that system to all of its parents.

We also analyzed the integrated protein proximity network using Louvain clustering (Blondel et al., 2008; Fortunato, 2009) (https://github.com/vtraag/louvain-igraph, v0.6.1) which partitions network nodes into a set of distinct clusters. We ran Louvain over 1,000 instances, each initialized with an independent random state, and selected the instance that maximized the global modularity as output by the algorithm. The partition used for the final MuSIC model included just two clusters which were strongly enriched for proteins with known subcellular locations in the cytoplasm versus nucleus, respectively (**Figure 1.2**). These clusters were added as parent systems of the top-layer systems found by CliXO (described above). In particular, a CliXO system was added as a child of a Louvain system if at least 50% of its proteins were in the Louvain system(s).

Systems in the MuSIC map were annotated by synthesizing prior literature with our own biological knowledge and reasoning. For systems highlighted in texts, we introduced a further quality control step in which we manually inspected the corresponding IF images and raw AP-MS spectra. This process prompted us to remove the protein RPL6 out of concerns for antibody correctness.

To label MuSIC systems as "known" or "putative" (**Figure 1.2**), MuSIC was compared to the cellular component branch of GO, filtered for the proteins under study. A system *s* was considered "known" if there existed a GO term *T* that was significantly enriched (FDR ≤ 0.001, hypergeometric statistic) for proteins in *s* with $Jaccard(s,T) \geq 0.4$, or if $Jaccard(s,T) = 1$, representing perfect agreement regardless of significance.

Estimation of system diameter

The diameter D_s of each MuSIC system *s* (**Figure 1.2**, size ladder) was estimated from the collection of diameters predicted from protein pairs in *s*:

$$D_s = 1.8 \times median_{p_1,p_2 \in s} D(p_1,p_2) \quad (1.5)$$

$$D(p_1,p_2) = 10^{-\hat{P}(p_1,p_2)} \quad (1.6)$$

Eqn. (1.6) is the inverse of eqn. (1.2). Note that the median predicted diameter of all protein pairs in *s* underestimates the true system diameter whenever *s* contains one or more subsystems. This underestimation occurs because some protein pairs are also assigned to common smaller subsystems. Here we found that robust estimates of D_s could be obtained using the factor $1.8 \times$ *median* in eqn. (1.5) above. This observation was made through an analysis of cellular components in GO, independent from MuSIC.

The 95% prediction interval for D_s was estimated using actual size measurements from the literature for nine components (**Figure 1.3a**, **Table S1.4**), as follows:

$$[log_{10}D_s - t \times SE, log_{10}D_s + t \times SE] \quad (1.7)$$

$$SE = \sqrt{\frac{\sum(log_{10}D_s - log_{10}D_{literature})^2}{n-1}} \sqrt{1 + \frac{1}{n}} \quad (1.8)$$

with *t* determined by the Student's *t*-distribution ($t = 2.306$ with $df = n - 1$, $n = 9$ components); SE is the standard error between predicted ($log_{10}D_s$) and measured sizes ($log_{10}D_{literature}$).

1.3 Figures

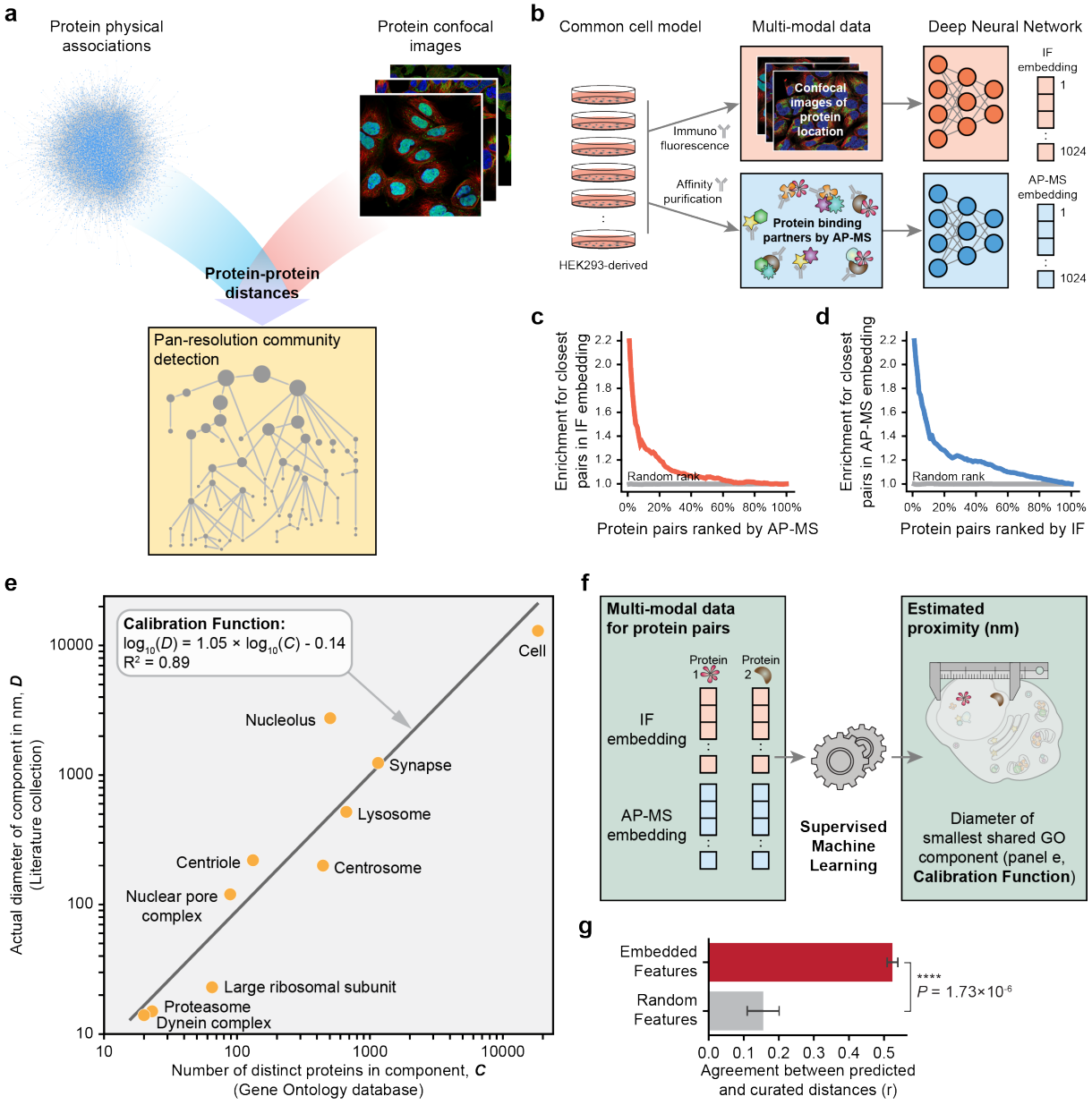


Figure 1.1: Fusing protein distances from protein image and interaction data. a, Overview of study. **b,** Generating an embedding for each protein from IF images and AP-MS data, respectively. **c, d,** Protein pairs ranked by similarity in AP-MS embedding enrich for the most similar protein pairs in IF (**c**), and vice versa (**d**). **e,** Calibrating physical diameter, *D*, of subcellular components against the number of proteins, *C*, assigned to the corresponding GO terms. **f,** Supervised model (random forest) estimates physical proximity (nm) of all pairs of proteins from their IF and AP-MS embeddings. **g,** Performance of model in recovering protein-protein distances in GO in five-fold cross validation (red, Pearson's *r*). Error bars show standard deviation in cross validation. Equivalent calculation for random feature sets (gray).

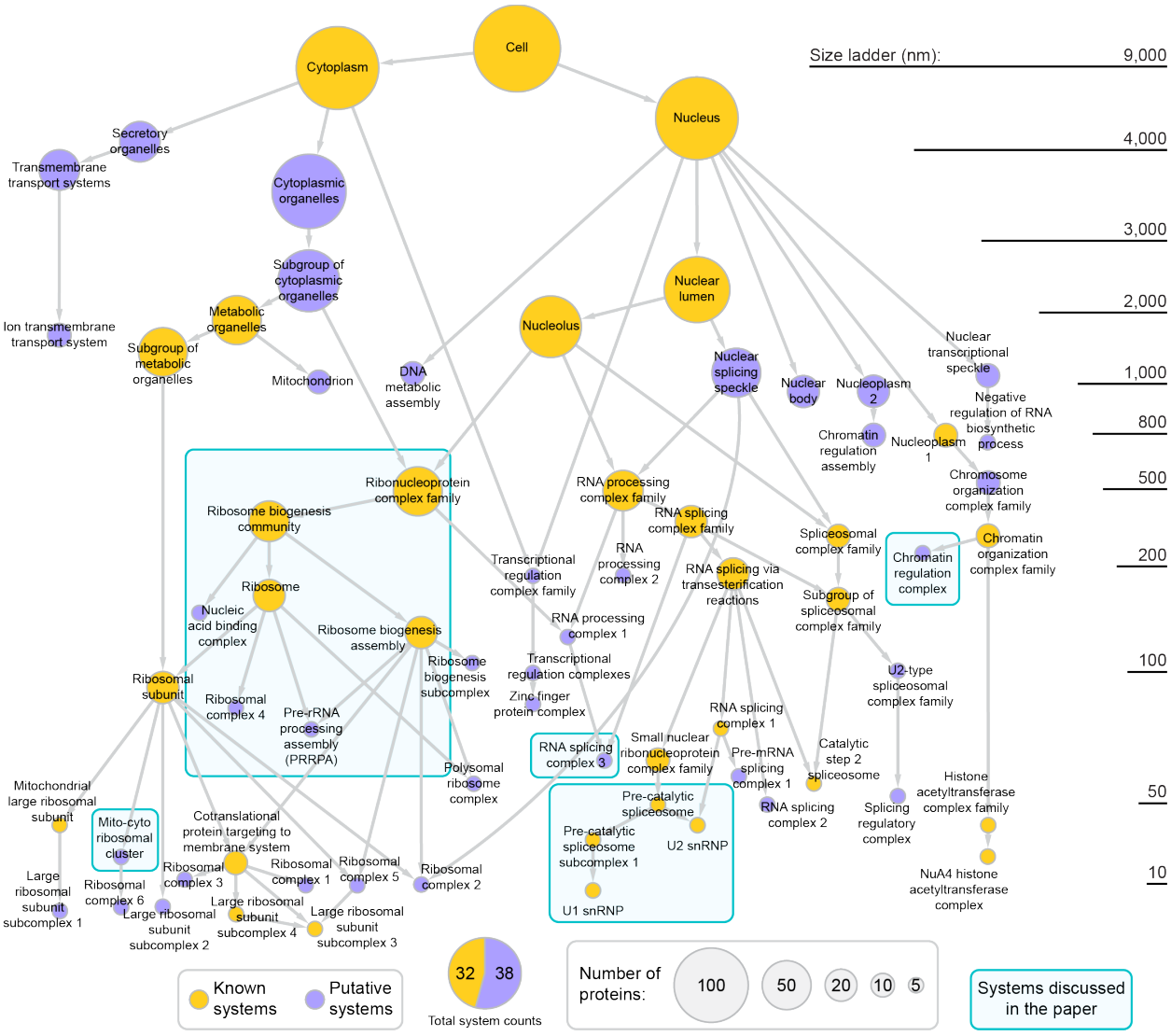
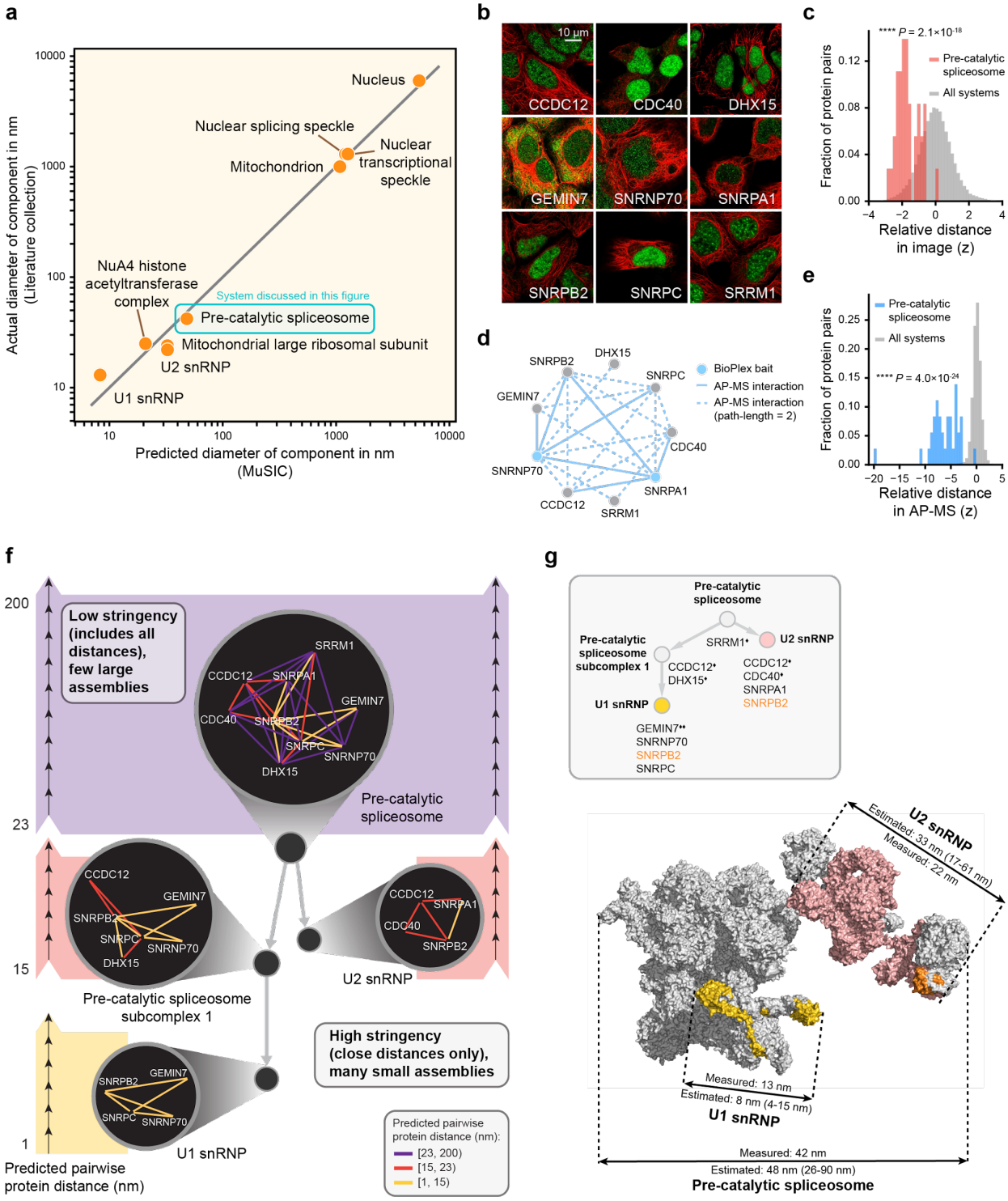


Figure 1.2: The Multi-scale Integrated Cell. MuSIC hierarchy, with nodes representing systems and arrows indicating containment of the lower system by the upper. Pie chart shows number of known (gold) or putative novel (purple) systems. Size of each circle is based on the number of proteins in the system. Systems highlighted by teal box are detailed in **Chapter 3**. The relative elevation of each system in the layout is determined based on the predicted diameter of the system in MuSIC (size ladder).

Figure 1.3: MuSIC predicts diameters of subcellular components. a, Comparison of predicted to actual diameter for components detailed in the literature and not used for calibration. **b,** Immunostained pre-catalytic spliceosome proteins (green, white text) versus cytoskeleton counterstain (red). **c,** Corresponding IF protein-protein distances shown as *z*-scores (red) against distribution of IF distances for all protein pairs (gray). **d,** BioPlex physical interaction network for proteins in pre-catalytic spliceosome. AP-MS interaction (path-length = 2) indicates protein pairs that interact with common affinity-tagged bait(s) outside the complex. **e,** Histogram as in (**c**), showing AP-MS rather than IF data. **f**, As the distance threshold increases (bottom to top), strongly associated protein systems are detected first and then subsequently expand to include proteins with moderate-to-weak associations. Each circle indicates a protein system, and edge colors (yellow, red, purple) indicate decreasing stringencies of association. **g**, Model of the pre-catalytic spliceosome 3D structure (left, Protein Data Bank code 6QX9) (Charenton et al., 2019) next to capture of pre-catalytic spliceosome in MuSIC (right). Proteins assigned the same colors across both types of maps. ♦, pre-catalytic spliceosome proteins (Deckert et al., 2006) captured by MuSIC but not included in structural model. ♦♦, protein important for the assembly of snRNPs. Note that SNRPB2 (orange) affiliates with both U1 and U2 subunits in MuSIC, as suggested by a previous study (Williams & Hall, 2011), whereas it is included in U2 subunit by the structural model.


a
Actual diameter of component in nm (Literature collection)
Predicted diameter of component in nm (MuSIC)
Nucleus
Nuclear splicing speckle
Nuclear transcriptional speckle
Mitochondrion
NuA4 histone acetyltransferase complex
System discussed in this figure
Pre-catalytic spliceosome
Mitochondrial large ribosomal subunit
U2 snRNP
U1 snRNP
b
10 μm
CCDC12
CDC40
DHX15
GEMIN7
SNRNP70
SNRPA1
SNRPB2
SNRPC
SRRM1
c
**** P = 2.1×10-18
Pre-catalytic spliceosome
All systems
Fraction of protein pairs
Relative distance in image (z)
d
BioPlex bait
AP-MS interaction
AP-MS interaction (path-length = 2)
e
Pre-catalytic spliceosome
All systems
**** P = 4.0×10-24
Fraction of protein pairs
Relative distance in AP-MS (z)
f
Low stringency (includes all distances), few large assemblies
Pre-catalytic spliceosome
Pre-catalytic spliceosome subcomplex 1
U2 snRNP
U1 snRNP
High stringency (close distances only), many small assemblies
Predicted pairwise protein distance (nm):
[23, 200)
[15, 23)
[1, 15)
Predicted pairwise protein distance (nm)
g
Pre-catalytic spliceosome
Pre-catalytic spliceosome subcomplex 1
U2 snRNP
U1 snRNP
SRRM1•
CCDC12•
DHX15•
CCDC12•
CDC40•
SNRPA1
SNRPB2
GEMIN7••
SNRNP70
SNRPB2
SNRPC
U2 snRNP
Estimated: 33 nm (17-61 nm)
Measured: 22 nm
Measured: 13 nm
Estimated: 8 nm (4-15 nm)
U1 snRNP
Measured: 42 nm
Estimated: 48 nm (26-90 nm)
Pre-catalytic spliceosome

1.4 Supplementary Figures

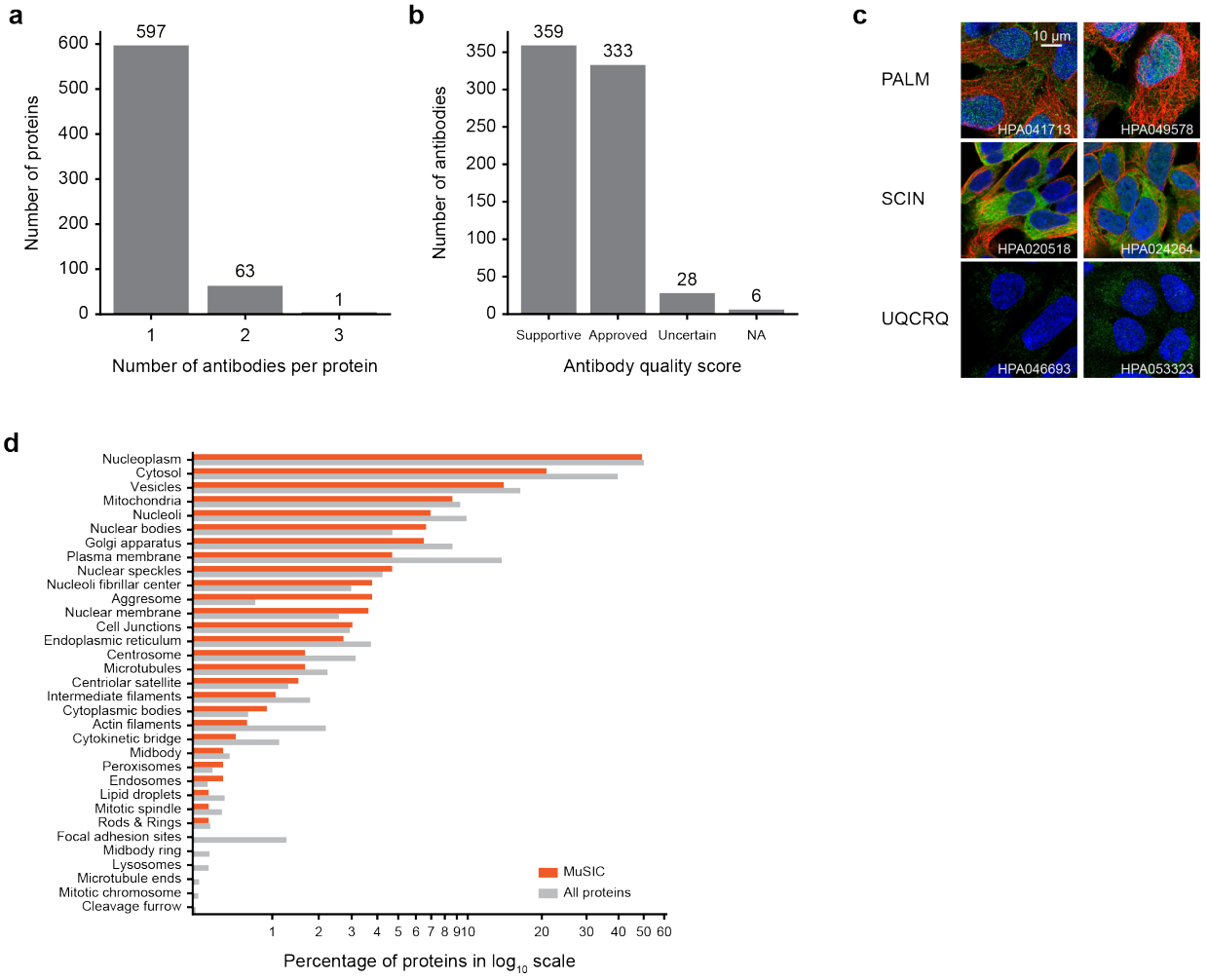
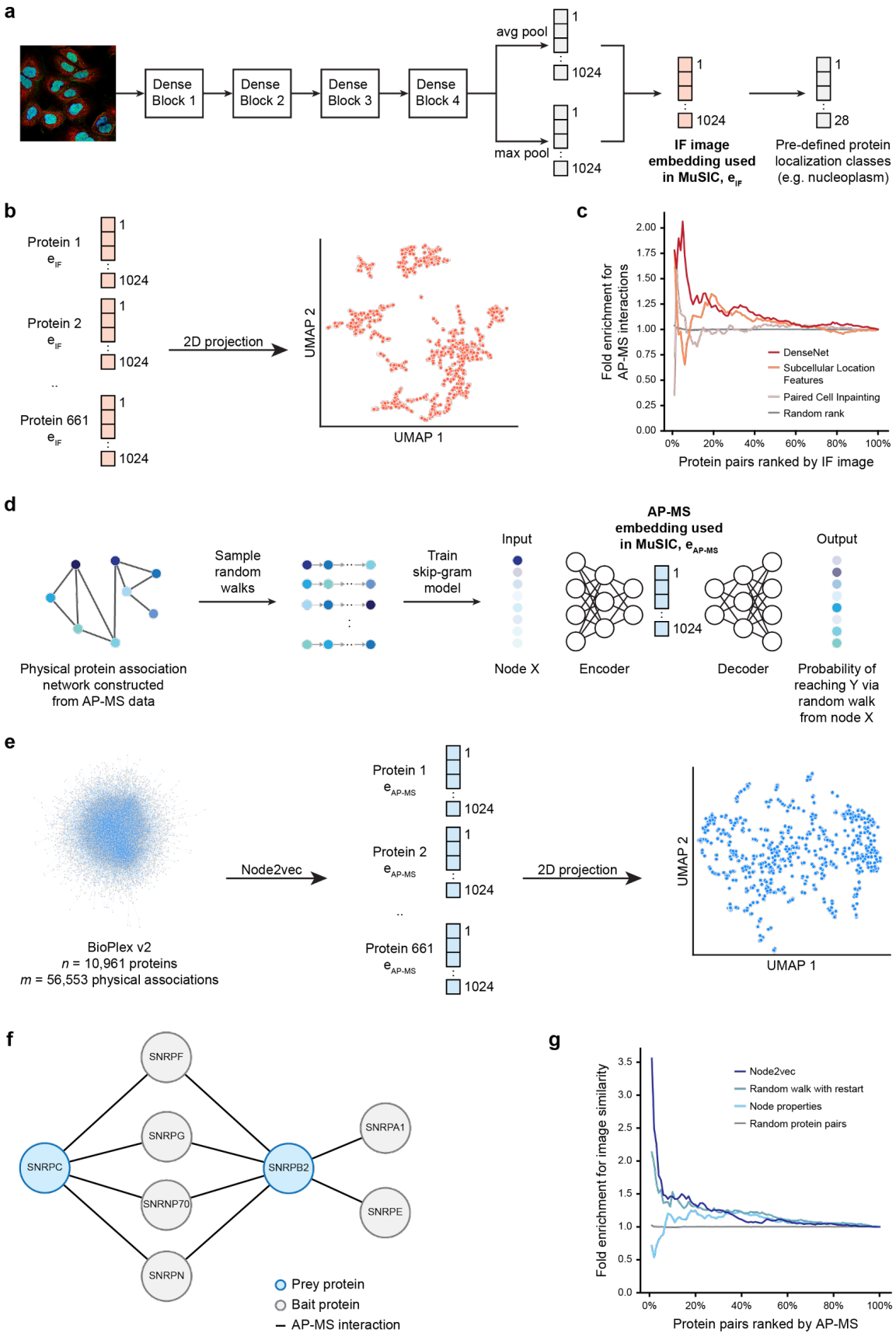


Figure S1.1: Characterization of image data used in this study. a, Histogram showing the distribution in number of antibodies per protein over the 661 proteins included in MuSIC. **b,** Histogram showing the distribution in antibody quality scores over the antibodies used in this study. **c,** IF images for alternative antibodies (columns) targeting the same protein (rows). Colors represent immunostained protein (green), cytoskeleton (red), or nucleus (blue). The images show high reproducibility even when different antibodies for the same target protein are used. **d,** Comparison of localizations for proteins in MuSIC (HEK293 cells, red) versus all proteins assayed by HPA in any cell line (grey). Localizations as defined by the HPA project (Thul et al., 2017).

Figure S1.2: Embedding immunofluorescence images and AP-MS data. a, Embedding immunofluorescence (IF) images using DenseNet. The 1024-dimension feature vector for each IF image was extracted from a DenseNet-121 (Huang et al., 2016) model trained to classify the IF image into one or several of 28 pre-defined protein localization classes from HPA. **b,** 2D visualization (UMAP, n_neighbors = 5) for the 1,451 image embeddings associated with the 661 proteins in MuSIC. **c,** Ability of different image embedding methods (colored curves) to generate image-image similarities (cosine similarity) in agreement with protein-protein interactions in BioPlex 2.0 (which records 281 interactions among the 661 proteins in MuSIC). **d,** Node2vec (Grover & Leskovec, 2016) workflow. The feature vector generated by node2vec captures the pattern of interaction neighborhood for the respective node in input network. **e,** Embedding AP-MS data using node2vec. The input network to node2vec was constructed by treating each protein as a node and assigning edges between protein pairs that were identified as physically interacting in the AP-MS data. The 2D visualization (UMAP, n_neighbors = 5) for AP-MS embeddings associated with 661 proteins in MuSIC is shown at right. **f,** Network showing all proteins (grey) that physically interact with SNRPC and SNRPB2 (blue) in BioPlex 2.0. SNRPC and SNRPB2 do not physically interact, but the cosine similarity of their embedded features is 0.93 due to shared interaction neighborhood. In many cases of two proteins with high node2vec similarity but without direct interaction in AP-MS data, we found that neither protein had yet been tagged as bait for an affinity purification experiment. In these cases, the node2vec embedding suggests gaps in existing AP-MS data. None of the proteins highlighted in blue were tagged as bait proteins in BioPlex 2.0. **g,** Ability of different AP-MS embedding methods to generate protein-protein similarities (cosine similarity) in agreement with protein pairwise similarities computed from HPA images (considering the top 281 protein pairs by image cosine similarity).


a
Dense Block 1
Dense Block 2
Dense Block 3
Dense Block 4
avg pool
max pool
1
1024
IF image embedding used in MuSIC, e_{IF}
28
Pre-defined protein localization classes (e.g. nucleoplasm)
b
Protein 1 e_{IF}
Protein 2 e_{IF}
Protein 661 e_{IF}
2D projection
UMAP 2
UMAP 1
c
Fold enrichment for AP-MS interactions
DenseNet
Subcellular Location Features
Paired Cell Inpainting
Random rank
Protein pairs ranked by IF image
d
Physical protein association network constructed from AP-MS data
Sample random walks
Train skip-gram model
Input
Node X
Encoder
AP-MS embedding used in MuSIC, e_{AP-MS}
Decoder
Output
Probability of reaching Y via random walk from node X
e
BioPlex v2
n = 10,961 proteins
m = 56,553 physical associations
Node2vec
Protein 1 e_{AP-MS}
Protein 2 e_{AP-MS}
Protein 661 e_{AP-MS}
2D projection
UMAP 2
UMAP 1
f
SNRPF
SNRPG
SNRPC
SNRNP70
SNRPN
SNRPB2
SNRPA1
SNRPE
Prey protein
Bait protein
AP-MS interaction
g
Fold enrichment for image similarity
Node2vec
Random walk with restart
Node properties
Random protein pairs
Protein pairs ranked by AP-MS

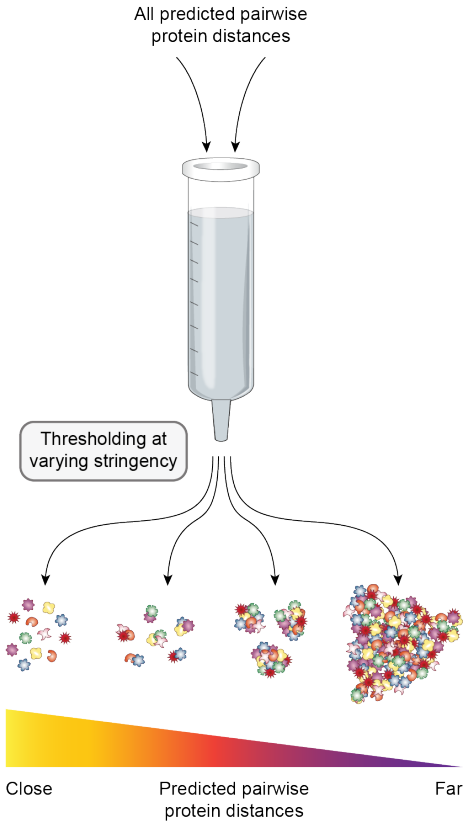
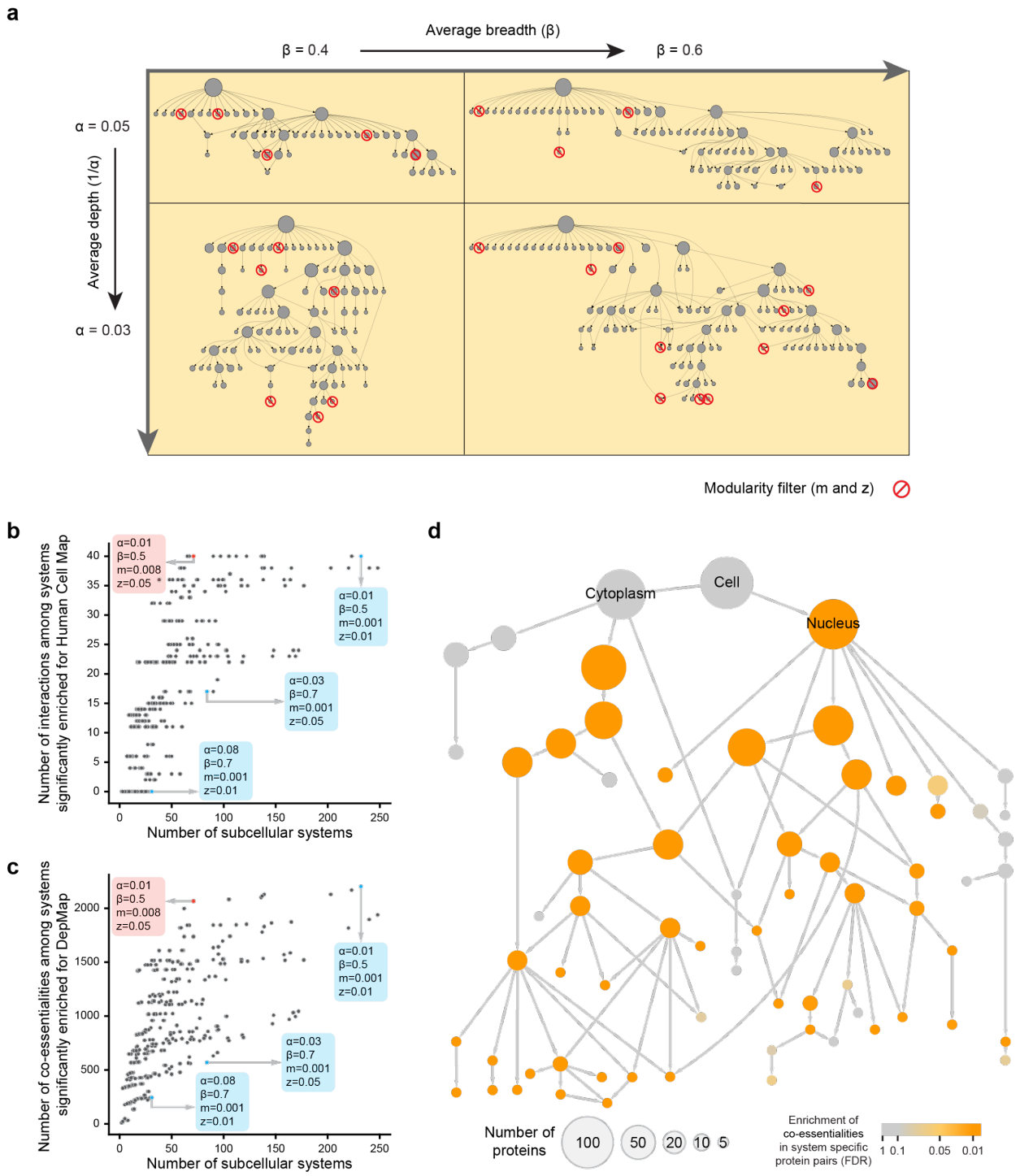


Figure S1.3: Multi-scale community detection. Using multi-scale community detection, protein systems of increasing sizes are discovered as the threshold for protein-protein distance is progressively increased.

Figure S1.4: Selection of parameters for community detection. a, CliXO community detection has four parameters (depth α, y-axis; breadth β, x-axis; minimum modularity m and modularity significance z, red circle backslash) that affect the sensitivity with which communities are identified and thus the size of the hierarchy. **b, c,** Dotplots in which each dot is a community hierarchy generated with a particular set of parameters. The selection for MuSIC is highlighted in red. This selection was among several that were optimal based on enrichment for protein-protein interactions in Human Cell Map (**b**) and co-essentialities from DepMap (**c**). Examples of other parameter sets are shown in blue. **d,** MuSIC map as in **Figure 1.2**, with system color showing enrichment for co-essentialities among protein pairs that are specific to that system. Enrichment of each system was assessed empirically, using 1000 randomized hierarchies, followed by Benjamini-Hochberg multiple test correction to obtain FDR (orange gradient).


a
Average breadth (β)
β = 0.4
β = 0.6
α = 0.05
Average depth (1/α)
α = 0.03
Modularity filter (m and z)
b
Number of interactions among systems significantly enriched for Human Cell Map
Number of subcellular systems
α=0.01 β=0.5 m=0.008 z=0.05
α=0.01 β=0.5 m=0.001 z=0.01
α=0.03 β=0.7 m=0.001 z=0.05
α=0.08 β=0.7 m=0.001 z=0.01
c
Number of co-essentialities among systems significantly enriched for DepMap
Number of subcellular systems
α=0.01 β=0.5 m=0.008 z=0.05
α=0.01 β=0.5 m=0.001 z=0.01
α=0.03 β=0.7 m=0.001 z=0.05
α=0.08 β=0.7 m=0.001 z=0.01
d
Cytoplasm
Cell
Nucleus
Number of proteins
100
50
20
10
5
Enrichment of co-essentialities in system specific protein pairs (FDR)
1
0.1
0.05
0.01

1.5 Author Contributions

Y.Q., E.L. and T.I. designed the study and developed the conceptual ideas. C.F.W. and W.O. generated image embeddings. E.L.H., L.P.V., T.Z., J.W.H. and S.P.G. generated AP-MS data and provided FLAG-HA-tagged clones. Y.Q. and J.M. designed the data integration approach. Y.Q. and F.Z. designed the community detection approach. Y.Q. implemented all computational methods and analyses. All authors contributed to developing ideas for data analyses and experimental designs. Y.Q. and T.I. wrote the manuscript with input from all other authors.

1.6 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in *Nature* 2021. Yue Qin, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, Fan Zheng, Michael Chen, Leah V. Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J. Lee, Wei Ouyang, Sophie N. Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F. Kreisberg, Steven P. Gygi, Jianzhu Ma, J. Wade Harper, Gene W. Yeo, Denis L. J. Lafontaine, Emma Lundberg & Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 600, 536–542 (2021). The dissertation author was the primary investigator and author of this paper.

We gratefully acknowledge helpful discussion and comments from Abraham Palmer, Cherie Ng, members of the Ideker laboratory, members of the Lundberg laboratory, the Human Protein Atlas, Jason Swedlow, and the anonymous referees of this work. We appreciate Michelle Dow for helping us improve the MuSIC GitHub repository and test the MuSIC pipeline. This work was supported by the National Institutes of Health (NIH) under grants P41 GM103504 and R01 HG009979 to T.I., U24 HG006673 to E.L.H., S.P.G, and J.W.H., U41 HG009889 and R01s HL137223 and HG004659 to G.W.Y., R50 CA243885 to J.F.K., by a gift from Google Ventures

to J.W.H. and S.P.G., by the Erling-Persson family foundation, Knut and Alice Wallenberg Foundation (2016.0204) and the Swedish Research Council (2017-05327) to E.L.

1.7 References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10), P10008.

Charenton, C., Wilkinson, M. E., & Nagai, K. (2019). Mechanism of 5' splice site transfer for human spliceosome activation. *Science*, 364(6438), 362–367.

Deckert, J., Hartmuth, K., Boehringer, D., Behzadnia, N., Will, C. L., Kastner, B., Stark, H., Urlaub, H., & Lührmann, R. (2006). Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Molecular and Cellular Biology*, 26(14), 5528–5543.

Fortunato, S. (2009). Community detection in graphs. *arXiv*. http://arxiv.org/abs/0906.0612

Go, C. D., Knight, J. D. R., Rajasekharan, A., Rathod, B., Hesketh, G. G., Abe, K. T., Youn, J.-Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L. Y., Popiel, E., Lambert, J.-P., Coyaud, É., Cheung, S. W. T., Rajendran, D., Wong, C. J., Antonicka, H., Pelletier, L., Palazzo, A. F., … Gingras, A.-C. (2021). A proximity-dependent biotinylation map of a human cell. *Nature*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). *MIT press Cambridge*.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. *International Conference on Knowledge Discovery & Data Mining*, 2016, 855–864.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6), 1515–1526.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *arXiv*. http://arxiv.org/abs/1608.06993

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655), 505–509.

Kramer, M., Dutkowski, J., Yu, M., Bafna, V., & Ideker, T. (2014). Inferring gene ontologies from pairwise similarity data. *Bioinformatics*, 30(12), i34–i42.

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., … Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49(12), 1779–1784.

Ouyang, W., Winsnes, C. F., Hjelmare, M., Cesnik, A. J., Åkesson, L., Xu, H., Sullivan, D. P., Dai, S., Lan, J., Jinmo, P., Galib, S. M., Henkel, C., Hwang, K., Poplavskiy, D., Tunguz, B., Wolfinger, R. D., Gu, Y., Li, C., Xie, J., … Lundberg, E. (2019). Analysis of the Human Protein Atlas Image Classification competition. *Nature Methods*, 16(12), 1254–1261.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.

Sullivan, D. P., Winsnes, C. F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., & Lundberg, E. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature Biotechnology*, 36(9), 820–828.

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1), D330–D338.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, 356(6340).

Williams, S. G., & Hall, K. B. (2011). Human U2B″ protein binding to snRNA stemloops. *Biophysical Chemistry*, 159(1), 82–89.

CHAPTER 2
Different Data Modalities Inform Different Scales of Cell Biology

2.1 Results

MuSIC needs both data types

We found that the majority of MuSIC systems were highly robust to minor disruptions in data (**Figure 2.1a**, jackknife resampling, **Methods**). In contrast, alternative MuSIC maps constructed with only one type of data were found to drop numerous systems. IF-only maps tended to robustly identify large systems such as organelles but falter for smaller subcomponents such as protein complexes, whereas AP-MS maps had the opposite behavior (**Figure 2.1b-d**).

MuSIC informs both data types

Notably, 30% of AP-MS protein interactions fell within a focused system of <100 proteins (**Figure 2.1e**). In each of these cases, such knowledge validates and provides cellular localization context for the protein-protein interaction. We found that such context also increases sensitivity for detection of protein interactions, some of which may have been overlooked in previous proteome-wide AP-MS due to the stringent scoring thresholds necessary to control for false discoveries. Focusing on protein pairs not reported to interact in the previous BioPlex study (Huttlin et al., 2017), we found that pairs in smaller systems nonetheless had significantly stronger AP-MS scores than pairs in larger systems ($P < 0.0001$, **Figure 2.1f**), suggesting an untapped trove of *bona fide* physical interactions.

2.2 Methods

Evaluation of system robustness

To evaluate the robustness of each system in the MuSIC map, we randomly removed 10% of the edges from the protein proximity network, and community detection was performed to construct a hierarchy using the same parameters as MuSIC (see **Pan-resolution community detection** in **Chapter 1**). This randomization procedure was repeated 300 times to obtain a pool of perturbed hierarchies, similar to statistical jackknifing (Efron, 1982). The "percentage recovery" of a MuSIC system s in a perturbed hierarchy H_P was calculated as:

$$\%\, recovery\ (s, H_P) = 100 \times Jaccard(s, s_P) \quad (2.1)$$

where s_P is the perturbed system best enriching for s in H_P. A MuSIC system s was considered to be "recovered" by H_P if the enrichment of s in s_P was significant (FDR ≤ 0.001, hypergeometric statistic) and $\%\ recovery(s, H_P) \geq 40\%$, or if $\%\ recovery(s, H_P) = 100\%$, representing perfect agreement regardless of significance. We computed $r(s)$, the robustness of system s (**Figure 2.1a**), as the fraction of all perturbed hierarchies that recovered s.

Dependence of systems on data types

To assess the dependence of each system on imaging data, we created an alternative protein proximity network using IF features only, with AP-MS features randomized (see **Random forest prediction** in **Chapter 1**). Subsequently, 10% of the edges in this protein network were randomly removed, and 300 jackknifed hierarchies were created from the protein-protein proximity network using IF features only, similar to the procedure described above (**Evaluation of system robustness**). These hierarchies were used to compute a robustness $r(s)$ for each system (**Figure 2.1b**), with computation of r also as described above. To assess the dependence of each system on AP-MS data, a reciprocal procedure was performed by generating an alternative integrated protein-protein proximity network using AP-MS features only, with IF features randomized (**Figure 2.1c**).

2.3 Figures

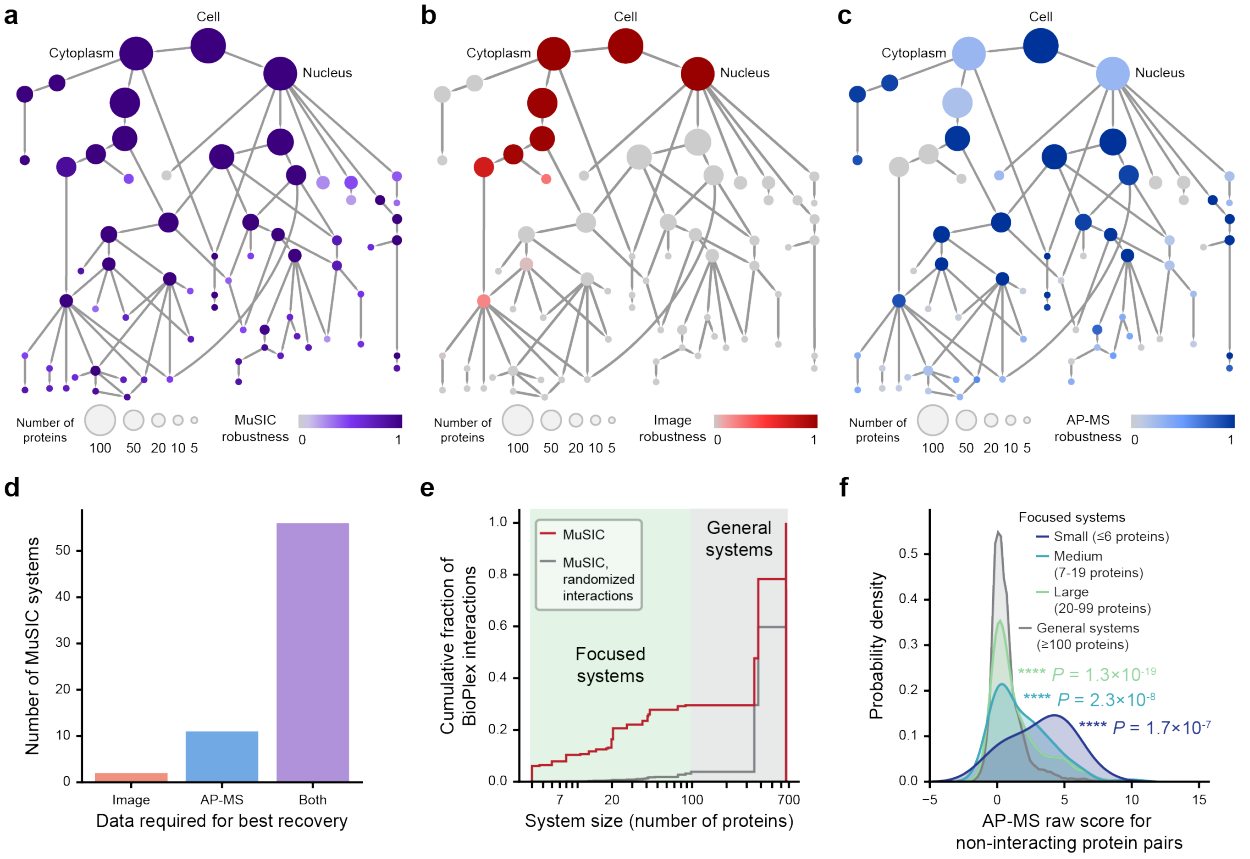


Figure 2.1: Different data, different scales of information. a-c, MuSIC hierarchy colored with system robustness when built using both IF and AP-MS data (full MuSIC, **a**), IF only (**b**), or AP-MS only data (**c**). Each hierarchy is a replica of the MuSIC hierarchy shown in **Figure 1.2**. **d**, Number of systems for which the highest robustness was obtained with IF, AP-MS, or both data types. **e**, Cumulative fraction of BioPlex protein interactions within MuSIC systems (red) versus random protein pairs (gray, 1000 randomizations). **f**, Distribution of AP-MS raw z-scores for protein pairs not labeled as interacting by BioPlex. P-values calculated against general systems ≥100 proteins.

2.4 Author Contributions

Y.Q., E.L. and T.I. designed the study and developed the conceptual ideas. Y.Q. implemented all computational methods and analyses. All authors contributed to developing ideas for data analyses. Y.Q. and T.I. wrote the manuscript with input from all other authors.

2.5 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *Nature* 2021. Yue Qin, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, Fan Zheng, Michael Chen, Leah V. Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J. Lee, Wei Ouyang, Sophie N. Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F. Kreisberg, Steven P. Gygi, Jianzhu Ma, J. Wade Harper, Gene W. Yeo, Denis L. J. Lafontaine, Emma Lundberg & Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 600, 536–542 (2021). The dissertation author was the primary investigator and author of this paper.

We gratefully acknowledge helpful discussion and comments from Abraham Palmer, Cherie Ng, members of the Ideker laboratory, members of the Lundberg laboratory, the Human Protein Atlas, Jason Swedlow, and the anonymous referees of this work. This work was supported by the National Institutes of Health (NIH) under grants P41 GM103504 and R01 HG009979 to T.I., U24 HG006673 to E.L.H., S.P.G, and J.W.H., U41 HG009889 and R01s HL137223 and HG004659 to G.W.Y., R50 CA243885 to J.F.K., by a gift from Google Ventures to J.W.H. and S.P.G., by the Erling-Persson family foundation, Knut and Alice Wallenberg Foundation (2016.0204) and the Swedish Research Council (2017-05327) to E.L.

2.6 References

Efron, B. (1982). The Jackknife, the Bootstrap and other resampling plans.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655), 505–509.

CHAPTER 3
Exploration of Novel Cell Biology Revealed by Cell Map

3.1 Results

Global validation of MuSIC by new AP-MS

Of the 661 proteins common to the IF and AP-MS datasets, 370 had not yet been affinity tagged as the central baits of an AP-MS experiment – rather, they had appeared in the list of preys isolated by another affinity-tagged protein. Therefore, as an immediate means of validating candidate systems in the MuSIC map, we created affinity tags for 134 former preys and performed AP-MS, resulting in the identification of 339 physical interactions (**Table S1.1**). We found that 44 of the 69 MuSIC systems were specifically enriched for the new interactions (64%, FDR < 0.1, **Figure 3.1a**), including 23 novel systems and those at very large and small scales in size. Additionally, 195 new interactions (58%) fell into MuSIC systems of <100 proteins, placing these into specific subcellular contexts (**Figure S3.1**).

Ribosomal systems at multiple scales

Among the novel systems validated by the additional interaction data was an assembly of seven proteins with an estimated diameter of 81 nm (95% prediction interval [43, 151]). We had tentatively named this system "Pre-ribosomal RNA (pre-rRNA) processing assembly" (PRRPA) based on synthesizing established pre-rRNA roles for two of its proteins (Chaudhuri et al., 2007; Yoshikatsu et al., 2015) (NVL, RPL13A) with supportive results from human high-throughput genetic screens (Tafforeau et al., 2013) (KRI1, NOC2L) and orthology to a pre-rRNA factor in budding yeast (Eppens et al., 2002) (REXO4). These proteins formed a MuSIC system due to IF similarity, with predominantly nucleolar localizations, and similarity of AP-MS network

neighborhoods (**Figure 3.1b, c**, **Figure S3.2a, b**). The AP-MS similarity was due to many indirect connections among these proteins (e.g., many network paths of length 2, **Figure S1.2f**), as most had not yet been tagged as BioPlex baits. To fill this gap, our new affinity purifications had directly targeted five PRRPA proteins, resulting in recovery of AP-MS interactions that were highly specific to this system (**Figure 3.1c, Figure S3.2c**). To explore the function of PRRPA proteins in pre-rRNA processing, we used small interfering RNAs (siRNA) to knockdown each of the corresponding proteins, noting that all knockdowns perturbed general ribosomal RNA maturation to some extent (**Figure 3.1d, Figure S3.3**). We then used RNA immunoprecipitation followed by quantitative polymerase chain reaction (RIP-qPCR) to show that these proteins bind the 45S pre-rRNA, consistent with a role in pre-rRNA processing (**Figure 3.1e**).

We also examined the larger scale system containing PRRPA, "Ribosome biogenesis community", with an estimated diameter of 347 nm (95% prediction interval [186, 646]). This system contained additional proteins not previously associated with ribosome biogenesis in humans (**Figure 3.1f**), seven of which we knocked down with targeted Dicer-substrate siRNAs (DsiRNA). All seven had varying effects on pre-rRNA processing upon knockdown, which stratified according to the specific pre-rRNA affected (**Figure 3.1g**, **Figure S3.4**). Finally, we had targeted three of these proteins as baits in our new AP-MS experiments (LIN28B, PRR3, ZNF689), each of which identified interaction partners that strongly enriched for proteins within the Ribosome biogenesis community (**Figure 3.1h**).

Another notable finding within ribosomal systems was abundant crosstalk between canonical subunits of the cytoplasmic and mitochondrial ribosomes ("Mito-cyto ribosomal cluster"; 20 nm, 95% prediction interval [11, 38]; **Figure S3.5a-d**). In the new affinity pull-downs, we tagged four of these proteins as baits (two cytoribosomal, two mitoribosomal), identifying five

within-system physical interactions, four interconnecting cytoplasmic and mitochondrial factors (**Figure S3.5e**). Such crosstalk has not been previously reported but may play a role in mitoribosome biogenesis, a poorly understood process (De Silva et al., 2015).

Exploration of chromatin and splicing

SRRM1 is an established splicing factor (Blencowe et al., 2000) which, in addition to its canonical placement in "RNA splicing complex 3" (71 nm, 95% prediction interval [38, 133]), participated in several additional MuSIC systems that were unexpected. One of these, "Chromatin regulation complex" (211 nm, 95% prediction interval [113, 393]), consisted of three histone acetyltransferases (HATs; DMAP1, JAZF1, and MORF4L1) (Piunti et al., 2019; UniProt Consortium, 2019) together with SATB1, a known DNA-binding protein that remodels chromatin through HAT recruitment (Pavan Kumar et al., 2006) (**Figure 3.2a**, **Figure S3.5f-h**). These functions suggested that SRRM1 and FAM120C, the remaining proteins in this system, might also have roles in regulation of chromatin. In support of this suggestion, we found that SRRM1 and FAM120C strongly associate with chromatin by an *in situ* fractionation assay (**Figure 3.2a, b**).

Returning to "RNA splicing complex 3," we noted that SRRM1 was associated with two other established factors in the major spliceosomal pathway, SNRNP70 (Pomeranz Krummel et al., 2009) and U2AF2 (Fleckner et al., 1997) (**Figure 3.2c, d**). A fourth member of this complex, RPS3A, was distinct from the first three in that it was a ribosomal protein (UniProt Consortium, 2019) not previously associated with major RNA splicing. However, analysis of published transcriptomic profiles (Van Nostrand et al., 2020) indicated that shRNA knockdown of RPS3A had very similar transcriptional effects as knockdown of SNRNP70 or U2AF2 (**Figure 3.2e**, **Figure S3.5i**). To test for a potential role in splicing, we subjected RPS3A to an enhanced UV crosslinking and immunoprecipitation assay (Van Nostrand et al., 2016, 2017) (eCLIP, **Figure**

3.2f), which identifies the RNA transcripts bound by a protein along with specific protein-binding sites, including intronic and exonic sequences and 3'/5' untranslated regions (UTR). Among the 866 reproducible and significant RNA binding regions identified (eCLIP peaks, **Table S3.1**), RPS3A predominantly bound to intronic regions (601 peaks, 69%) with a pattern very similar to that of canonical splicing regulators and distinct from that of other ribosomal proteins (**Figure 3.2g**). Moreover, when clustering the RNA binding profile for RPS3A together with all 223 eCLIP profiles available in the public domain (Van Nostrand et al., 2020), RPS3A robustly clustered with canonical splicing regulators (92% recovery in jackknife resampling, **Figure 3.2h**), in support of an alternative role for this protein.

3.2 Methods

Global system validation using new AP-MS data

We constructed stable HEK293T cell lines for 134 bait proteins (**Table S1.1**) with C-terminal FLAG-HA-tags based on the human ORFeome v8.1 (http://horfdb.dfci.harvard.edu/) (Yang et al., 2011) as previously described (Huttlin et al., 2015, 2017, 2021). Cell pellets were lysed using 50 mM Tris-HCl pH 7.5, 300 mM NaCl, 0.5% (v/v) NP40 buffer, and cell debris were removed with centrifugation and filtration. Mouse monoclonal anti-HA agarose resins (Sigma-Aldrich, clone HA-7), immobilized and pre-washed, were incubated with cell lysates at 4 ℃ for 4 hours. After removing supernatant, precipitates were washed four times with lysis buffer and two times with PBS (pH 7.2). Elution was performed in two steps by adding 250 μg/mL HA peptide in PBS at 37°C followed by TCA precipitation. Eluted samples were analyzed by LC-MS using Q-Exactive mass spectrometers (Thermo Fisher). Each bait protein was analyzed with a single biological replicate for the affinity purification step and technical duplicates for the LC-MS step, yielding four replicates in total. MS/MS spectra were analyzed using the Sequest algorithm (Eng

et al., 1994) to match peptide sequences from the Uniprot database (UniProt Consortium, 2019) supplemented by signatures for green fluorescent protein (negative control), the FLAG-HA-tag, and common contaminants. Identified peptides and proteins were further filtered using the target-decoy method (Elias & Gygi, 2007) to control FDR. High confidence protein interactions were identified using the ComPASS algorithm (Behrends et al., 2010; Sowa et al., 2009) on merged technical duplicates, followed by ComPASS-Plus analysis (Huttlin et al., 2015).

To examine per-system enrichment for the new AP-MS data in MuSIC map, the assignment of proteins to systems was shuffled while keeping the overall hierarchy structure and number of proteins per system the same, resulting in 1,000 random hierarchies. For each system, we calculated the empirical p-value (North et al., 2002) for the number of new AP-MS interactions among the "system-specific" protein pairs, defined as protein pairs in that system but not in children of that system. P-values were corrected using Benjamini-Hochberg multiple test correction to obtain an FDR for per-system enrichment (**Figure 3.1a**).

Analysis of mature rRNA by TapeStation

HEK293T cells were plated the first day and transfected (Lipofectamine RNAiMAX Transfection Reagent, Thermo Fisher) with small interfering RNA (siRNA, purchased from SIRNA - MISSION® siRNA, see **Table S3.2** for sequences used) against proteins of interest (**Figure 3.1d**) the next day, followed by 72 hour incubation. Protein knockdown was assessed using either Simple Western assay (WES, ProteinSimple) or SDS-PAGE (see **Table S3.3** for list of antibodies, **Figure S3.3a-f**). Protein knockdowns were observed for all proteins except RPL7L1, which was excluded from further analysis here. Total RNA was extracted using TRIzol (Invitrogen) followed by Direct-zol™ RNA Microprep. RNA intensity profiles (**Figure S3.3g**) were analyzed using a TapeStation system (Agilent). Abundances of 28S and 18S rRNA were

determined by fitting a Gaussian function to the RNA intensity profiles and computing the area under this curve. The raw ratios of 28S/18S rRNA were normalized to that of samples treated with scrambled siRNA to obtain relative ratios (**Figure 3.1d**).

Assessment of pre-rRNA binding using RIP-qPCR

Biological duplicates of stable cell lines expressing C-terminal FLAG-HA-tagged proteins of interest (**Figure 3.1e**) were created using ORFeome v.8.1 clones (Yang et al., 2011) in HEK293T cells. Lentiviral transfected cells were maintained in puromycin (2.5 µg/mL) selection. At least 20 million cells were UV crosslinked (400 mJoules/cm^2, 254 nm) followed by cell lysis and sonication, as previously described (Van Nostrand et al., 2016), then DNase treated. 2.5% of samples were saved as input RNA and anti-HA antibody-bound beads were incubated with the remainder of cell lysates at 4℃ overnight. IP samples were washed 3 times with eCLIP high salt wash buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate) and 5 times with eCLIP wash buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate). RNA was extracted from input and IP samples using Trizol™ LS reagent (Invitrogen) followed by Direct-zol RNA Miniprep purification (Zymo Research). RNA (0.8% lysate equivalent for input and 250 ng for IP samples) was reverse transcribed with oligodT and random hexamer priming using the SuperScript III First-Strand Synthesis System (Thermo Fisher). Equal volumes of cDNA were quantified via qPCR by incubation for 10 min at 95℃ followed by 40 cycles of [15 sec at 95℃; 15 sec at 55℃; 1 min at 60℃]. Abundance for 45S pre-rRNA was assessed during qPCR with primer set 5'-CCTGCTGTTCTCTCGCGCGTCCGAG-3' and 5'-AACGCCTGACACGCACGGCACGGAG-3' (forward and reverse) (Grandori et al., 2005). The average *Ct* value of qPCR technical triplicates was used for follow-up analysis. Each target protein was first normalized based on its respective

input to obtain $\Delta Ct = Ct_{IP} - (Ct_{Input} - log_2 40)$, then normalized by the average ΔCt of DMAP1 to obtain $\Delta\Delta Ct_{Target} = \Delta Ct_{Target} - \overline{\Delta Ct_{DMAP1}}$, which was used as the $\log_2$ fold enrichment of 45S pre-rRNA bound by the target protein (**Figure 3.1e**).

Analysis of pre-rRNA processing by northern blotting

Cell lines used for pre-rRNA analysis (HeLa ref. ATCC CCL-2; HEK293 ref. ATCC CRL-1573) were purchased from ATCC and confirmed by short tandem repeat (STR) profiling. Cells were grown in DMEM at 37°C under 5% CO_2. Dicer-Substrate Short Interfering RNAs (DsiRNAs) targeting genes of interest (**Figure 3.1g**, see **Table S3.2** for sequences used) were purchased from Integrated DNA Technologies. Cells were transfected with lipofectamine RNAiMAX (Thermo Fisher) with 20 nM silencers and incubated for 3 days. Total RNA extraction and northern analysis were performed as described previously (Tafforeau et al., 2013) (**Figure S3.4**, see **Table S3.4** for probes used). For mature rRNA detection, the gel was stained with ethidium bromide. Signal was quantified with a Fuji FLA-200 phopshorimager and normalized to signal from cells treated with scrambled DsiRNA (**Figure 3.1g**).

Assessment of chromatin-protein association by *in situ* fractionation

In situ fractionation was performed to verify protein-chromatin interactions (**Figure 3.2b**) according to a previously described procedure (Sawasdichai et al., 2010). Briefly, HEK293 cells were washed twice with 1x phosphate buffered saline (PBS, 8 mM Na_2HPO_4, 2 mM NaH_2PO_4, 150 mM NaCl, pH 7.2) and incubated for 5 min in Cytoskeleton Buffer (10 mM PIPES pH 6.8, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 1 mM EGTA, 0.5% Triton X-100) and then for 5 min in Cytoskeleton Stripping Buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% vol/vol Tween 20, 0.5% vol/vol sodium deoxycholate). Cells were washed with PBS three times

and immunostained immediately. The detergent treatment led to a permeabilization of all cellular membranes, thus removing all proteins except those bound to chromatin (Sawasdichai et al., 2010).

RNA-seq data analysis

Fastq files containing raw reads for each shRNA knockdown (**Figure 3.2e**) were downloaded from ENCODE (https://www.encodeproject.org/). Reads were aligned to hg19 using STAR version 2.7.1a (--outFilterType BySJout --outFilterMultimapNmax 10 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 4 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 100000). The expression level for each transcript was quantified using featureCounts v1.6.3 (-T 50 -B -p -s 2). Differential expression was analyzed using DESeq2 v1.28.1. Gene set enrichment analysis (**Figure S3.5i**) was performed using the MSigDB webserver (Subramanian et al., 2005) with Gene Ontology Biological Process.

Assessment of bound RNA by eCLIP analysis

HEK293T cells were transfected using lentivirus harboring a C-terminal FLAG-HA-tagged RPS3A clone from ORFeome v.8.1 (Yang et al., 2011). Two biological replicates of stable cell lines were created and maintained in 2.5 μg/mL puromycin. The eCLIP experiments were performed as previously described (Van Nostrand et al., 2016, 2017). Briefly, >20 million cells were collected for UV crosslink (400 mJoules/cm^2, 254 nm), followed by cell lysis, sonication, and RNase I treatment. Anti-HA antibody (BioLegend #901501) was incubated with cell lysates at 4℃ overnight. 2% of samples were saved as paired input before immunoprecipitation (IP) steps. IP samples were washed, followed by RNA dephosphorylation (FastAP, Thermo Fisher; T4 PNK, NEB) and 3' RNA adaptor ligation (T4 RNA ligase, NEB). IP and input samples were run on a PAGE Bis-Tris protein gel and subsequently transferred to a nitrocellulose membrane. Region

starting from the protein size up to 75 kDa above was excised from the membrane for proteinase K (NEB) treatment and column purification (Zymo). RNA obtained from input samples was also dephosphorylated and ligated to 3' RNA adaptors as performed previously to IP samples. Final RNA samples were reverse transcribed, ligated to a 3' DNA adaptor (T4 RNA ligase, NEB), and PCR amplified to obtain the final library for next generation sequencing. Following sequencing, raw reads were aligned to GRCh38 and analyzed following a previously published pipeline (Van Nostrand et al., 2016, 2017, 2020). Consistent with the ENCODE standard (Van Nostrand et al., 2020), reads aligning to artifact-enriched or repetitive genomic regions were removed, producing 866 reproducible and significant peaks of aligned reads at IDR cutoff of 0.01, $P \leq 0.001$, and fold enrichment ≥ 8. Genic regions of eCLIP peaks were annotated based on overlap with GENCODE v26 transcripts following the priority order consistent with the previous study (Van Nostrand et al., 2020) (**Figure 3.2g, h**; **Table S3.1**).

Hierarchical clustering of eCLIP profiles

Genic region profiles for 223 eCLIP experiments were obtained from Van Nostrand et al (Van Nostrand et al., 2020). The "intron" category was defined by eCLIP peaks annotated with any of the following genic regions: 5' splice site, non-coding 5' splice site, 3' splice site, non-coding 3' splice site, proximal intron, non-coding proximal intron, distal intron, non-coding distal intron. The "miRNA" category included all eCLIP peaks annotated to either miRNA or proximal miRNA regions. Hierarchical clustering of genic region profiles was performed using the fastcluster (Müllner & Others, 2013) Python package (metric='euclidean', method='ward') (**Figure 3.2h**).

To assess the robustness of the cluster containing RPS3A ($C_{observed,RPS3A}$), we used a statistical jackknife approach in which we randomly dropped 10% of eCLIP profiles and calculated the recovery rate *r* using the following equation:

$$r = \frac{|C_{observed,RPS3A} \cap C_{sampled,RPS3A}|}{|C_{observed,RPS3A}|} \quad (3.1)$$

where $C_{sampled,RPS3A}$ represents eCLIP profiles that clustered with RPS3A after jackknifing. The above procedure was repeated 1000 times, after which % recovery was estimated as $100 \times \bar{r}$, with $\bar{r}$ denoting the average of all *r*. To assess statistical significance, we randomly selected 77 eCLIP profiles (same number that clustered with RPS3A in $C_{observed,RPS3A}$), drawing from the profiles not clustering with RPS3A. We then calculated *r* for these eCLIP profiles over 1000 random samplings. The p-value was assessed with a two-sided Wilcoxon signed-rank test.

3.3 Figures

Figure 3.1: Exploration of MuSIC using additional pull-downs and functional assays. a, MuSIC map as in **Figure 1.2**, with system color showing enrichment for new AP-MS interactions (FDR, blue gradient). **b,** IF images for proteins in PRRPA system. Green, immunostained protein; red, cytoskeleton. **c**, PRRPA AP-MS interactions, display as in **Figure 1.3d**. **d,** Mature 28S/18S rRNA ratio under siRNAs targeting each PRRPA protein (green) versus scrambled siRNA (gray), $n = 3$ biological replicates. FDR from two-sided t-test with Benjamini-Hochberg (BH) correction. Error bars show standard deviation. **e,** Enrichment of 45S pre-rRNA bound by FLAG-HA-tagged proteins (x axis), measured using RIP-qPCR normalized to DMAP1 ($n = 2$ stable cell lines). FDR from one-sided t-test with BH correction. Error bars as in (**d**). N.S. for non-significant. **f,** Categorization of proteins in "Ribosome biogenesis community" by whether they have been previously identified in human ribosome biogenesis. Excludes PRRPA proteins described in previous panels. **g,** Heatmap summarizing northern blot analysis of intermediate RNA products during pre-rRNA processing (rows), under DsiRNA targeting the respective gene (columns). Heatmap color shows the percentage of each pre-rRNA species with respect to the scramble control. As controls, cells were depleted for UTP18, a known ribosome biogenesis factor, or a non-targeting scramble silencer. Effects from independent silencers against a particular target were highly consistent. **h,** For protein baits in new AP-MS experiments (x axis), the fraction of interacting preys that fall within the Ribosome biogenesis community (blue bars) or elsewhere (gray bars) is shown. Only new AP-MS interactions were considered for this analysis. RNPS1 does not belong to Ribosome biogenesis community and serves as a negative control.

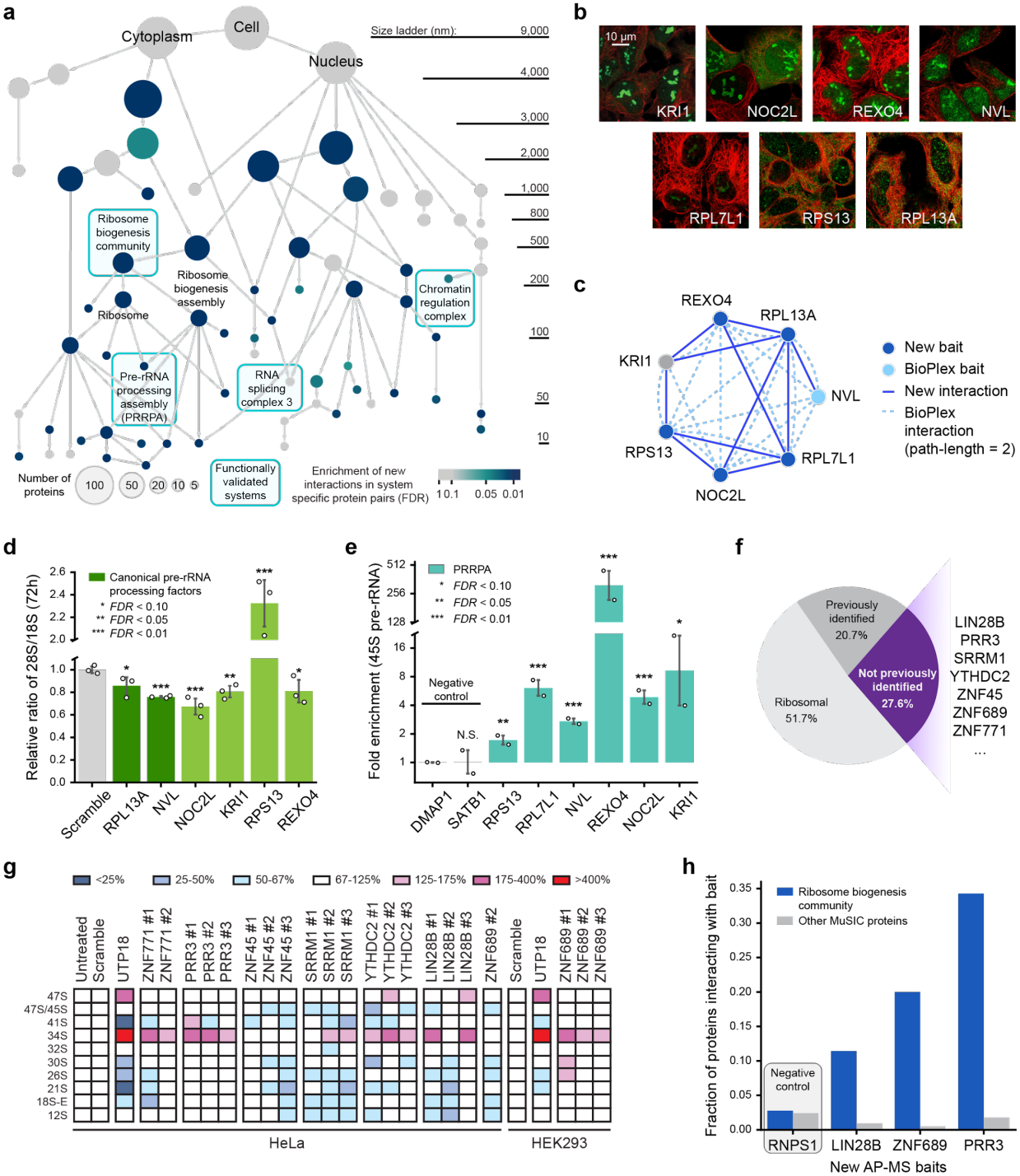
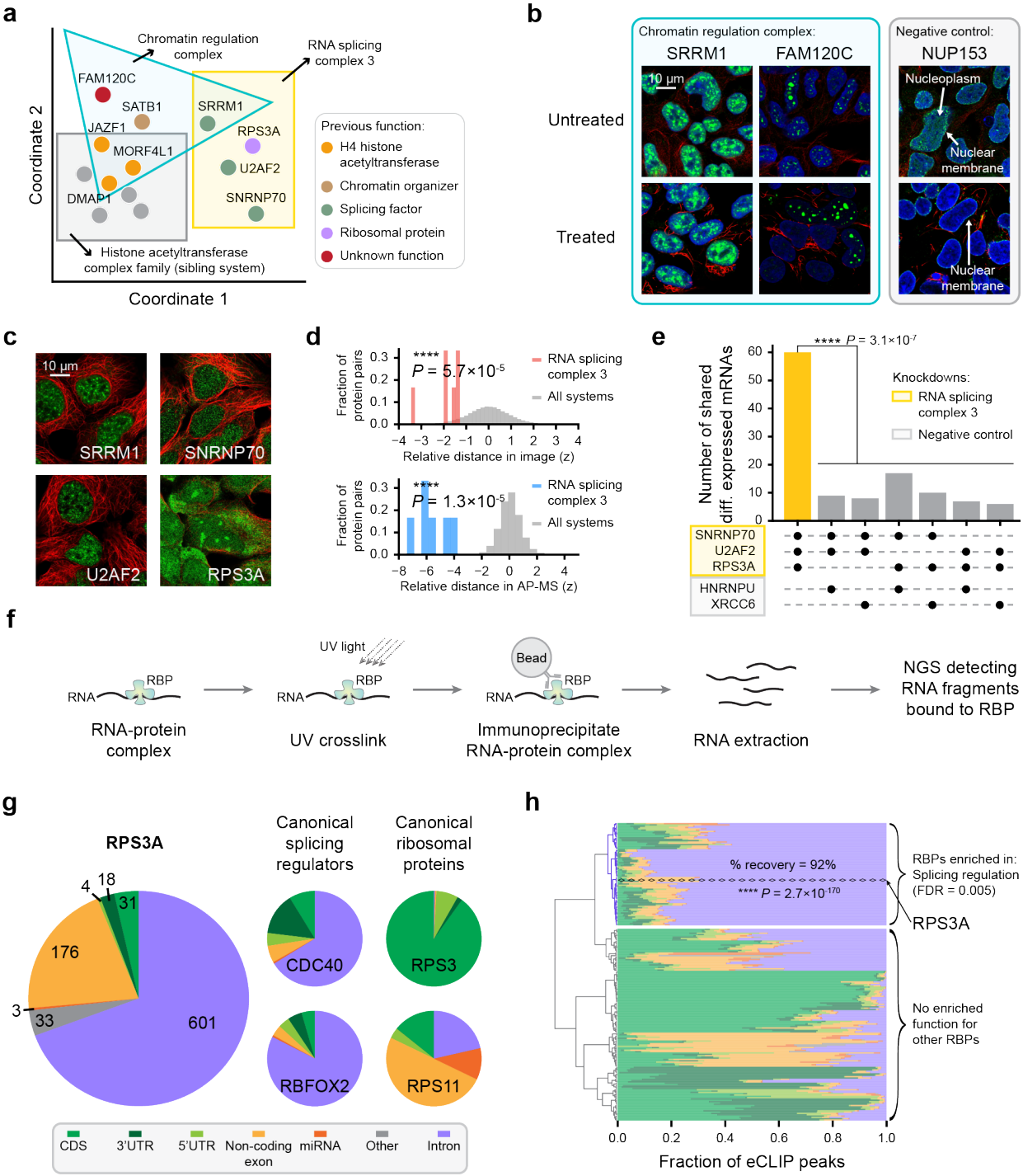

a
Cytoplasm
Cell
Nucleus
Size ladder (nm):
9,000
4,000
3,000
2,000
1,000
800
500
200
100
50
10
Ribosome biogenesis community
Ribosome biogenesis assembly
Ribosome
Chromatin regulation complex
Pre-rRNA processing assembly (PRRPA)
RNA splicing complex 3
Number of proteins
100
50
20
10
5
Functionally validated systems
Enrichment of new interactions in system specific protein pairs (FDR)
1
0.1
0.05
0.01
b
10 µm
KRI1
NOC2L
REXO4
NVL
RPL7L1
RPS13
RPL13A
c
REXO4
RPL13A
KRI1
NVL
RPS13
RPL7L1
NOC2L
New bait
BioPlex bait
New interaction
BioPlex interaction (path-length = 2)
d
Relative ratio of 28S/18S (72h)
Canonical pre-rRNA processing factors
* FDR < 0.10
** FDR < 0.05
*** FDR < 0.01
Scramble
RPL13A
NVL
NOC2L
KRI1
RPS13
REXO4
e
Fold enrichment (45S pre-rRNA)
PRRPA
* FDR < 0.10
** FDR < 0.05
*** FDR < 0.01
Negative control
N.S.
DMAP1
SATB1
RPS13
RPL7L1
NVL
REXO4
NOC2L
KRI1
f
Previously identified
20.7%
Not previously identified
27.6%
Ribosomal
51.7%
LIN28B
PRR3
SRRM1
YTHDC2
ZNF45
ZNF689
ZNF771
...
g
<25%
25-50%
50-67%
67-125%
125-175%
175-400%
>400%
Untreated
Scramble
UTP18
ZNF771 #1
ZNF771 #2
PRR3 #1
PRR3 #2
PRR3 #3
ZNF45 #1
ZNF45 #2
ZNF45 #3
SRRM1 #1
SRRM1 #2
SRRM1 #3
YTHDC2 #1
YTHDC2 #2
YTHDC2 #3
LIN28B #1
LIN28B #2
LIN28B #3
ZNF689 #2
Scramble
UTP18
ZNF689 #1
ZNF689 #2
ZNF689 #3
47S
47S/45S
41S
34S
32S
30S
26S
21S
18S-E
12S
HeLa
HEK293
h
Fraction of proteins interacting with bait
Ribosome biogenesis community
Other MuSIC proteins
Negative control
RNPS1
LIN28B
ZNF689
PRR3
New AP-MS baits

Figure 3.2: MuSIC reveals proteins in chromatin and splicing. a, 2D projection (spring embedding) of distances among proteins in chromatin regulation, RNA splicing, and histone acetyltransferase complexes. Colored frames organize proteins assigned to each complex, with protein color indicating previously assigned functions. **b,** Immunofluorescent proteins (columns) imaged in HEK293 cells, untreated (top) or treated (bottom) with *in situ* fractionation to remove soluble cytoplasmic and loosely held nuclear proteins. Chromatin-binding proteins remain after treatment. Green, immunostained protein; red, cytoskeleton; blue, nucleus. **c,** IF images showing similar nucleoplasm signals among a community of four proteins identified by MuSIC, "RNA splicing complex 3." **d,** Corresponding distributions of protein-protein distances (z-scores) for IF (red) or AP-MS (blue) data, calibrated to all pairwise distances (gray distribution). **e,** Comparison of 500 top differentially expressed mRNAs (absolute fold change) resulting from shRNA knockdown of each of five genes. Bar chart shows number of differential mRNAs shared by different gene groups indicated by black dots beneath each bar. One-sided one sample t-test. **f,** eCLIP workflow. **g,** Pie charts categorizing significant eCLIP peaks by type of genomic region (colored slices). Results for RPS3A (left large pie) compared to proteins with well-characterized functions (small pies, right). **h,** Hierarchical clustering of RPS3A eCLIP profile (dashed line) with all 223 ENCODE eCLIP profiles (Van Nostrand et al., 2020) from 150 proteins. Proteins robustly clustering with RPS3A (92% recovery from 1000 jackknife resamplings) are significantly enriched for splicing regulators. Hypergeometric test with Benjamini-Hochberg correction against ENCODE RBP function library (Van Nostrand et al., 2020). Color consistent with (**g**). CDS, coding sequence. NGS, next-generation sequencing. RBP, RNA binding protein. UTR, untranslated region.


a
Chromatin regulation complex
RNA splicing complex 3
FAM120C
SATB1
JAZF1
MORF4L1
DMAP1
SRRM1
RPS3A
U2AF2
SNRNP70
Histone acetyltransferase complex family (sibling system)
Coordinate 1
Coordinate 2
Previous function:
H4 histone acetyltransferase
Chromatin organizer
Splicing factor
Ribosomal protein
Unknown function
b
Chromatin regulation complex:
SRRM1
FAM120C
Negative control:
NUP153
Untreated
Treated
10 μm
Nucleoplasm
Nuclear membrane
c
10 μm
SRRM1
SNRNP70
U2AF2
RPS3A
d
Fraction of protein pairs
**** P = 5.7×10-5
RNA splicing complex 3
All systems
Relative distance in image (z)
**** P = 1.3×10-5
Relative distance in AP-MS (z)
e
**** P = 3.1×10-7
Number of shared diff. expressed mRNAs
Knockdowns:
RNA splicing complex 3
Negative control
SNRNP70
U2AF2
RPS3A
HNRNPU
XRCC6
f
RBP
RNA
UV light
Bead
RNA-protein complex
UV crosslink
Immunoprecipitate RNA-protein complex
RNA extraction
NGS detecting RNA fragments bound to RBP
g
RPS3A
4
18
1
31
176
3
33
601
Canonical splicing regulators
Canonical ribosomal proteins
CDC40
RPS3
RBFOX2
RPS11
CDS
3'UTR
5'UTR
Non-coding exon
miRNA
Other
Intron
h
% recovery = 92%
**** P = 2.7×10-170
RBPs enriched in: Splicing regulation (FDR = 0.005)
RPS3A
No enriched function for other RBPs
Fraction of eCLIP peaks

3.4 Supplementary Figures

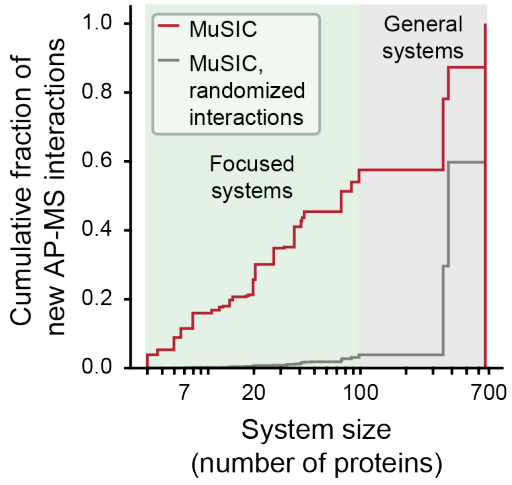


Figure S3.1: Cumulative fraction of new AP-MS interactions in MuSIC. Cumulative fraction of newly identified protein interactions within MuSIC systems (red) versus random protein pairs (gray, 1000 randomizations).

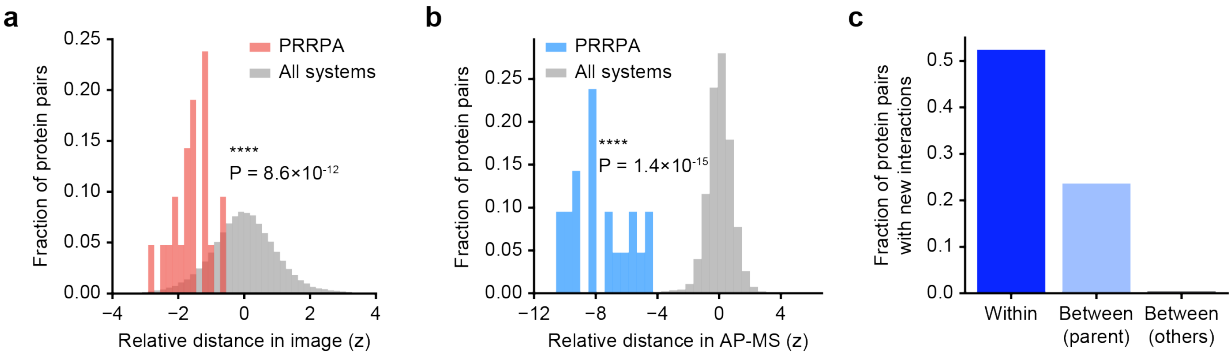


Figure S3.2: Supporting analyses for PRRPA. a, b, Distributions of protein-protein distance z-scores among the seven proteins in the PRRPA system for IF (**a**, red) or AP-MS (**b**, blue) modalities, calibrated to all such distances, respectively (gray). **c,** The specific recovery of new AP-MS interactions within PRRPA is shown (dark blue bar), in comparison to interactions between proteins in PRRPA and other proteins organized under the same parent systems ("Ribosome" and "Ribosome biogenesis assembly", light blue bar), or between proteins in PRRPA and those organized elsewhere in MuSIC (gray bar).

Figure S3.3: Knockdown of PRRPA proteins for functional assay. a-f, Western blot analysis (**a-b**, Simple Western assay; **c-f**, SDS-PAGE) of target protein abundance after treating HEK293T cells with respective siRNA for 72 hours (**Tables S3.2, S3.3**). The siRNAs highlighted in red were selected to assess the perturbation of mature rRNA ratio (28S/18S rRNA) when knocking down target protein. **g,** TapeStation analysis of total RNA extracted from HEK293T cells treated with selected siRNA (panels **a-f**) for 72 hours. The quantities for 28S rRNA and 18S rRNA were determined with Gaussian curve fitting. Raw 28S/18S rRNA ratio is labeled in the respective total RNA profile (related to **Figure 3.1d**).

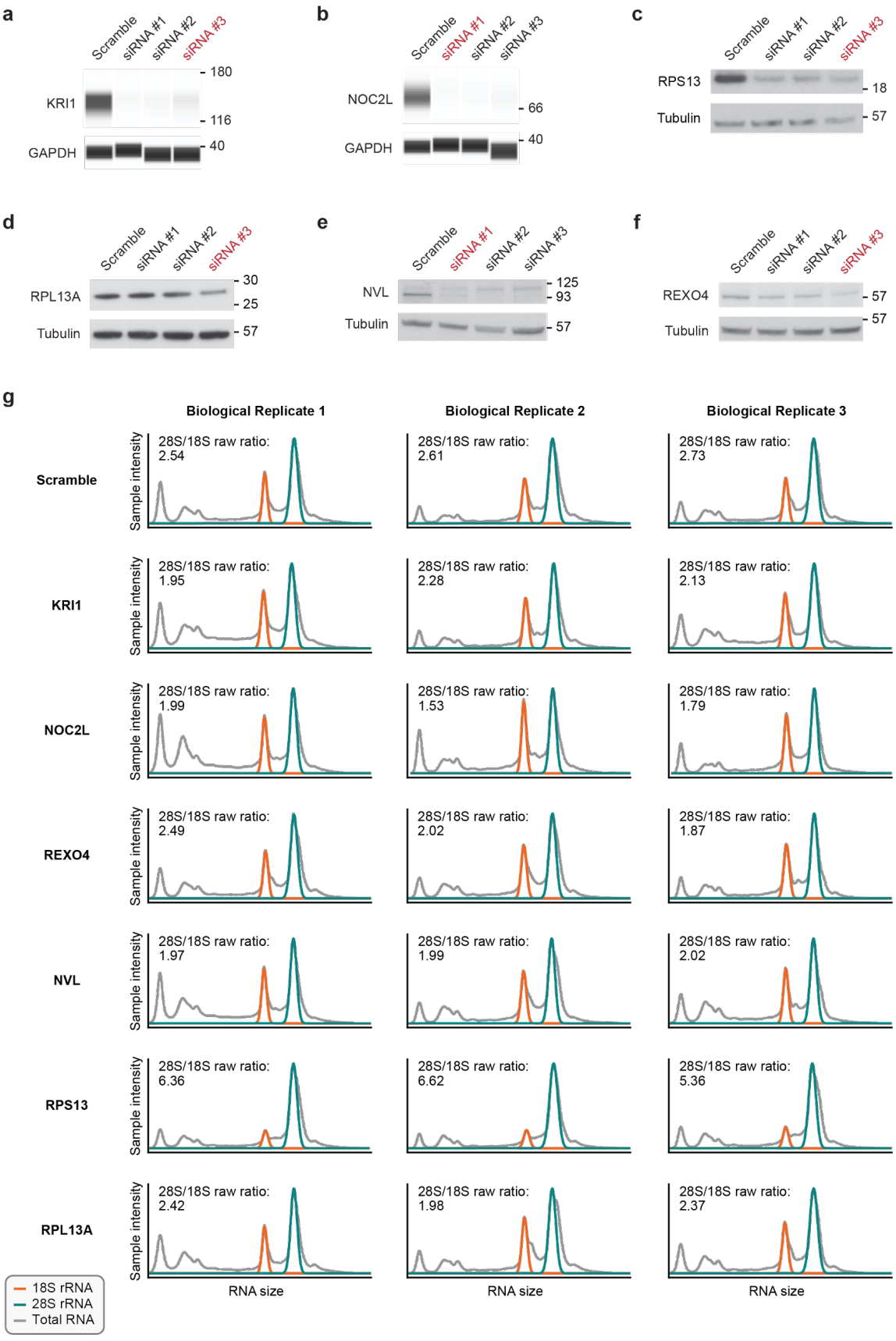

a
Scramble
siRNA #1
siRNA #2
siRNA #3
KRI1
GAPDH
180
116
40
b
NOC2L
66
c
RPS13
18
Tubulin
57
d
RPL13A
30
25
e
NVL
125
93
f
REXO4
g
Biological Replicate 1
Biological Replicate 2
Biological Replicate 3
Scramble
KRI1
NOC2L
REXO4
NVL
RPS13
RPL13A
Sample intensity
28S/18S raw ratio:
2.54
2.61
2.73
1.95
2.28
2.13
1.99
1.53
1.79
2.49
2.02
1.87
1.97
1.99
2.02
6.36
6.62
5.36
2.42
1.98
2.37
RNA size
18S rRNA
28S rRNA
Total RNA

Figure S3.4: Functional assays for proteins in "Ribosome biogenesis community". a, Structure of human pre-rRNA and probes used for northern blot. In eukaryotes, 3 out of 4 mature rRNAs (the 18S, 5.8S, and 28S rRNAs) are produced from a single long polycistronic precursor (47S) synthesized by RNA polymerase I. The mature rRNAs are interspersed with the 5' and 3' external transcribed spacers (ETS) and internal transcribed spacer (ITS) 1 and 2. The probes used in the northern blotting (5'-ETS, ITS1, and ITS2) are indicated and color-coded. **b,** Total RNA extracted from the indicated cell line, which was transfected with a DsiRNA specific to the target protein for 72 hours, analyzed by northern blotting with probes specific to the 5'-ETS, ITS1, and ITS2 sequences (**Table S3.4**). As controls, cells were either untreated, transfected with a scrambled silencer, or transfected with a silencer targeting UTP18 (positive control involved in small ribosomal subunit biogenesis). Heatmap color shows the percentage of each pre-rRNA species with respect to the scramble control.

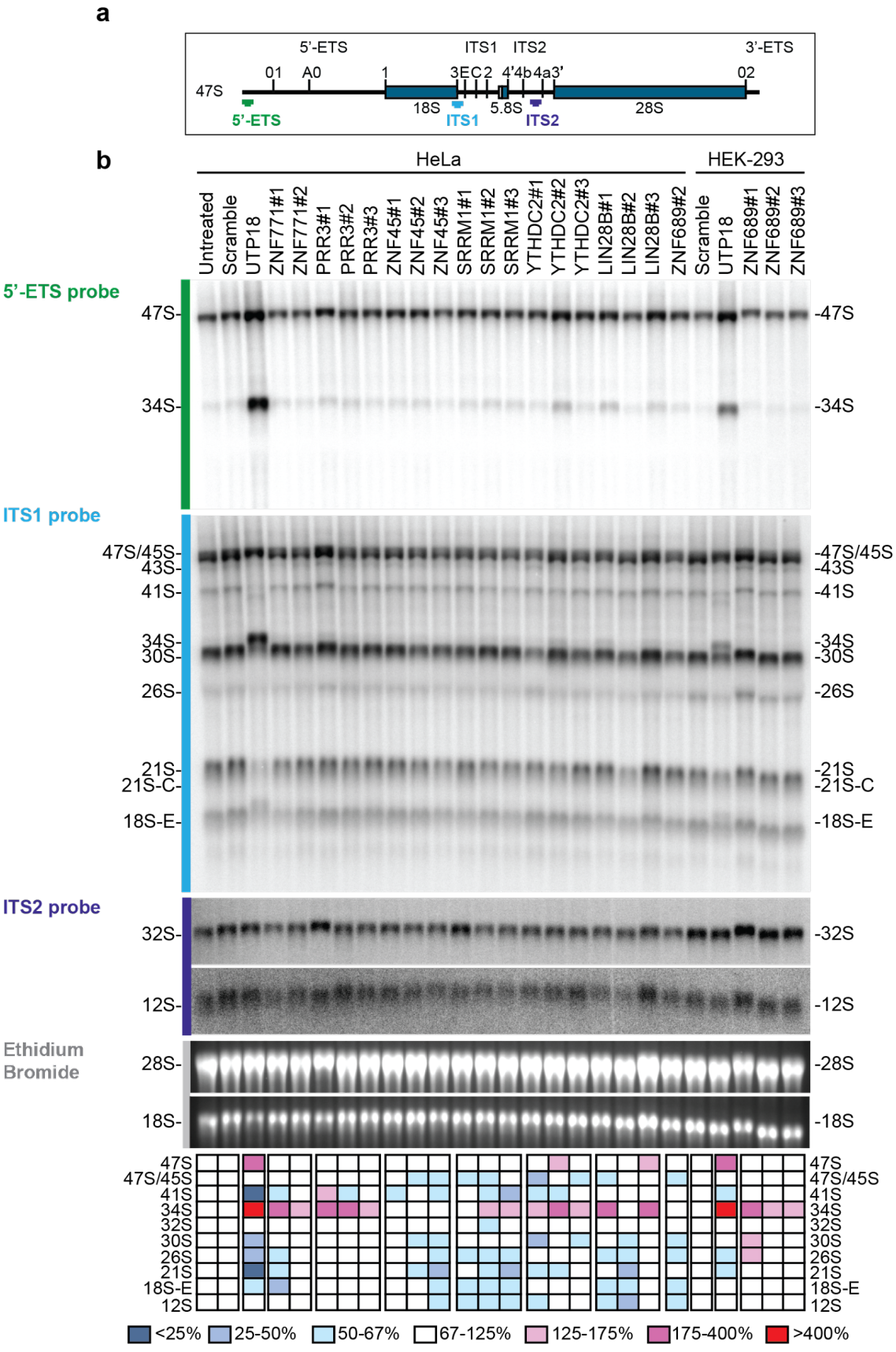
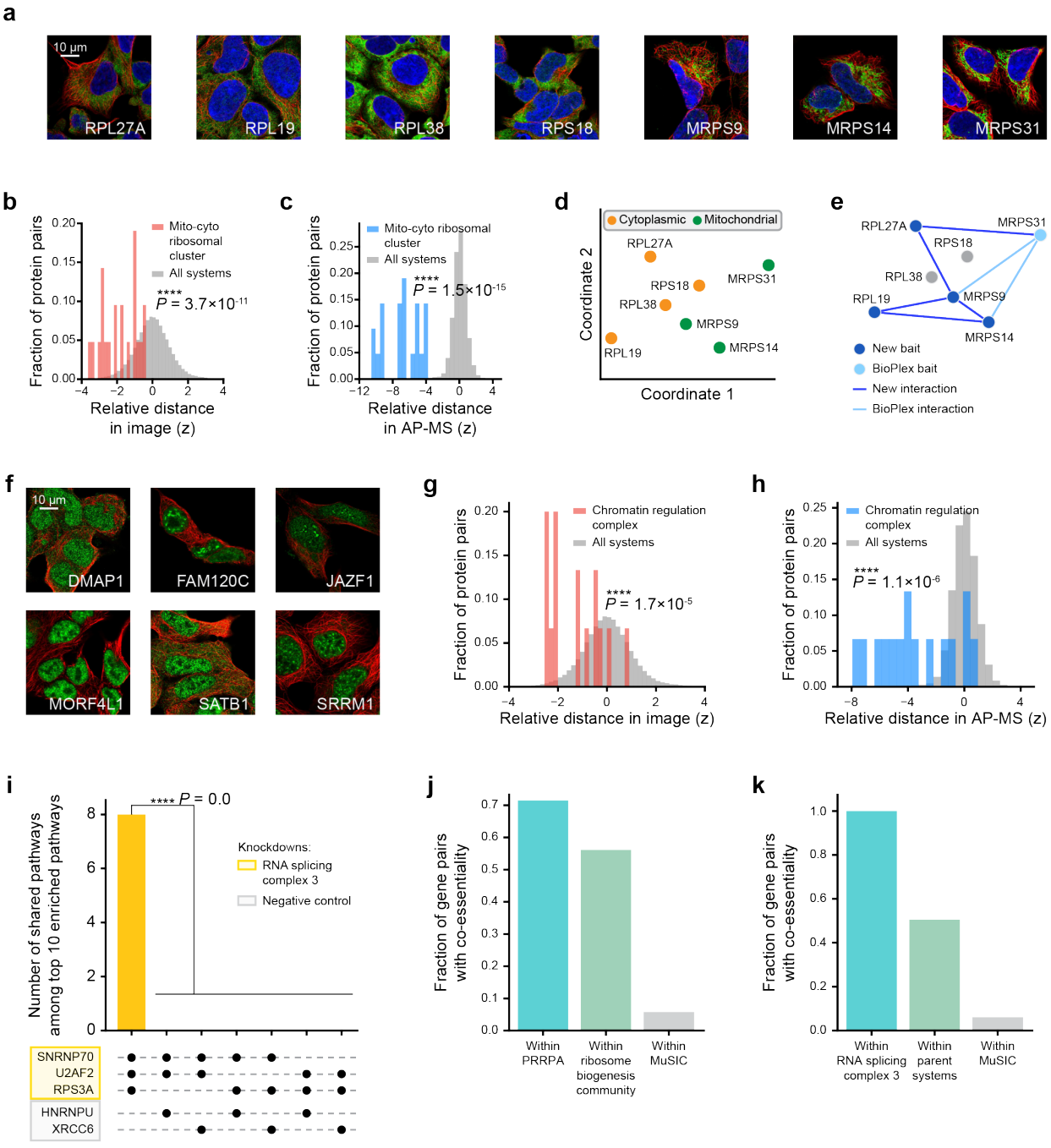

a
5'-ETS
ITS1
ITS2
3'-ETS
01
A0
1
3
E
C
2
4'
4b
4a
3'
02
47S
18S
5.8S
28S
5'-ETS
ITS1
ITS2
b
HeLa
HEK-293
Untreated
Scramble
UTP18
ZNF771#1
ZNF771#2
PRR3#1
PRR3#2
PRR3#3
ZNF45#1
ZNF45#2
ZNF45#3
SRRM1#1
SRRM1#2
SRRM1#3
YTHDC2#1
YTHDC2#2
YTHDC2#3
LIN28B#1
LIN28B#2
LIN28B#3
ZNF689#2
Scramble
UTP18
ZNF689#1
ZNF689#2
ZNF689#3
5'-ETS probe
47S
34S
ITS1 probe
47S/45S
43S
41S
34S
30S
26S
21S
21S-C
18S-E
ITS2 probe
32S
12S
Ethidium Bromide
28S
18S
47S
47S/45S
41S
34S
32S
30S
26S
21S
18S-E
12S
<25%
25-50%
50-67%
67-125%
125-175%
175-400%
>400%

Figure S3.5: Supporting analyses for MuSIC systems. a, IF images showing similar cytoplasmic staining for proteins in "Mito-cyto ribosomal cluster". Cytoplasmic staining is dim for MRPS9, MRPS14 and MRPS31 compared to their predominant mitochondrial locations. Colors represent immunostained protein (green), cytoskeleton (red) and nucleus (blue). **b, c,** Corresponding distributions of protein-protein distance z-scores for IF (**b**, red) or AP-MS (**c**, blue), calibrated to all such distances, respectively (gray). **d,** 2D projection of proteins in this system as in **Figure 3.2a**. Proteins colored according to known affiliations to cytoplasmic ribosome or mitochondrial ribosome. **e**, Validated AP-MS interactions in Mito-cyto ribosomal cluster. Note that only one out of seven proteins was previously tagged as bait in BioPlex v2 (light blue node), thus most physical associations (dark blue edges) among protein pairs were newly identified in this study. **f,** IF images showing similar nucleoplasm and nuclear speckles signals among proteins in the "Chromatin regulation complex." Color as in (**a**). **g, h,** Similar analysis for Chromatin regulation complex as in (**b**) and (**c**). **i,** Comparison among the top 10 pathways (GO Biological Process) returned from Gene Set Enrichment Analysis using the top 500 differentially expressed transcripts. Bar chart shows number of enriched pathways shared by different gene groups indicated by black dots beneath each bar. One-sided one-sample t-test. **j,** Degree of co-essentiality for gene paris within PRRPA (teal bar) shown in comparison to remaining pairs of genes assigned to the more general system that contains it, "Ribosome biogenesis community" (green bar), as well as all other gene pairs in MuSIC (grey bar). **k,** Similar analysis as in (**j**) for "RNA splicing complex 3." Parent systems are "RNA processing complex 1" and "RNA splicing complex family".


a
10 μm
RPL27A
RPL19
RPL38
RPS18
MRPS9
MRPS14
MRPS31
b
Fraction of protein pairs
Mito-cyto ribosomal cluster
All systems
**** P = 3.7×10-11
Relative distance in image (z)
c
Fraction of protein pairs
Mito-cyto ribosomal cluster
All systems
**** P = 1.5×10-15
Relative distance in AP-MS (z)
d
Cytoplasmic
Mitochondrial
Coordinate 2
Coordinate 1
RPL27A
MRPS31
RPS18
RPL38
MRPS9
RPL19
MRPS14
e
RPL27A
MRPS31
RPS18
RPL38
RPL19
MRPS9
MRPS14
New bait
BioPlex bait
New interaction
BioPlex interaction
f
10 μm
DMAP1
FAM120C
JAZF1
MORF4L1
SATB1
SRRM1
g
Fraction of protein pairs
Chromatin regulation complex
All systems
**** P = 1.7×10-5
Relative distance in image (z)
h
Fraction of protein pairs
Chromatin regulation complex
All systems
**** P = 1.1×10-6
Relative distance in AP-MS (z)
i
Number of shared pathways among top 10 enriched pathways
**** P = 0.0
Knockdowns:
RNA splicing complex 3
Negative control
SNRNP70
U2AF2
RPS3A
HNRNPU
XRCC6
j
Fraction of gene pairs with co-essentiality
Within PRRPA
Within ribosome biogenesis community
Within MuSIC
k
Fraction of gene pairs with co-essentiality
Within RNA splicing complex 3
Within parent systems
Within MuSIC

3.5 Author Contributions

Y.Q., E.L. and T.I. designed the study and developed the conceptual ideas. E.L.H., L.P.V., T.Z., J.W.H. and S.P.G. generated AP-MS data and provided FLAG-HA-tagged clones. Y.Q. implemented all computational methods and analyses. S.M.B. and G.W.Y. generated and analyzed RIP-qPCR data. L.W. and D.L.J.L. generated and analyzed northern blot data. C.F.W., A.B. and E.L. generated and analyzed *in situ* fractionation data. M.L.G. and G.W.Y. generated and analyzed eCLIP data. Y.Q., M.C., K.L. and J.J.L. performed the rest of experiments. All authors contributed to developing ideas for data analyses and experimental designs. Y.Q. and T.I. wrote the manuscript with input from all other authors.

3.6 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in *Nature* 2021. Yue Qin, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, Fan Zheng, Michael Chen, Leah V. Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J. Lee, Wei Ouyang, Sophie N. Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F. Kreisberg, Steven P. Gygi, Jianzhu Ma, J. Wade Harper, Gene W. Yeo, Denis L. J. Lafontaine, Emma Lundberg & Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 600, 536–542 (2021). The dissertation author was the primary investigator and author of this paper.

We gratefully acknowledge helpful discussion and comments from Abraham Palmer, Cherie Ng, members of the Ideker laboratory, members of the Lundberg laboratory, the Human Protein Atlas, Jason Swedlow, and the anonymous referees of this work. We thank the Cell Profiling facility and Dr. Charlotte Stadler at the Science for Life Laboratory for help with the *in situ* fractionation. This work was supported by the National Institutes of Health (NIH) under grants P41 GM103504 and R01 HG009979 to T.I., U24 HG006673 to E.L.H., S.P.G, and J.W.H., U41

HG009889 and R01s HL137223 and HG004659 to G.W.Y., R50 CA243885 to J.F.K., by a gift from Google Ventures to J.W.H. and S.P.G., by the Erling-Persson family foundation, Knut and Alice Wallenberg Foundation (2016.0204) and the Swedish Research Council (2017-05327) to E.L., and by the Belgian Fonds de la Recherche Scientifique (F.R.S./FNRS), the Université Libre de Bruxelles (ULB), the European Joint Programme on Rare Diseases ('RiboEurope' and 'DBAcure'), the Région Wallonne (SPW EER) ('RIBOcancer'), the Internationale Brachet Stiftung, and the Epitran COST action (CA16120) to D.L.J.L.

3.7 References

Behrends, C., Sowa, M. E., Gygi, S. P., & Harper, J. W. (2010). Network organization of the human autophagy system. *Nature*, 466(7302), 68–76.

Blencowe, B. J., Baurén, G., Eldridge, A. G., Issner, R., Nickerson, J. A., Rosonina, E., & Sharp, P. A. (2000). The SRm160/300 splicing coactivator subunits. *RNA*, 6(1), 111–120.

Chaudhuri, S., Vyas, K., Kapasi, P., Komar, A. A., Dinman, J. D., Barik, S., & Mazumder, B. (2007). Human ribosomal protein L13a is dispensable for canonical ribosome function but indispensable for efficient rRNA methylation. *RNA*, 13(12), 2224–2237.

De Silva, D., Tu, Y.-T., Amunts, A., Fontanesi, F., & Barrientos, A. (2015). Mitochondrial ribosome assembly in health and disease. *Cell Cycle*, 14(14), 2226–2250.

Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3), 207–214.

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11), 976–989.

Eppens, N. A., Faber, A. W., Rondaij, M., Jahangir, R. S., van Hemert, S., Vos, J. C., Venema, J., & Raué, H. A. (2002). Deletions in the S1 domain of Rrp5p cause processing at a novel site in ITS1 of yeast pre-rRNA that depends on Rex4p. *Nucleic Acids Research*, 30(19), 4222–4231.

Fleckner, J., Zhang, M., Valcárcel, J., & Green, M. R. (1997). U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes & Development*, 11(14), 1864–1872.

Grandori, C., Gomez-Roman, N., Felton-Edkins, Z. A., Ngouenet, C., Galloway, D. A., Eisenman, R. N., & White, R. J. (2005). c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature Cell Biology*, 7(3), 311–318.

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11), 3022–3040.e28.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655), 505–509.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., … Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2), 425–440.

Müllner, D., & Others. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9), 1–18.

North, B. V., Curtis, D., & Sham, P. C. (2002). A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics*, 71(2), 439–441.

Pavan Kumar, P., Purbey, P. K., Sinha, C. K., Notani, D., Limaye, A., Jayani, R. S., & Galande, S. (2006). Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. *Molecular Cell*, 22(2), 231–243.

Piunti, A., Smith, E. R., Morgan, M. A. J., Ugarenko, M., Khaltyan, N., Helmin, K. A., Ryan, C. A., Murray, D. C., Rickels, R. A., Yilmaz, B. D., Rendleman, E. J., Savas, J. N., Singer, B. D., Bulun, S. E., & Shilatifard, A. (2019). CATACOMB: An endogenous inducible gene that antagonizes H3K27 methylation activity of Polycomb repressive complex 2 via an H3K27M-like mechanism. *Science Advances*, 5(7), eaax2887.

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K. W., Li, J., & Nagai, K. (2009). Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*, 458(7237), 475–480.

Sawasdichai, A., Chen, H.-T., Hamid, N. A., Jayaraman, P.-S., & Gaston, K. (2010). In situ subcellular fractionation of adherent and non-adherent mammalian cells. *Journal of Visualized Experiments: JoVE*, 41. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3156070/

Sowa, M. E., Bennett, E. J., Gygi, S. P., & Harper, J. W. (2009). Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2), 389–403.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550.

Tafforeau, L., Zorbas, C., Langhendries, J.-L., Mullineux, S.-T., Stamatopoulou, V., Mullier, R., Wacheul, L., & Lafontaine, D. L. J. (2013). The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of Pre-rRNA processing factors. *Molecular Cell*, 51(4), 539–551.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818), 711–719.

Van Nostrand, E. L., Nguyen, T. B., Gelboin-Burkhart, C., Wang, R., Blue, S. M., Pratt, G. A., Louie, A. L., & Yeo, G. W. (2017). Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP). *Methods in Molecular Biology*, 1648, 177–200.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6), 508–514.

Yang, X., Boehm, J. S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S. R., Alkan, O., Bhimdi, T., Green, T. M., Johannessen, C. M., Silver, S. J., Nguyen, C., Murray, R. R., Hieronymus, H., Balcha, D., Fan, C., Lin, C., … Root, D. E. (2011). A public genome-scale lentiviral expression library of human ORFs. *Nature Methods*, 8(8), 659–661.

Yoshikatsu, Y., Ishida, Y.-I., Sudo, H., Yuasa, K., Tsuji, A., & Nagahama, M. (2015). NVL2, a nucleolar AAA-ATPase, is associated with the nuclear exosome and is involved in pre-rRNA processing. *Biochemical and Biophysical Research Communications*, 464(3), 780–786.

EPILOGUE

4.1 Discussion

In classical image analysis of cells, locations of proteins are identified in reference to a handful of subcellular markers. Measurements of pairwise protein proximity are made by fluorescently labeling multiple proteins in the same image (Stryer, 1978), a combinatorial process that is difficult to scale to more than a few proteins. Here we have developed a systematic means of measuring pairwise protein proximity, *via* the neural network embeddings of each individual immunofluorescent protein. In turn, this systematic accumulation of proximity measurements allows us to move from a closed library of subcellular locations to an open approach in which both existing and new structures are identified *de novo* from inherent structure in data. Such image analysis can be integrated with other data modalities, as we demonstrated here with AP-MS, leading to recovery of biological structures across a range of scales (**Figure 1.3a**) and discoveries of novel protein communities which we have physically and functionally validated (**Figures. 3.1, 3.2**). Moreover, while the imaging field is accustomed to thinking about physical sizes and intracellular distances, the notion that protein interactions can provide a complementary measure of intracellular distance is, to our knowledge, new to this study.

What about when the data disagree? While nearly a third of observed AP-MS interactions fall within the same focused system of <100 proteins, more than two thirds do not (**Figure 2.1e**). For example, PPP6R1, a phosphatase, and NPAS1, a helix-loop-helix transcription factor, interact directly by AP-MS but were placed in different organelles in MuSIC map related to their distinct image locations (PPP6R1, Cytoplasm; NPAS1, Nucleus; **Figure 4.1a**). Such discrepancies may indicate rare physical association of the proteins in a common compartment despite more abundant locations in distinct others, as might be expected from pleiotropic roles or stages of protein

maturation (Alberts et al., 2002). Alternatively, the two proteins could interact transiently or periodically (e.g., cell cycle, circadian rhythms) or discrepancies might derive from independent errors or biases, such as the fact that IF detects endogenous proteins whereas AP-MS detects over-expressed tagged proteins. Some disagreement between IF and AP-MS can clearly be tolerated by the system, such as correct assignment of GEMIN7 and SNRNP70 to the U1 snRNP (Gubitz et al., 2004; Yong et al., 2002) (**Figure 1.3g**), despite only partial overlap in the image distributions of the two proteins (**Figure 4.1b**). In this case, correct assignment was facilitated by the physical associations of these proteins observed by AP-MS.

Systems in MuSIC reside at a variety of physical scales, bridging and exceeding the ranges of IF and AP-MS (**Figure 2.1a-d**). What does it mean for two components to occupy different scales in the map? Although tightly correlated with the number of protein species, here the scale of a component reflects the estimated proximities among protein members. For example, analysis of protein-protein proximities at broad scale identified the pre-catalytic spliceosome, whereas decreasing the distance threshold led to the recovery of two smaller scale components, the U1 and U2 snRNPs (**Figure 1.3f, g**). Just as the physical proximity becomes weaker with increasing scale, one would expect the same to be true for functional association. Such a structure-function relationship indeed has some support within the ribosome biogenesis and RNA splicing branches of MuSIC explored in this study: within each of these areas, gene co-essentiality, a measure of joint protein function (Wang et al., 2017), is strongest among genes assigned to the same small systems, weaker within the larger scale systems that contain them, and near zero for unrelated groups of genes (**Figure S3.5j, k**). Components at different physical scales may also map naturally to different types of assay for functional exploration. For example, we used the 28S/18S rRNA ratio as a general readout affected by most proteins belonging to the Ribosome biogenesis

community. Conversely, probing specific rRNA precursors can implicate specific ribosome biogenesis subsystems, such as binding of a protein to 45S pre-rRNA (suggesting involvement in early-stage ribosome biogenesis, **Figure 3.1e**) or changes in 34S pre-rRNA abundance resulting from a protein knockdown (suggesting an early maturation defect associated with loss of function of small-subunit processome, **Figure S3.4**, **Figure 3.1g**) (Tafforeau et al., 2013). We expect future validation of MuSIC systems to draw from a broad range of functional assays at the molecular, pathway, and cellular level.

As the MuSIC map is further developed to cover all >20,000 proteins, a key question will be how to handle cellular dynamics. While some MuSIC systems correspond to constitutive cellular structures, others correspond to dynamic assembly pathways such as those related to ribosome biogenesis (**Figure 3.1**). In these cases, is it preferable to work towards a single unified map of human cell components, a strategy taken by the Gene Ontology project (Ashburner et al., 2000; The Gene Ontology Consortium, 2019), or to create separate maps that capture differing architectures across cell types and states? We believe that an attractive middle road is to create one, or no more than a few, reference maps of widely conserved cell components, with context-specific additions or deletions indicated as annotations. For specific cellular contexts that prove very different from the norm, separate maps could be constructed and served alongside the major ones.

Regardless, here we focused on HEK293-derived cells, a widely used model system for gaining biological insights that are generally applied (Go et al., 2021; Huttlin et al., 2015, 2017; Thul et al., 2017). Previous studies have shown that approximately 70% of proteins have consistent localization across cell lines (Thul et al., 2017) and that about 50% maintain their physical interactions (Huttlin et al., 2021); thus we expect that a cell reference map in HEK293 will partially

generalize to other cell types and states, with attention paid to communities prone to dynamics. Notably, many systems in the MuSIC map are significantly co-regulated in protein expression across diverse human cell lines (**Figure 4.1c**), suggesting these systems are indeed relevant to other biological contexts.

As these maps evolve, we note the synergy achieved in integrating HPA and BioPlex, two large-scale mapping efforts that might have otherwise progressed independently. Such coordination should continue and might also encompass collaborative dataset design, for instance by adopting a common set of human cell lines and proteins targeted across projects. Furthermore, new protein systems might arise with inclusion of yet additional modalities of data, such as proximity-dependent labeling (Gingras et al., 2019; Go et al., 2021; Kalocsay, 2019; Rhee et al., 2013; Varnaitė & MacNeill, 2016; Youn et al., 2018), cross-linking mass spectrometry (Leitner et al., 2016) or cryo-electron microscopy (Rout & Sali, 2019). It will be interesting to explore synergies with these other platforms, all of which might be calibrated to measure molecular distances and, in turn, contribute to integrated maps of the multi-scale cell.

4.2 Figures

Figure 4.1: Heterogeneity in the MuSIC map. a, b, Examples of proteins with strong AP-MS protein interactions that have very different IF localization patterns. Colors represent immunostained protein (green) and cytoskeleton (red). **c,** Protein co-abundance for MuSIC systems, calculated from the median Pearson correlation of pairwise protein abundance over 375 diverse cell lines (Nusinow et al., 2020). The plot shows all systems with less than 20 proteins, having co-abundance measurements for >50% of protein pairs. Significance was assessed empirically, using 1000 randomized MuSIC hierarchies, followed by Benjamini-Hochberg multiple test correction to obtain FDR (color of bar). Protein co-abundance for a system provides evidence for its presence in cell types beyond HEK293. Larger systems tend to correspond to high-level compartments and organelles found in most human cells.


a
b
c
10 µm
PPP6R1
NPAS1
10 µm
GEMIN7
SNRNP70
Large ribosomal subunit subcomplex 4
Large ribosomal subunit subcomplex 3
Cotranslational protein targeting to membrane system
Ribosomal complex 1
Ribosomal complex 5
Ribosomal complex 2
Ribosomal complex 3
RNA splicing complex 2
Splicing regulatory complex
U2-type spliceosomal complex family
Ribosomal complex 4
Subgroup of spliceosomal complex family
Spliceosomal complex family
Catalytic step 2 spliceosome
Large ribosomal subunit subcomplex 2
U2 snRNP
U1 snRNP
PRRPA
Pre-catalytic spliceosome
RNA splicing complex 1
Mitochondrion
Small nuclear ribonucleoprotein complex family
Pre-catalytic spliceosome subcomplex 1
Chromatin regulation assembly
Mito-cyto ribosomal cluster
NuA4 histone acetyltransferase complex
RNA splicing complex 3
Polysomal ribosome complex
Pre-mRNA splicing complex 1
Mitochondrial large ribosomal subunit
Histone acetyltransferase complex family
DNA metabolic assembly
Ribosome biogenesis subcomplex
Chromosome organization complex family
Large ribosomal subunit subcomplex 1
Nucleoplasm 1
Chromatin organization complex family
Chromatin regulation complex
Nuclear transcriptional speckle
Nucleic acid binding complex
Negative regulation of RNA biosynthetic process
Ribosomal complex 6
RNA processing complex 1
RNA processing complex 2
FDR
[0, 0.01]
(0.01, 0.05]
(0.05, 0.1]
Non-significant
0.0
0.2
0.4
0.6
0.8
Median protein co-abundance
(Pearson's r, 375 cell lines)

4.3 Author Contributions

Y.Q., E.L. and T.I. designed the study and developed the conceptual ideas. All authors contributed to developing ideas for data analyses and experimental designs. Y.Q. and T.I. wrote the manuscript with input from all other authors.

4.4 Acknowledgements

Epilogue, in full, is a reprint of the material as it appears in *Nature* 2021. Yue Qin, Edward L. Huttlin, Casper F. Winsnes, Maya L. Gosztyla, Ludivine Wacheul, Marcus R. Kelly, Steven M. Blue, Fan Zheng, Michael Chen, Leah V. Schaffer, Katherine Licon, Anna Bäckström, Laura Pontano Vaites, John J. Lee, Wei Ouyang, Sophie N. Liu, Tian Zhang, Erica Silva, Jisoo Park, Adriana Pitea, Jason F. Kreisberg, Steven P. Gygi, Jianzhu Ma, J. Wade Harper, Gene W. Yeo, Denis L. J. Lafontaine, Emma Lundberg & Trey Ideker. A multi-scale map of cell structure fusing protein images and interactions. *Nature* 600, 536–542 (2021). The dissertation author was the primary investigator and author of this paper.

We gratefully acknowledge helpful discussion and comments from Abraham Palmer, Cherie Ng, members of the Ideker laboratory, members of the Lundberg laboratory, the Human Protein Atlas, Jason Swedlow, and the anonymous referees of this work. This work was supported by the National Institutes of Health (NIH) under grants P41 GM103504 and R01 HG009979 to T.I., U24 HG006673 to E.L.H., S.P.G, and J.W.H., U41 HG009889 and R01s HL137223 and HG004659 to G.W.Y., R50 CA243885 to J.F.K., by a gift from Google Ventures to J.W.H. and S.P.G., by the Erling-Persson family foundation, Knut and Alice Wallenberg Foundation (2016.0204) and the Swedish Research Council (2017-05327) to E.L., and by the Belgian Fonds de la Recherche Scientifique (F.R.S./FNRS), the Université Libre de Bruxelles (ULB), the European Joint Programme on Rare Diseases (‘RiboEurope’ and ‘DBAcure’), the Région

Wallonne (SPW EER) ('RIBOcancer'), the Internationale Brachet Stiftung, and the Epitran COST action (CA16120) to D.L.J.L.

4.5 References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Compartmentalization of Cells. *Garland Science*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29.

Gingras, A.-C., Abe, K. T., & Raught, B. (2019). Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Current Opinion in Chemical Biology*, 48, 44–54.

Go, C. D., Knight, J. D. R., Rajasekharan, A., Rathod, B., Hesketh, G. G., Abe, K. T., Youn, J.-Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L. Y., Popiel, E., Lambert, J.-P., Coyaud, É., Cheung, S. W. T., Rajendran, D., Wong, C. J., Antonicka, H., Pelletier, L., Palazzo, A. F., … Gingras, A.-C. (2021). A proximity-dependent biotinylation map of a human cell. *Nature*.

Gubitz, A. K., Feng, W., & Dreyfuss, G. (2004). The SMN complex. *Experimental Cell Research*, 296(1), 51–56.

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11), 3022–3040.e28.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655), 505–509.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., … Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2), 425–440.

Kalocsay, M. (2019). APEX Peroxidase-Catalyzed Proximity Labeling and Multiplexed Quantitative Proteomics. *Methods in Molecular Biology*, 2008, 41–55.

Leitner, A., Faini, M., Stengel, F., & Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences*, 41(1), 20–32.

Nusinow, D. P., Szpyt, J., Ghandi, M., Rose, C. M., McDonald, E. R., 3rd, Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D. K., Jedrychowski, M., Golji, J., Porter, D. A., Rejtar, T., Wang, Y. K., Kryukov, G. V., Stegmeier, F., Erickson, B. K., Garraway, L. A., Sellers, W. R., & Gygi, S. P. (2020). Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell*, 180(2), 387–402.e16.

Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., & Ting, A. Y. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science*, 339(6125), 1328–1331.

Rout, M. P., & Sali, A. (2019). Principles for Integrative Structural Biology Studies. *Cell*, 177(6), 1384–1403.

Stryer, L. (1978). Fluorescence energy transfer as a spectroscopic ruler. *Annual Review of Biochemistry*, 47, 819–846.

Tafforeau, L., Zorbas, C., Langhendries, J.-L., Mullineux, S.-T., Stamatopoulou, V., Mullier, R., Wacheul, L., & Lafontaine, D. L. J. (2013). The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of Pre-rRNA processing factors. *Molecular Cell*, 51(4), 539–551.

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1), D330–D338.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, 356(6340).

Varnaitė, R., & MacNeill, S. A. (2016). Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics*, 16(19), 2503–2518.

Wang, T., Yu, H., Hughes, N. W., Liu, B., Kendirli, A., Klein, K., Chen, W. W., Lander, E. S., & Sabatini, D. M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell*, 168(5), 890–903.e15.

Yong, J., Pellizzoni, L., & Dreyfuss, G. (2002). Sequence-specific interaction of U1 snRNA with the SMN complex. *The EMBO Journal*, 21(5), 1188–1196.

Youn, J.-Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., Bagci, H., Rathod, B., MacLeod, G., Eng, S. W. M., Angers, S., Morris, Q., Fabian, M., Côté, J.-F., & Gingras, A.-C. (2018). High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell*, 69(3), 517–532.e11.