A proximity biotinylation map of a human cell

Christopher D. Go[1,2*], James D.R. Knight[1*], Archita Rajasekharan[3], Bhavisha Rathod[1], Geoffrey G. Hesketh[1], Kento T. Abe[1,2], Ji-Young Youn[1], Payman Samavarchi-Tehrani[1], Hui Zhang[4], Lucie Y. Zhu[4], Evelyn Popiel[2], Jean-Philippe Lambert[1,5], Étienne Coyaud[6,7], Sally W.T. Cheung[1], Dushyandi Rajendran[1], Cassandra J. Wong[1], Hana Antonicka[3], Laurence Pelletier[1,2], Brian Raught[6,8], Alexander F. Palazzo[4], Eric A. Shoubridge[3], and Anne-Claude Gingras[1,2]

[1] Lunenfeld-Tanenbaum Research Institute, Toronto, ON, Canada
[2] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[3] Montreal Neurological Institute and Department of Human Genetics, McGill University, Montreal, QC, Canada
[4] Department of Biochemistry, University of Toronto, Toronto, ON, Canada.
[5] Present address: Department of Molecular Medicine and Cancer Research Centre, Université Laval, Quebec, Canada; CHU de Québec Research Center, CHUL, Quebec, QC, Canada.
[6] Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
[7] Present address: PRISM INSERM U1192, Université de Lille, Villeneuve d'Ascq, France
[8] Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

* These authors contributed equally to the manuscript

Data deposition: Mass spectrometry data have been deposited in the Mass spectrometry Interactive Virtual Environment (MassIVE, http://massive.ucsd.edu).

Correspondence:
e-mail : gingras@lunenfeld.ca, Phone: (416) 586-5027 Fax: (416) 586-8869

INTRODUCTION

Compartmentalization is an essential characteristic of eukaryotic cells, ensuring that cellular processes are partitioned to defined subcellular locations. High throughput microscopy[1] and biochemical fractionation coupled with mass spectrometry[2-6] have helped to define the proteomes of multiple organelles and macromolecular structures. However, many compartments have remained refractory to such methods, partly due to lysis and purification artefacts and poor subcompartment resolution. Recently developed proximity-dependent biotinylation approaches such as BioID and APEX provide an alternative avenue for defining the composition of cellular compartments in living cells (e.g. [7-10]). Here we report an extensive BioID-based proximity map of a human cell, comprising 192 markers from 32 different compartments that identifies 35,902 unique high confidence proximity interactions and localizes 4,145 proteins expressed in HEK293 cells. The recall of our localization predictions is on par with or better than previous large-scale mass spectrometry and microscopy approaches, but with higher localization specificity. In addition to assigning compartment and subcompartment localization for many previously unlocalized proteins, our data contain fine-grained localization information that, for example, allowed us to identify proteins with novel roles in mitochondrial dynamics. As a community resource, we have created humancellmap.org, a website that allows exploration of our data in detail, and aids with the analysis of BioID experiments.

BODY

Proximity-dependent labelling approaches have rapidly grown in popularity, as they provide a robust way to label the environment in which a protein resides in living cells[7,8]. In the most widely used of these techniques, BioID, a mutant *E. coli* biotin ligase – BirA* (R118G) – is fused in-frame with the coding sequence of a bait polypeptide of interest, and the resulting fusion protein expressed in cultured cells. While BirA* can activate biotin to biotinoyl-AMP, the abortive mutant enzyme exhibits a reduced affinity for the activated molecule. A reactive intermediate is thus released into the local environment that can react with free epsilon amine groups on nearby lysine residues[7]. This ability for BirA* to label a local environment has led to BioID being employed by multiple laboratories to define the composition, and in some cases the overall organization, of both membrane-bound and membraneless organelles (e.g. [7-10]).

Here, we set out to map a human cell by profiling markers (consisting of full-length proteins or targeting sequences) from 32 cellular compartments. These compartments include the cytosolic face of all membrane-bound organelles, the ER lumen, subcompartments of the nucleus and mitochondria, major membraneless organelles such as the centrosome and the nucleolus, and the main cytoskeletal structures (actin, microtubules and intermediate filaments). Several proteins were also queried throughout the endomembrane system to identify components enriched at locales along its continuum (e.g. early versus late endosomes).

A total of 234 candidate markers were selected from the literature with the goal of having several independent markers for each subcellular compartment (**Figure 1A; Supplementary Table 1)**. Each of these compartment markers was tagged with BirA*, stably integrated in HEK293 Flp-In T-REx cells, and processed for BioID (see Methods for details). SAINTexpress[11]

was used to identify high confidence proximity interactors by scoring against a set of negative controls that localize non-specifically to the cytoplasm and nucleus. Only high-confidence interactors (i.e. those passing a 1% FDR threshold) were considered for downstream analysis (**Supplementary Table 2**). Reproducibility across replicate analysis of the same marker was high overall, with a mean R^2 of 0.95 (**Supplementary Figure 1**). Quality control for each marker included immunofluorescence microscopy to confirm expected localization (**Figure 1B; Supplementary Table 1)** and Gene Ontology (GO) enrichment analysis of the resulting high-confidence proximity interactors to ensure enrichment of the expected cellular component (CC) terms (**Supplementary Table 3**). With the notable exception of the Golgi lumen (for which all baits selected for profiling remained trapped in the ER), all selected compartments were successfully characterized with multiple baits. Altogether, 192 of the candidate markers passed quality control, identifying 7192 prey proteins prior to filtering (**Figure 1C**).

Collectively these baits yielded 4,424 high confidence proximity interactors at a 1% FDR. 8.9% of these interactions were previously reported in either the BioGRID[12] or IntAct[13] protein interaction databases. However, if only the top 25 most abundant preys for each bait are considered, 21.1% of interactions were previously reported (**Supplementary Table 4**). This is consistent with BioID being primarily a proximity labelling technique, where the most proximal (and therefore most extensively labelled) preys should be enriched for direct interactors and protein complex components.

Comparison of bait BioID profiles produced a clustering of baits generally consistent with their expected compartments (**Supplementary Figure 2**), further attesting to the overall quality of the dataset. As expected from previous studies[8,14], baits that profile different faces of the same organelle show distinct proximity profiles, since labelling is restricted to the side exposed to the enzyme. Using the Jaccard index as a measure of similarity (where 1 indicates complete overlap of the preys recovered and 0 no overlap on a linear scale), we find a median Jaccard index of 0.541 amongst the eight baits used to profile the ER lumen and a much smaller median Jaccard index of 0.069 between the ER lumen and 19 ER membrane-bound cytosolic facing baits. Similarly, in the mitochondria, the median Jaccard index for the five matrix baits is 0.587, while the median value between the matrix and six non-matrix mitochondrial baits is 0.065. This suggests that our approach is successfully achieving sub-compartment resolution.

To localize prey proteins, we employed a strategy we previously used to define proteins localizing to P-bodies and stress granules[10]. This approach exploits the correlated behaviour of untagged, endogenous preys as they are profiled across multiple baits: preys that have the same signature across multiple baits are more likely to be close to each other, for example within the same organelle or structure (**Figure 1D**). A straightforward utilization of this principle is to perform pairwise correlation between prey proteins from their bait signatures, clustering the results and manually analyzing cluster composition. This process identified clearly defined and expected (sub)compartments (**Figure 1E**; **Supplementary Table 5** for diagonal and off diagonal cluster annotation). As we have previously described[10], a more sophisticated approach to annotate prey localizations uses spatial analysis of functional enrichment (SAFE)[15] and Non-Negative Matrix factorization (NMF)[16], and we applied those same techniques here (see **Figure**

2A and Methods for details). Using these pipelines, SAFE localized 3,252 of the 4,424 high confidence prey proteins to 23 compartments (**Supplementary Figure 3; Supplementary Table 6)**, while NMF localized 4,145 preys to 20 compartments (**Figure 2B; Supplementary Table 7)**. 54% of localizations assigned by SAFE have previously been reported in GO and 50% for NMF. When both SAFE and NMF made a prediction, they were consistent in 88% of cases (**Supplementary Table 8**). While we initially targeted 32 compartments and subcompartments for mapping, we did not resolve all of them in our analysis. This was particularly evident for subcompartments associated with (endo)membranes, consistent with the fluid nature of these compartments that result in baits localizing or labelling more than their primary (selected) compartment. Accurately defining profiles for these subcompartments will require additional BioID on markers for those compartments, or chemical/genetic perturbation experiments, so that subtle differences distinguishing the primary residence of prey proteins can be detected.

This caveat aside, compartments that were identified by our NMF and SAFE pipelines show enrichment for expected domains and motifs (**Supplementary Table 9**). For example, for NMF, the plasma membrane shows enrichment for PH, immunoglobulin, RhoGAP, RhoGEF and tyrosine kinase domains (4.4 to 7.6-fold enrichment) while the adjacent cell junction compartment enriches for PDZ and FERM domains (8.9 to 13.1-fold). The various nuclear sub-compartments are similarly distinguished by their enriched domains, with the chromosome compartment enriched for the KRAB domain, C2H2 zinc fingers, bromodomains and PWWP domain (9.9, 5.3, 6.3 and 8.3-fold, respectively), the nucleolus for the DEAD and helicase domain (9.7 and 6.0-fold) and the nuclear body compartment for the RNA Recognition Motif (RRM) and G-patch domains (10.1 and 8.8-fold). Interestingly, there were almost no enriched domains shared between the compartments. Compartments also showed clear enrichment for specific motifs (coiled-coiled, disordered, signal peptides and transmembrane), and, in contrast with domains, these were often shared between compartments (**Supplementary Figure 3B and C**).

We next compared our predictions with those made by microscopy and fractionation studies. After removing Human Protein Atlas (HPA[1], www.proteinatlas.org) annotations from GO to prevent self-validation for that dataset, the recovery of known protein localizations for NMF and SAFE was similar to HPA, but our approach performed better than fractionation approaches (**Figure 2C**, **Supplementary Table 10**). However, this analysis ignores the fact that gross localization annotations (e.g. nucleus versus cytoplasm) are more likely to be known for a protein than specific sub-compartment localizations. Consistent with this, making more specific predictions negatively impacts the ability to recover known localizations, regardless of the technique employed (**Supplementary Figure 4A**). To account for this, we binned localizations into precision tiers using the information content of the corresponding GO term (see Methods for details), with tier 1 being the most specific localizations (containing terms such as “peroxisome” and “spliceosome”) and tier 5 the least specific (e.g. “cytoplasm” and “nucleus”). By binning localizations into tiers, we found that our predictions are much more specific than those of other approaches with, for example, 73% of proteins localized to the tier 3 bin or better for NMF and 54% for SAFE, versus 17-25% for the other data sets (**Figure 2D; Supplementary Table 10**). High recall of known localizations with increased localization

specificity is, therefore, a marked advantage of our approach. With our localization methodologies, the likelihood of successfully localizing a prey increases as the number of baits it was detected with increases, while the detected abundance (spectral count) does not appear to have a significant impact (**Supplementary Figure 4B and C**). As we expand this project in the future to incorporate additional baits, we can expect further refinement in our ability to correctly localize preys.

To assess the accuracy of our localization predictions, we experimentally tested our localization predictions for several previously uncharacterized or poorly characterized groups of proteins, including the solute carriers ("SLC"), Rab family GTPases, transmembrane ("TMEM") proteins and uncharacterized proteins ("ORF" and "FAM"). The localization of 66 proteins was assessed by immunofluorescence of transiently expressed GFP-tagged constructs or by reciprocal BioID. 86% (57/66) of the predicted localizations were supported by one or both approaches (**Figure 3A, Supplementary Figure 5; Supplementary Table 11**). These results support our localization assignments and the ability for our analysis pipelines to correctly predict the localization of poorly characterized proteins.

Multiple proteins in our dataset have NMF profiles that suggest they may localize to multiple compartments (**Supplementary Table 7**), a phenomenon that could partially be due to proteins moonlighting between distinct compartments as discussed by others[1,17]. Alternatively, these could represent proteins that localize to, or exchange between contiguous compartments (e.g. cell junction and plasma membrane, early and late endosomal components, nucleoplasm and chromatin or nuclear bodies), or that are found at contact sites between compartments, a process that is becoming increasingly appreciated[18]. For example, contacts between the mitochondria and ER are critical for lipid and calcium exchange and play critical roles in mitochondrial dynamics[19,20,21]. To investigate whether our dataset could provide evidence for new proteins at mitochondria-ER contact sites, we selected proteins with NMF scores indicating a primary localization in the ER membrane or mitochondrial outer membrane, evidence for a secondary localization in the opposite compartment, and relatively low evidence for localization in other compartments. This revealed a list of 17 proteins, that included both SAR1A and SAR1B, two GTPases associated with ER exit sites that have recently been shown to regulate the size of mitochondria-ER contact sites[22], and RMDN3 (aka PTPIP51), a protein that participates as a mitochondria-ER tether with VAPB to regulate autophagy through calcium signaling[23,24] (**Figure 3B**). Consistent with the extensive lipid transfer between the mitochondria and ER, six of these proteins have functions in lipid and cholesterol homeostasis, and four in calcium signalling **(Supplementary Table 12)**. For further study, we selected proteins with an existing ER literature annotation (SAR1A, SAR1B, APOL2, C18orf32, CHMP7 and PPP1R15B), and performed BioID to localize them more precisely. While strongly enriching for expected ER components, SAR1A, SAR1B, C18orf32 and CHMP7 (and to a much lesser extent APOL2 and PPP1R15B) recovered major outer mitochondrial proteins such as AKAP1, MAVS, HK1 and OCIAD1, as well as other predicted ER-mitochondrial candidates found in **Figure 3B**, further validating their likely localization to ER-mitochondrial contact sites (**Supplementary Figure 6; Supplementary Table 12**). Interestingly, these baits also detected the mitochondria fission components DNM1L

(orthologous to yeast Drp1[25]) and INF2 (a formin that mediates actin-dependent fision[26]), suggesting they may play a role in mitochondrial fusion/fission events.

To test whether the proteins identified above may be implicated in mitochondrial dynamics, we first expressed GFP-tagged versions of these proteins and quantified the percentage of cells with fragmented mitochondrial morphology. In this assay, MARCH5/MITOL (a ubiquitin E3 ligase that positively regulates membrane fission[27]) and DNM1L served as positive controls and induced mitochondrial fragmentation as expected (**Figure 3C, 3D**). Expression of three other proteins, APOL2 (apolipoprotein L2), C18orf32 (a protein that traffics to lipid droplets[28]) and CHMP7 (an ESCRT-III component) also strongly induced mitochondrial fragmentation. To evaluate the consequence of depletion of the two strongest hits on mitochondrial morphology, C18orf32 and CHMP7, we used siRNA-mediated depletion. C18orf32 and CHMP7 depletion induced a striking hyperfused mitochondrial phenotype, suggesting that they are important in mitochondrial homeostasis (**Figure 3E, 3F**). How these proteins participate in mitochondrial dynamics remains to be defined.

To allow the community to explore and benefit from our data, and to have a repository for the BioID data generated by our lab and collaborators going forward, we created the humancellmap.org. Here the community can search and view data on profiled baits, identified preys and organelles, and explore interactive 2D maps/networks of the NMF and SAFE data. Help documentation describing all available features at the site can be found at humancellmap.org/help. As we profile and incorporate more baits in the future, the site will be updated as new data is available.

A key feature of the site is the ability to upload user BioID data and compare it against our database. This can help to localize a query bait to specific cell compartments based on its prey similarity signature to our dataset, and help identify those preys that are most specific to the bait queried.

To illustrate this capability, we performed BioID on a regulatory subunit of PI3 kinase (PIK3R1), an SH2 domain-containing adaptor that recruits PI3 kinase to activated receptor complexes at the plasma membrane. Analyzing the BioID data through the analysis module of the humancellmap.org revealed that the most similar baits in our database localize to the plasma membrane and cell junction. Importantly, while 27 high-confidence proximity interactions were detected with this bait, the specificity metric revealed highly specific proximity interactions with PI3 kinase catalytic subunits, insulin receptor substrate proteins (IRS2, IRS4) as well as to the scaffold protein GRB2, as expected[29] (**Supplementary Figure 7A, 7B**). We also reanalyzed a previously published BioID bait, RNGTT[10], a nuclear protein involved in mRNA capping. The analysis module reports a nuclear localization, with specific interactions including several RNA Polymerase II subunits and components of the catalytic subunit of the PP4 phosphatase, as previously reported by affinity purification coupled to mass spectrometry[30,31] (**Supplementary Figure 7C**).

This strategy can be expanded to the analysis of poorly characterized proteins. For instance, we performed BioID on some of the uncharacterized proteins for which we predicted a localization in this study (**Supplementary Table 13**), and used the analysis module at the humancellmap.org to aid interpretation of the results. Both FAM171A1 and FAM171B were predicted to localize to the cell junction and plasma membrane and, consistent with this, their BioID profiles were most similar to junctional and plasma membrane baits, while specific preys included several cytoskeletal proteins, inline with a previous study that showed a reduction of actin stress fibers following knockdown of FAM171A1[32] (**Supplementary Figure 8A**). Similarly, MTFR2 (FAM54A) was associated with the mitochondrial outer membrane and peroxisome as a prey protein, with a weak signature at the mitochondrial inner membrane/mitochondrial intermembrane space. As a bait, the analysis module reports that it is most similar to peroxisomal baits, followed by mitochondrial outer and inner membrane baits, supporting its predicted localization. MTFR1, SLC25A46 and VPS13D were found to be highly specific interactions to MTFR2, consistent with the mitochondrial fragmentation previously observed upon overexpression of MTFR2[33] (**Supplementary Figure 8B**). These results further attest to the applicability of the humancellmap.org resource for the exploration of BioID datasets.

Globally, this first version of the humancellmap.org provides a framework for the interpretation of proximity-dependent biotinylation data, and the exploration of subcellular neighbourhoods by cell biologists. Our own data can be explored directly through the site, or it can be employed to assist in the interpretation of a user's BioID experiments as outlined above. While the current version of the humancellmap.org is a static view of a single cell type (HEK293) using a relatively small yet well characterized set of baits, future versions will explore higher density coverage of baits (i.e. by merging organelle-specific datasets within the humancellmap.org), other cell types, and dynamic events, thereby supplementing other proteomics and cell biological resources.

Acknowledgements

We thank Zhen-Yuan Lin for profiling PIK3R1 and members of the Gingras lab for helpful discussion and advice throughout the project, and Jin Zhang and many cell biologists for helpful suggestions regarding bait selection.

Work in the Gingras lab was supported by a Canadian Institutes of Health Research (CIHR) Foundation Grant (FDN 143301). EAS is supported by a grant from the CIHR (MOP-133530). Proteomics work was performed at the Network Biology Collaborative Centre at the Lunenfeld-Tanenbaum Research Institute, a facility supported by Canada Foundation for Innovation funding, by the Ontario Government, and by Genome Canada and Ontario Genomics (OGI-139). This research was enabled in part by support provided by Compute Canada (www.computecanada.ca). C.D.G. was supported by a CIHR Banting studentship. A.-C.G. is the Canada Research Chair in Functional Proteomics and the Lea Reichmann Chair in Cancer Proteomics.

Author Contributions

A.-C.G., C.D.G. and J.D.R.K. conceived the project.

A.-C.G., C.D.G. and J.D.R.K. wrote the paper with input from B.R., G.G.H, P.S.T, J.-Y.Y., J.-P.L and E.C.

C.D.G. generated most of the BioID constructs and cell lines and performed BioID experiments and immunofluorescence studies.

J.D.R.K., C.D.G., G.G.H. and A.-C. G. performed data analysis.

J.D.R.K. created the humancellmap.org website.

K.A. contributed the cell model and illustrations.

A.R., H.A. and E.S performed mitochondrial morphology experiments and analyzed results.

B.R. generated constructs and cell lines for BioID and testing predictions.

G.G.H., K.A., J.-Y.Y., P.S.-T., H.Z., L.Y.Z., E.P., J.-P.L., E.C., S.C., L.P., B.R. and A.P. contributed constructs and cell lines.

A.-C.G. supervised the project.

Figure/Table legends

Fig. 1: Procedure for dataset generation and localization rationale. (**a**) Compartments specifically targeted for profiling by BioID. **Bold** numbers on the schematic correspond to the indices on the legend. The *italicized* numbers in brackets next to the compartment name indicate the number of baits used to profile that compartment after quality control (QC). (**b**) QC and dataset-scoring pipeline. Bait performance was assessed by immunofluorescence (IF) of FLAG-tagged baits alongside endogenous compartment markers (to exclude baits with gross mislocalization) and GO term enrichment of significant proximal interactors following SAINTexpress[11] scoring against control replicates. All IF images are available with their corresponding bait reports at the humancellmap.org. BioID replicate reproducibility plots for each bait can be found in **Supplementary Figure 1**. **(c)** QC filtered out 42 baits from the original 234, and SAINT analysis yielded 4,424 high confidence proximity interactors from the total 7,192 proteins identified by mass spectrometry. (**d**) Rationale for prey-association based localization. In BioID, bait proteins label their proximal environment in a distance-dependent manner. The relative labelling of preys across baits is therefore dependent on the proximity of those prey proteins to each other *in situ.* In other words, each prey produces a "signature" across baits that it will share with proteins localizing to the same locales. This correlation can be used to assign localizations to preys based on the previously known localization of preys with a similar signature. (**e**) Correlation prey-prey heat map. Correlation between preys across baits was calculated from spectral counts using the Pearson coefficient, and preys were clustered by Euclidean distance and complete linkage. The heat map was manually annotated by performing GO enrichment on cluster components. See **Supplementary Table 5** for annotations and GO enrichments of the highlighted clusters.

Fig. 2: Localization of proteins using prey-prey information. (**a**) Pipelines for localizing prey proteins using Spatial Analysis of Functional Enrichment (SAFE[15]) and Non-negative Matrix Factorization (NMF[16]). In SAFE, preys with a spectral count correlation across baits ≥ 0.65 are considered "interactors" and these pairs are used to generate a network that is annotated for GO:CC terms. In NMF, the bait-prey spectral count matrix is reduced to a compartment-prey matrix and compartments are then defined using GO:CC for the compartments most abundant preys. A 2D network is generated in parallel from the compartment-prey matrix using t-SNE[34]. (**b**) NMF-based map of the cell generated with t-SNE. Each prey is coloured to indicate its primary localization. An interactive version of the map can be viewed at "humancellmap.org/explore/maps". (**c**) Performance of the NMF- and SAFE-based procedures against the immunofluorescence-based Human Protein Atlas (HPA[1], www.proteinatlas.org) and the fractionation studies of Christoforou[2] and Itzhak[3]. The vertical axis indicates the number of genes assigned to a previously known localization (GO:CC term). (**d**) Specificity of localization assignments. GO:CC terms were binned into specificity tiers (1: most specific; 5: least specific), and the number of genes assigned to each tier was quantified for our pipelines and the published studies under comparison.

Fig. 3: Novel protein localizations and identification of proteins at mitochondrial-ER contact sites. (**a**) Summary of experimental validations for predicted localizations of proteins by immunofluorescence (IF) microscopy and BioID. Confidence rankings were annotated as follows: "supported primary" indicates proteins that matched the NMF and SAFE prediction; "supported consistent" indicates proteins that matched the NMF and SAFE prediction, but did not have an endogenous compartment marker for the immunofluorescence microscopy; "supported BioID" indicates those proteins for which BioID data enriched for the NMF and SAFE prediction by GO analysis; "contradiction" indicates proteins that failed to localize to the predicted localization made by NMF and SAFE; "inconclusive" indicates proteins that had no clear subcellular compartment localization. (**b**) Heat map of genes that have a primary localization at the mitochondrial outer membrane or at the ER membrane/nuclear outer membrane-ER membrane and a secondary localization to the other compartment as computed by NMF. To be included on the heat map, genes required an NMF score of at least 0.15 in the compartments of interest, a score ratio of at least 0.4 between the primary and secondary localization, and a score ratio of at least 2 between the compartments of interest and all other compartments. Bolded genes indicate those selected for mitochondrial morphology assays in the following panels. A yellow dot on the right side of the plot indicates proteins involved in lipid and cholesterol homeostasis, while a pink dot indicates calcium signalling. (**c**) Mitochondrial morphology altered by transient expression of GFP-tagged CHMP7 and C18orf32 proteins is monitored by confocal immunofluorescence (IF) microscopy in HeLa cells. Cells were fixed and then probed with antibodies to GFP and COXIV (see Methods for details). The white box indicates the zoomed area displayed in the rightmost panels. Scale bars, 10 µm. **(d)** Quantification of HeLa cells with fragmented mitochondrial morphology upon overexpression of GFP-tagged proteins. Negative controls are coloured in orange including untransfected, GFP alone and CCDC47 (ER protein); positive controls MARCH5 and DNM1L are coloured in green; test candidates are coloured in blue. Experiments were done in biological triplicate with an average of ~150 cells counted per sample, statistical confidence of mitochondrial fragmentation was calculated using the Student's t-test and error bars represent standard deviation (see Methods). **(e)** IF of mitochondrial morphology in human primary fibroblasts that have GFP, CHMP7 and C18orf32 targeted siRNA knockdown. The white box indicates the zoomed area displayed in the bottom panels. The mitochondrial marker is an antibody against cytochrome C (see Methods). Scale bars, 10 µm. **(f)** Quantification of primary fibroblast mitochondria morphology upon siRNA mediated knockdown. Fraction of cells with hyperfused, fragmented and intermediate mitochondrial morphology are displayed in blue, red and purple. Experiments were done in biological triplicate with 100-150 cells counted per sample and error bars represent standard deviation (see Methods).

Supplementary Fig. 1: BioID bait replicate reproducibility. The spectral count for significant preys is plotted between replicates for each bait in the core dataset of 192 baits. The bait name and R^2 are listed above each plot.

Supplementary Fig. 2: Bait similarity and localization. The Jaccard index was calculated between each pair of baits in the core dataset using the significant preys for those baits. Baits

were clustered using the Euclidean distance and complete linkage method, and clusters optimized using the CBA package in R. The gradient next to the bait labels indicates whether a bait shares an expected localization with both adjacent baits (red), one adjacent bait (light red) or neither adjacent bait (white). Major clusters were manually annotated based on the expected localization of the components.

Supplementary Fig. 3: SAFE-based map of the cell and motif enrichment. **a**) SAFE-based map of the cell generated from preys with a Pearson correlation score of 0.65 or higher and plotted using Cytoscape with a spring-embedded layout. Each prey is coloured to indicate its primary localization (domain in SAFE terminology) as indicated in the legend. An interactive version of the map can be viewed at "humancellmap.org/explore/maps" and toggling from NMF to SAFE on the bottom menu. **b**) Pfam regions/motifs enriched in the indicated NMF ranks. The heat map value represents the $\log_2$ fold-change between the genes localized to the rank and all preys in the dataset. **c**) Pfam regions/motifs enriched in the indicated SAFE domains. The heat map value represents the $\log_2$ fold-change between the genes localized to the rank and all preys in the dataset.

Supplementary Fig. 4: Localization benchmarking. **a**) Percentage of genes localized to a previously known compartment for each specificity tier using our NMF and SAFE pipelines, compared with the Human Protein Atlas (HPA[1], www.proteinatlas.org) and the fractionation studies of Christoforou[2] and Itzhak[3]. Specificity tiers were defined by binning GO:CC terms based on their Information Content (IC) as defined in **Methods**. Tier 1 terms are the most specific and Tier 5 the least specific. **b-c**) Percentage of preys localized to a previously known compartment relative to the number of baits they were detected with for NMF and SAFE respecitively. **c-d**) Percentage of preys localized to a previously known compartment relative to the average number of spectral counts they were seen with for NMF and SAFE. Preys were binned by spectral count. The left tick mark for each data point indicates the lower bound for the bin (inclusive) and the right tick mark the upper bound (exclusive).

Supplementary Fig. 5: Localization prediction validation strategy and examples. Confidence rankings are defined as follows: "supported primary" indicates proteins that matched the NMF and SAFE prediction; "supported consistent" indicates proteins that matched the NMF and SAFE prediction but did not have an endogenous compartment marker for the immunofluorescence microscopy; "contradiction" indicates proteins that failed to localize to the predicted localization made by NMF and SAFE; "inconclusive" indicates proteins that had no clear subcellular compartment localization. Representative IF images are shown, markers used are on the respective panel. NMF scores across the defined ranks/categories/compartments are displayed as seen on humancellmap.org with the highest NMF category corresponding to the localization prediction.

Supplementary Fig. 6: Dotplot view of BioID data for mito-ER contact site candidates highlighting recovery of mitochondrial fission machinery, mito-ER tethers and outer mitochondrial membrane proteins. Asterisks on the heat map indicate spectral counts for prey

genes corresponding to the bait that were removed by SAINT as peptides from the bait itself confound accurately evaluating the abundance of itself as an interactor.

Supplementary Fig. 7: Analysis module at humancellmap.org. **a**) Screenshot of the analysis report for the bait PIK3R1. Red circles indicate: 1) Baits from the cell map are sorted from most similar to least similar as calculated by the Jaccard distance. 2) The ten most similar baits to the query in the cell map. 3) The average spectral count for each prey averaged across all baits in the cell map database. 4) Expected localizations of the ten most similar baits. 5) Overlap/similarity metrics between the query bait and the top ten most similar baits in the cell map. The distance is the Jaccard distance, with a score of 0 for complete prey overlap and 1 for no overlap. The intersection refers to the number of shared preys, and the union refers to the combined number of preys between the query and the indicated bait. 6) The most specific preys for the query. The specificity score is calculated as the fold enrichment of a prey in the query relative to the average across the cell map baits used for the comparison. 7) The specificity score calculated against the top ten most similar baits to the query. 8) The specificity score calculated against all baits in the cell map. 9) Links to open the heat map or specificity plots at the interactive viewer at ProHits-viz. 10) Links for data downloads. **b**) Specificity plot for PIK3R1. **c**) Specificity plot for RNGTT.

Supplementary Fig. 8: Exploratory analysis of FAM171A1 and MTFR2. BioID was performed on these two baits and the SAINT-processed data was analyzed using the analysis module at the humancellmap.org. a) Specificity plot of FAM171A1 showing the high abundance and/or specificity of cytoskeletal proteins. b) Specificity plot of MTFR2 showing the high specificity of proteins involved in mitochondrial dynamics.

Supplementary Table 1: Bait descriptions and bait quality control summary.

Supplementary Table 2: Mass spectrometry samples and results for the core BioID dataset.

Supplementary Table 3: Enriched terms for significant proximal proteins.

Supplementary Table 4: Recovered protein interactors.

Supplementary Table 5: Enriched GO cellular component terms in clusters.

Supplementary Table 6: SAFE localization predictions.

Supplementary Table 7: NMF localization predictions.

Supplementary Table 8: NMF and SAFE compartment definitions.

Supplementary Table 9: NMF and SAFE domain and motif enrichment.

Supplementary Table 10: Benchmarking.

Supplementary Table 11: Predicted localization validation descriptions.

Supplementary Table 12: Mass spectrometry samples and results for the ER-Mito BioID dataset.

Supplementary Table 13: Mass spectrometry samples and results for the prediction BioID dataset.

Supplementary Table 14: List of reagents

Methods

Selection of compartment markers

We aimed at selecting at least three independent baits (we refer to them as "compartment markers") for all major membrane-bound and membraneless organelles in HEK293 cells, as well as for all cytoskeletal elements. For complex organelles, such as the nucleus and the mitochondrion, distinct markers were selected to profile their major subcompartments (e.g. matrix, inner membrane and outer membrane for the mitochondria). These markers were selected by manual literature curation (e.g. they have previously been used as fluorescent recombinant proteins or sequence tags to mark selected structures), from proteins reported as high quality markers in the Human Protein Atlas[1], commercially used as compartment markers for immunofluorescence (e.g. Cell Signaling Technology), or following advice from cell biology experts. The list of the constructs used can be found in **Supplementary Table 1**.

The selection of the BirA*-FLAG location (N- or C-terminus) for each marker was as follows: if the selected marker had previously been used successfully for fluorescent-protein tagging and microscopy, we kept the location of the tag identical. For proteins where such information was not available (i.e. they were used as endogenous markers), the structural organization of the protein was taken into consideration (for example, if a critical domain or motif such as prenylation, was present at one of the termini, the other terminus was used for tagging). Lastly, for transmembrane-containing proteins, the membrane topology was analyzed from both the literature and using the Protter tool[35], and the tag integrated on the side of the membrane where compartment labelling was desired. In 6 cases, both N and C-terminal fusions of the same protein were generated.

Selected markers were subcloned as in-frame fusions by Gateway cloning in the pcDNA5-FLAG-BirA* backbone (with fusion of the marker at either the N- or C-terminus). When no appropriate entry Gateway construct was available, entry clones were generated by PCR amplification from cDNA constructs (Mammalian Gene Collection; MGC). "Open" Gateway constructs destined to N-terminal fusions were first "closed" by PCR amplification and recloned as closed entries to prevent cloning scars[36]. Sequence tags[37] were PCR amplified from relevant cDNA or Gateway ORF clones of the full-length proteins, or from oligo annealing, and inserted into the pcDNA5-FLAG-BirA* backbone. All constructs generated by PCR amplification were validated by Sanger sequencing.

Cell line generation for BioID

For BioID, the parental cell line, HEK293 Flp-In T-REx 293 (Invitrogen), was grown at 37°C in DMEM high glucose supplemented with 5% Fetal Bovine Serum, 5% Cosmic calf serum and 100 U/ml Pen/Strep (growth media). The parental cell lines are routinely monitored for mycoplasma contamination and have been authenticated by STR analysis with The Center for Applied Genomics Genetic Analysis Facility.

For the generation of stable cell lines, HEK293 Flp-In T-REx cells were transfected using the jetPRIME transfection reagent (Polyplus Cat# CA89129-924). Cells were seeded at 250,000 cells/well in a 6-well plate in 2 ml growth media (day 0). The following day (day 1), cells were transfected with 100 ng of pcDNA5-FLAG-BirA* bait construct and 1 µg of pOG44 in 200 µl of jetPRIME buffer mixed with 3 µl of jetPrime reagent (of this mix, 200 µl was added to the cells as per the manufacturer's protocol). On day 2, transfected cells were passaged to 100 mm plates. On day 3, hygromycin was added to the growth media (final concentration of 200 µg/ml). This selection media was changed every 2–3 days until clear visible colonies were present, at which point the colonies were pooled. Cells were then scaled up to 150 mm plates. Cells were grown to 70% confluence before induction of protein expression using 1 µg/ml tetracycline, and the media supplemented with 50 µM biotin for protein labelling. Cells were harvested 24 h later as follows: cell media was decanted, cells were washed once with 5 ml PBS per 150 mm plate and then harvested by scraping in 1 ml PBS. Cells from one or two 150 mm plates were pelleted at 233 RCF for 5 min, the supernatant aspirated, and pellets frozen on dry ice. Cell pellets were stored at -80°C until further processing.

BioID

Two different BioID protocols were implemented and are described below. The protocol used for each bait can be found in **Supplementary Table 2**.

Protocol 1 (high stringency washes; highSDS): Cell pellets from one 150 mm plate were lysed in a modified RIPA buffer containing $MgCl_2$ (modRIPA + $MgCl_2$: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 1% Triton X-100, 1 mM EGTA, 0.1% SDS, Sigma-Aldrich protease inhibitors P8340 1:500 (v:v), and 0.5% Sodium deoxycholate) at 1:10 (pellet weight in g : lysis buffer volume in ml). After lysis buffer addition, 1 µl of benzonase (EMD, CA80601-766, 250 U) was added to each sample, and cell pellets were incubated on a nutator at 4^{o}C for 20 min. Lysates were sonicated (3 x 10-second bursts with 2 seconds rest) on ice at 65% amplitude using a Qsonica with a CL-18 probe. Lysates were centrifuged for 30 min at 20,817 RCF at 4°C. After centrifugation, lysate supernatants were added to pre-washed streptavidin-sepharose beads (GE Cat# 17-5113-01; 30 µl bed volume of pre-washed beads per sample), and biotinylated proteins were affinity-purified at 4^{o}C on a nutator for 3 h. After affinity purification, streptavidin sepharose beads were pelleted (400 RCF, 1 min), and the supernatant removed. Streptavidin beads were then transferred to a new microfuge tube in 1 ml of 2% SDS Wash Buffer (2% SDS, 25 mM Tris-HCl pH 7.5). All subsequent washes used 1 ml of the indicated buffer with a centrifugation force of 400 RCF for 1 min. Beads were washed twice with modRIPA +$MgCl_2$ (without protease inhibitors or sodium deoxycholate), and three times with 50 mM ammonium bicarbonate buffer (pH 8). All buffer was removed from the final wash, and 1 µg of mass spectrometry grade trypsin/Lys-C mix (Promega CAT# V5071) in 60 µl of 50 mM ammonium bicarbonate was added to each sample. Proteins were digested on beads overnight at 37 ^{o}C on a rotator. The following day, an additional 0.5 µg trypsin/Lys-C mix was added to samples that were further digested at 37 ^{o}C on a rotator for 2 h. Each sample was spun down at 400 RCF for 1 min to pellet beads, and the supernatant was transferred to a new 1.5 ml microcentrifuge tube. Beads were then washed with 30 µl of HPLC-grade water (Caledon Laboratory Chemicals

CAT# 7732-18-5), centrifuged at 400 RCF for 1 min to pellet beads, and the supernatant pooled with digested peptides collected previously (this step was repeated once). Samples were centrifuged at 16,000 RCF for 5 min and 100 µl was transferred to a new microfuge tube. Samples were acidified by adding 4 µl of 50% formic acid (final concentration of 2% formic acid) and dried in a centrifugal evaporator. Dried peptides were stored at -80 ^{o}C.

Protocol 2 (lower stringency washes; lowSDS): This follows the same steps as Protocol 1, except for the details listed below. Cell pellets from two 150 mm plates were lysed in modified RIPA buffer containing EDTA (modRIPA + EDTA: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, 1 mM EGTA, 0.1% SDS, Sigma-Aldrich protease inhibitors P8340 1:500 (v:v), and 0.5% Sodium deoxycholate) at 1:10 (pellet weight in g : lysis buffer volume in ml). After affinity purification, streptavidin beads were transferred to a new microfuge tube in 1 ml of modRIPA +EDTA (without protease inhibitors or sodium deoxycholate). All subsequent washes used 1 ml of a buffer with a centrifugation force of 400 RCF for 1 min. The beads were washed once more with modRIPA +EDTA (without protease inhibitors or sodium deoxycholate), twice with an NP-40 wash buffer (10% glycerol, 50 mM HEPES-KOH pH 8.0, 100 mM KCl, 2 mM EDTA, 0.1% NP-40) and three times with 50 mM ammonium bicarbonate (pH 8) buffer. All of the buffer was removed from the final wash, and 1 µg of mass spectrometry grade trypsin (Sigma-Aldrich T6567) in 200 µl of 50 mM ammonium bicarbonate was added to each sample. Samples were digested on beads overnight at 37 ^{o}C on a rotator. After the addition of an additional 0.5 µg of trypsin and 2 h incubation, the digested peptides were transferred to a new 1.5 ml microcentrifuge tube. Beads were then washed with 150 µl of HPLC-grade water (Caledon Laboratory Chemicals CAT# 7732-18-5), centrifuged at 400 RCF for 1 min to pellet beads, and the supernatant pooled with digested peptides collected previously. The water wash and collection of the supernatant were repeated once more. Digested peptides were centrifuged at 16,000 RCF for 5 min and 470 µl collected into a new microfuge tube. Samples were dried in a centrifugal evaporator, and dried peptides were stored at -80 ^{o}C.

Mass spectrometry analysis

Dried peptides were resuspended in 20 µl of 5% formic acid and centrifuged at 16,000 RCF for 1 min. 5 µl were injected via autosampler in a 12 cm analytical fused silica capillary column (0.75 µm internal diameter, 350 µm outer diameter). The column was made in house using a laser puller (Sutter Instrument Co., model P-2000; heat = 280, FIL = 0, VEL = 30, DEL = 200), packed with C18 reversed-phase material (Reprosil-Pur 120 C18-AQ, 3 µm; Dr. Maische), and connected in-line to a NanoLC-Ultra 2D plus HPLC system (Eksigent, Dublin, USA). The system was equipped with a nanoelectrospray ion source (Proxeon Biosystems, Thermo Fisher Scientific) delivering the sample to an Orbitrap Elite Hybrid Ion Trap-Orbitrap mass spectrometer (Thermo Fisher Scientific). The HPLC program delivered the following percentages of buffer B (0.1% formic acid in acetonitrile) to buffer A (0.1% formic acid in water) at the described flow rates over a 130 min gradient. The start of the HPLC program loaded the sample onto the column with a flow rate of 400 µl/min with 5% buffer B for 14 min followed by a drop in flow rates from 400 µl/min to 200 µl/min using a linear gradient from 5% to 2% buffer B for 1 min. Next, a linear gradient from 2% to 35% buffer B began eluting the sample into the mass

spectrometer at 200 µl/min for 90 min, followed by another linear gradient from 35 to 80% buffer B over 5 min, and maintaining 80% buffer B for 5 min to elute the remaining analytes. The final stages of the HPLC program had a flow rate of 200 µl/min using a linear gradient from 80% to 2% buffer B over 3 min, and a quick re-equilibration of the column for 12 min at 200 µl/min with 2% buffer B.

The Orbitrap Elite Hybrid Ion Trap-Orbitrap mass spectrometer was operated with Xcalibur 2.0 software in data-dependent acquisition mode with the following parameters: one centroid MS (mass range 400 to 2000) followed by MS2 on the top 10 most abundant ions with a dynamic exclusion of 20 s (general parameters: activation type = CID, isolation width = 2 m/z, normalized collision energy = 35, activation Q = 0.25, activation time = 10 ms. The minimum signal required was 1000, the repeat count = 1, repeat duration = 30 s, exclusion size list = 500, exclusion duration = 15 s, exclusion mass width (Da) = low 0.6, high 1.2). To decrease carry over between samples on the autosampler, the analytical column was washed three times using a “sawtooth” gradient of 35% acetonitrile with 0.1% formic acid to 80% acetonitrile with 0.1% formic acid, holding each gradient for 5 min, three times per gradient. Following washes, quality control on the column and machine performance were assessed by loading 30 fmol BSA tryptic peptide standard (Michrom Bioresources Inc. Fremont, CA) with 60 fmol α-Casein tryptic digest. The HPLC program for the quality control ran a shortened 60 min gradient with the following percentages of buffer B and flow rates: 9 min at 400 µl/min with 5% buffer B, 1 min going from 400 µl/min to 200 µl/min using a linear gradient from 5 to 2% buffer B, 30 min at 200 µl/min using a linear gradient from 2 to 35% buffer B, 5 min at 200 µl/min using a linear gradient from 35 to 80% buffer B, 5 min at 200 µl/min with 80% buffer B, 5 min at 200 µl/min using a linear gradient from 80 to 2% buffer B and 5 min at 200 µl/min with 2% buffer B.

Mass spectrometry data analysis

Samples analyzed on the Orbitrap Elite Hybrid Ion Trap-Orbitrap mass spectrometer were converted to mzML using ProteoWizard (3.0.4468)[38] and analyzed using the iProphet[39] pipeline implemented within ProHits[40] as follows. The database consisted of the HEK293 sequences in the RefSeq protein database (version 57) supplemented with “common contaminants” from the Max Planck Institute http://141.61.102.106:8080/share.cgi?ssid=0f2gfuB and the Global Proteome Machine (GPM; http://www.thegpm.org/crap/index.html) with the addition of sequences from common fusion proteins and epitope tags. The search database consisted of forward and reverse sequences (labeled “gi|9999” or “DECOY”); in total, 72,226 entries (including decoys) were searched. Spectra were analyzed separately using Mascot (2.3.02; Matrix Science) and Comet (2012.01 rev.3)[41] for trypsin specificity with up to two missed cleavages; deamidation (NQ) or oxidation (M) as variable modifications; single-, double-, and triple-charged ions allowed, mass tolerance of the parent ion to 12 ppm; and the fragment bin tolerance at 0.6 amu. The resulting Comet and Mascot search results were individually processed by PeptideProphet[42], and peptides were assembled into proteins using parsimony rules first described in ProteinProphet[43] into a final iProphet protein output using the Trans-Proteomic Pipeline (TPP; Linux version, v0.0 Development trunk rev 0, Build 201303061711). TPP options were 1) general options: -p0.05 -x20 -PPM -d"DECOY", 2) iProphet options: –

ipPRIME and 3) PeptideProphet options: –pP. All proteins with a minimal iProphet protein probability of 0.05 were parsed to the relational module of ProHits. Note that for analysis with SAINT (see below), only proteins with iProphet protein probability ≥ 0.95 were considered, corresponding to an estimated protein level false-discovery rate (FDR) of approximately 0.5%.

SAINT file processing

For each prey protein identified in an affinity purification experiment, SAINT calculates the probability of it being a true interaction by using spectral counting (semi-supervised clustering, using a number of negative control runs). SAINTexpress[11] analysis was performed using version exp3.6.1 with two biological replicates per bait. Two separate SAINT analyses were performed for the two BioID protocols. For the baits used with BioID Protocol 1, 322 bait protein samples (162 baits) were analyzed alongside 70 negative control runs, consisting of purifications from untransfected cells or cells expressing BirA*-FLAG, or BirA*-FLAG-GFP. For BioID Protocol 2, 52 bait protein samples (26 baits) were analyzed alongside 16 negative control runs, consisting of purifications from untransfected cells or cells expressing BirA*-FLAG, or BirA*-FLAG-GFP. No compression of the controls was performed and default parameters for SAINTexpress were used. A 1% Bayesian FDR cutoff was used to select confident proximity interactors. The two SAINT files for the core dataset were combined into a single file for downstream analysis, and non-human contaminants were removed from the final report, as were baits with less than 5 preys. SAINTexpress was also used in a separate analysis for the proximity proteomes of the "Prediction" baits; Protocol 1 controls described above were used for this analysis, using the same parameters as above.

Data deposition

Datasets consisting of raw files and associated peak lists and results files have been deposited in ProteomeXchange through partner MassIVE (http://proteomics.ucsd.edu/ProteoSAFe/datasets.jsp) as complete submissions. Additional files include the sample description, the peptide/protein evidence and the complete SAINTexpress output for each dataset, as well as a "README" file that describes the dataset composition and the experimental procedures associated with each submission. The different datasets generated here were submitted as independent entries.

Dataset 1 (see **Supplementary Table 2**):
Go_BioID_humancellmap_HEK293_lowSDS_core_dataset_2019
MassIVE **ID** MSV000084359 and PXD015530

Dataset 2 (see **Supplementary Table 2**): Go_BioID_humancellmap_HEK293_ highSDS_core_dataset _2019
MassIVE ID MSV000084360 and PXD015531

Dataset 3 (see **Supplementary Table 13**): Go_BioID_humancellmap_HEK293_prediction_2019
MassIVE ID MSV000084369 and PXD015554

Dataset 4 (see **Supplementary Table 12**): Go_BioID_humancellmap_HEK293_ER-mito_candidates_2019
MassIVE ID MSV000084357 and PXD015528

Negative control samples were deposited in the Contaminant Repository for Affinity Purification[44] (CRAPome.org) and assigned samples numbers CC1100 to CC1185 (see **Supplementary Table 2**); this will be part of the next release of the database.

Immunofluorescence (IF) microscopy for bait quality control

For quality control of stable cell lines expressing BirA*-FLAG-tagged baits, HEK293 Flp-In T-REx cells were seeded directly on 12 mm poly-L-lysine coated coverslips (Corning, Product # 354085). The next day, cells were treated with 1 µg/ml tetracycline and media was supplemented with 50 µM biotin for 24 h. Media was aspirated, and cells were washed with PBS supplemented with 200 µM $CaCl_2$, 100 µM $MgCl_2$, prior to fixation with 4% formaldehyde in PBS for 10 min, and washing three times in TBS-T (Tris-buffered saline and 0.1% v/v). The cells were then treated for 10 min in permeabilization buffer (0.1% Triton X-100 in TBS-T), followed by 3 washes in TBS-T and incubation at room temperature in blocking buffer (5% BSA w/v in TBS-T). Samples were incubated with primary antibodies in blocking buffer in a humidified chamber for 1 h: anti-FLAG M2 (1:2000 dilution, Sigma Aldrich, F3165) and an endogenous compartment marker antibody from rabbit (**Supplementary Table 14** for list of antibodies used), or anti-FLAG from rabbit (1:500 dilution, Sigma Aldrich, F7425) and an endogenous compartment marker antibody from mouse. All samples were then washed 3 times in blocking buffer before incubation with blocking buffer containing secondary antibodies in a dark, humidified chamber for 1 h with one of the combination of antibodies and dyes listed here: (1) anti-rabbit coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11034), anti-mouse coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21422), Streptavidin-coupled to Alexa Fluor 647 (1:2500, Invitrogen, S32357); (2) anti-mouse coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21428), Streptavidin-coupled to Alexa Fluor647 (1:2500, Invitrogen, S32357); (3) anti-mouse coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21422), DAPI (1:2000), Streptavidin-coupled to Alexa Fluor 647 (1:2500, Invitrogen, S32357); or (4) anti-mouse coupled to Alexa Fluor555 (1:1000, Invitrogen, A21422), Phalloidin-coupled to Alexa Fluor 488 (1:1000, Invitrogen, A12379), Streptavidin-coupled to Alexa Fluor 647 (1:2500, Invitrogen, S32357). Following incubation, samples were washed 3 times with TBS-T. Each coverslip was mounted on a glass slide using ~4 µl of ProLong Gold Antifade Mountant (Thermo Fisher Scientific, CAT #P36930). Samples were then cured, lying flat, overnight in the dark, followed by storage in the dark at 4°C. Images were acquired on a Nikon C1Si Confocal Microscope using a 60x objective lens magnification and 3x field zoom.

In some instances, ice-cold methanol (MeOH) was used as a fixative to better visualize microtubules and facilitate the use of specific antibodies only amenable to MeOH fixation conditions. Ice-cold MeOH addition and incubation at -20°C for 30 min was used to fix and permeabilize cells after the first initial wash. After the cells were washed 3 times with TBS-T,

the protocol continued as described above with the addition of blocking buffer. When Wheat Germ Agglutinin (WGA)-coupled to Alexa Fluor488 (1:250, Invitrogen, W11261) was used as a counterstain, all steps were performed with samples chilled on ice, employing ice-cold buffers and in the dark. After the initial wash, cells were incubated with a solution containing WGA-coupled to Alexa Fluor488 in PBS containing 200 µM $CaCl_2$, 100 µM $MgCl_2$ for 10 min. After the samples were washed twice with this solution, the protocol was as described for formaldehyde fixation.

GO enrichment analysis

GO enrichments were performed using g:Profiler[45]. Enrichments were performed considering gene lists as unordered, allowing only genes with annotations, using all significant proximal interactors as background, a max p-value of 0.01 and the g:SCS multiple test correction method. For bait QC, it should be noted that DHFR2 did not match its expected compartment enrichment but was allowed into our analysis pipeline as its large list of proximal interactors was deemed to be informative for localization purposes. NPM1 and KDM1A had an expected GO:CC enrichment profile when the max p-value was relaxed from 0.01 to 0.05.

Databases used for analysis

The BioGRID[12] human database v3.5.169 was downloaded on 13/2/2019. Human gene annotations were downloaded from Gene Ontology (GO) on 15/2/2019 (GO version date 1/2/2019). The GO hierarchy (release date 13/2/2019) was downloaded from GO[46,47] on 15/2/2019. The UniProt database[48] was downloaded on 21/2/2019. The IntAct[13] human database was downloaded on 13/2/2019. Human protein domain annotations and motifs were retrieved from Pfam[49] (version 32) on 21/2/2019.

Jaccard index

The Jaccard index is the overlap between two sets (*A, B*) calculated as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard distance is defined as 1 – *J(A,B)*.

Top 25 proximity interactors

The top 25 interacting preys for each bait were determined from their length-normalized spectral counts. For bait *i* and prey *j* this value was calculated by first subtracting the prey's average spectral count found in control samples from its abundance with bait *i*, then multiplying by the median prey length of all preys across bait *i* and dividing by the length of prey *j*.

$$NormSpec_{i,j} = \frac{\widetilde{length}}{length_j} \cdot (AvgSpec_{i,j} - AvgSpec_{control,j})$$

Prey-prey correlation

The SAINTexpress file was processed using the correlation tool at ProHits-viz[50] with an FDR score filter of 0.01 and an abundance cut-off of 0. If a prey passed the FDR cut-off for one bait, its abundance across all other baits was used in the analysis. Control average values were subtracted from replicate spectral counts and these control-subtracted values used for correlation. After Pearson correlation scores were calculated between preys, complete-linkage clustering was performed using the Euclidean distance between preys, and cluster order was optimized using the CBA package in R.

SAFE

A network was built from prey-prey correlation data using ProHits-viz as described above. Networks were built in Cytoscape[51], version 3.6.1 using a spring embedded layout. All preys passing an FDR cutoff of 0.01 were included in this analysis. After performing correlation, we considered preys to be interaction pairs if they passed a required correlation cut-off. This cut-off was set to 0.5 to 0.9 in increments of 0.05 for testing with SAFE[52] as we could not know a priori what an ideal cut off would be although manual assessment suggested something in this range would be suitable. SAFE requires annotations for network nodes and for each node we created a list of all known GO cellular compartment terms supplemented with their parent terms. When running SAFE, we also tested several percentile neighbourhood radii for each network, ranging from 3 to 10 in increments of 0.5. With these parameters, we sought to maximize the number of preys being assigned to a domain with a known GO term for that prey. A prey was considered assigned to a correct domain if one of its GO terms (or a parent of those terms) was found within the terms assigned to its predominant domain. After manually inspecting the SAFE results, we felt the optimal annotation was generated from the network built with a correlation cut-off of 0.65 with a neighbourhood radius of 4.5. This resulted in a network with 24 domains (one of which is “unknown”), where 60.2% (2351/3903) of genes were assigned to a domain with a known GO term. The complete definition of each domain was determined by the GO cellular compartment terms resulting from an enrichment of all preys with a primary localization to the domain in question. We also selected a representative term(s) for each domain as its compartment ID for localization and assessment purposes.

NMF

Non-negative matrix factorization (NMF) is an approach to create a compressed and simplified version of an $n \times m$ dimension matrix **V**, such that **V** ≈ **WH**, where **W** has dimensions $n \times r$ and **V** has dimensions $r \times m$, and both matrices consist entirely of non-negative entries[53]. Given an interaction matrix of n preys and m baits, where $V_{i,j}$ is the spectral count of prey i with bait j, the minimal rank r of the factorization is sought that sufficiently summarizes this input matrix. In

our case, we seek r << m. The matrix **W** can then be thought of as a compressed form of our input matrix, whereby instead of displaying a prey's profile across all baits, it shows how preys profile across ranks. A simple way to think of a rank in the context of our dataset is that it may represent a collection of baits that convey redundant information. In contrast to the input matrix that may show several data points indicating a prey is detected highly with each nuclear bait, for example, we might expect a single entry in the matrix indicating it was detected highly in the nucleus. Preys that behave similarly across baits would be expected to have similar profiles across ranks. Preys that only behave similarly across a subset of baits would still be expected to show a similar profile across a single or subset or ranks, while being free to show a different profile across the remaining ranks. Our input matrix had dimensions 4424 × 192 for the 192 baits in the data set and 4424 preys passing an FDR cutoff of 1%. Prey spectral counts had their average value in controls subtracted and were then rescaled from 0–1 across baits as we wanted each prey to be considered of equal weight. NMF, as implemented optimizing the squared Frobenius norm initialized by Nonnegative Double Singular Value Decomposition (NNDSVD) with L1 regularization in the *-scikit-learn* Python package[54], version 0.18.1, was then performed on this matrix for *r* = 10, 11...30. For each NMF run, GO cellular compartment terms were assigned to the resulting ranks by taking the top preys for each rank in the **W** matrix (up to 100 maximum) and profiling with g:Profiler[55] using our complete prey list as background. A prey could contribute to the enrichment process in an NMF rank if it was most abundant in that rank or within 25% of its maximum within that rank, and if it had a value of at least 0.25. These values were set to try and ensure there was sufficient evidence that a prey truly belonged to a rank. To determine the optimal number of ranks to use for NMF, we sought to maximize the number of preys assigned a known localization and minimize the overlap in GO terms between ranks. A prey was considered assigned to a correct rank if one of its known GO terms (or a parent of those terms) was found within the terms assigned to its rank. To determine the overlap in GO terms between ranks, we calculated the Jaccard distance between GO terms for each pair of ranks (where 0 would indicate complete overlap and 1 no overlap). While several NMF ranks performed well, we selected 20 ranks after manual inspection. Analysis with 20 ranks resulted in 87.6% of preys assigned to a rank with a previously known GO term and 74.8% of preys assigned to a rank where one of the top 5 GO terms was previously known, with the worst rank overlap at a Jaccard distance of 0.31. After defining the optimal rank number, each prey was assigned to its best rank for visualization and assessment purposes, and a representative GO term or terms was/were chosen to identify the rank, and also for visualization and assessment purposes. Since at most only the top 100 preys in a rank were used for its definition, we used the remaining preys localized to the rank to assess the ability of this approach to correctly localize proteins. 48.0% of these preys were localized to a previously known compartment (based on GO:CC annotations, **Supplementary Table 7**), giving us confidence in the procedure.

A network was built from the pairwise prey Euclidean distance matrix derived from the NMF **W** matrix using t-Distributed Stochastic Neighbor Embedding (t-SNE)[34]. t-SNE was performed using the Matlab script available at http://lvdmaaten.github.io/tsne/. It was run with the number of initial dimensions equal to the number of NMF ranks (20) and a perplexity of 20 for a maximum of 1000 iterations.

Information content

The information content (IC) of each GO cellular component term was calculated as -log(*p*), where *p* is the probability a gene has an annotation, i.e. the number of genes with the annotation divided by the total number of genes in GO. Annotations occurring in 1% of genes or less (189 / 18858 total genes in GO) were placed in our highest specificity IC tier (bin 1). Bins 2-5 corresponded to annotations occurring in 2%, 10%, 25% or > 25% of genes.

Dataset comparison

The Human Protein Atlas (HPA) subcellular localization data was downloaded on 15/3/2019 from www.proteinatlas.org/about/download, and is based on the Human Protein Atlas[56] version 18.1 and Ensembl version 88.38. All HPA entries in the subcellular localization table have an associated gene name and all localization terms are based on GO. Fractionation-based localizations from Christoforou et al.[2] were retrieved from Supplementary dataset 1, tab 2, column AI ("Final Localization Assignment"). Their localization terms were mapped to the closest GO term. Although their dataset is for mouse genes, more localizations were known if we assumed their genes were human and compared against the human GO database, so this was used for our assessment. Fractionation-based localizations from Itzhak et al.[3] were retrieved from Supplementary file 1, tab 2, columns E and F ("Compartment" and "Subcompartment"). Localization terms were mapped to the closest GO term. All genes from these datasets with their assigned and corresponding GO IDs are listed in **Supplementary Table 10**. Localization tiers were defined using the information content of each GO term as defined in the information content section. When genes were assigned multiple localizations, the lowest information content term (i.e. least specificity) was used for binning that gene into a localization tier.

Enrichments

Enrichment scores (p-values) for domain and motif enrichment were calculated for each NMF rank and SAFE domain using Fisher's exact test. Of the 4424 genes in our NMF analysis, 4368 had domain information available and 4301 had motif information available in Pfam. Of the 3903 genes in our SAFE analysis, 3855 had domain information available and 3809 had motif information available in Pfam. All genes with available information were used as background for the enrichment tests. The FDR was controlled by using the Benjamini–Hochberg procedure for an FDR of 1%.

Validation of the localization predictions by immunofluorescence microscopy and BioID

Selected targets for validation were cloned in Gateway compatible pcDNA5-GFP and pcDNA5-FLAG-BirA* backbones (with tags at either N- or C-terminus as described for the selection of bait quality control above) and localizations validated by immunofluorescence microscopy and GO enrichment as described above (see **Supplementary Table 11** for the list of tested baits).

GFP-tagged constructs were transiently transfected into HeLa cells (ATCC, CCL-2) using the jetPRIME transfection reagent (Polyplus Cat# CA89129-924). Cells were seeded at 250,000 cells/well in a 6-well plate in 2 ml growth media. The following day, cells were transfected with 400 ng of pcDNA5-GFP-tagged construct and 40 µl of jetPRIME buffer mixed with 0.8 µl of jetPrime reagent. The next day formaldehyde fixation, as described above, was used with the following alterations. Samples were incubated with primary antibodies in blocking buffer in a humidified chamber for 1 h. The primary antibodies used were anti-GFP from mouse (1:500 dilution, Roche, CAT# 11814460001) and an endogenous compartment marker antibody from rabbit (refer to **Supplementary Table 14** for list of antibodies used), or anti-GFP from rabbit (1:2000 dilution, abcam, ab290) and an endogenous compartment marker antibody from mouse. Samples were then incubated with blocking buffer containing secondary antibodies in a dark, humidified chamber for 1 h with one of the combination of antibodies and dyes listed here: (1) DAPI (1:2000), anti-rabbit coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11034), anti-mouse coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21422); (2) DAPI (1:2000), anti-mouse coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21428); or (3) DAPI (1:2000), anti-rabbit coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11034), Phalloidin-coupled to Alexa Fluor647 (1:1000, Invitrogen, A22287). Images were acquired on a Nikon C1Si Confocal Microscope using a 60x objective lens magnification and 1x or 2x field zoom.

BioID was performed on selected targets as described above for cell line generation, BioID Protocol 1, mass spectrometry data analysis and SAINT file processing. For the baits used with BioID Protocol 1, 20 bait protein samples (10 baits) were analyzed alongside 74 negative control runs, consisting of purifications from untransfected cells or cells expressing BirA*-FLAG, or BirA*-FLAG-GFP. GO enrichments were performed using g:Profiler[45]. Enrichments were performed considering gene lists as unordered, allowing only genes with annotations, using a max p-value of 0.05 and the g:SCS multiple test correction method.

Confidence levels of co-localization immunofluorescence images with respect to predicted localizations were assessed and corroborated by three individuals. Confidence rankings were annotated as follows: “supported primary” indicates proteins that matched the primary NMF and SAFE prediction; “supported consistent” indicates proteins that matched the primary NMF and SAFE prediction but did not have an endogenous compartment marker for the immunofluorescence microscopy; “contradiction” indicates proteins that failed to localize to the predicted localizations made by NMF and SAFE; “inconclusive” indicates proteins that had no clear subcellular compartment localization.

Cell culture for mitochondrial fragmentation assays

Primary fibroblasts (Cell bank at Montreal Children’s Hospital) and HeLa cells were grown in high-glucose DMEM supplemented with 10% fetal bovine serum, at 37°C in an atmosphere of 5% CO_2. Stealth RNAi duplex constructs (Invitrogen) were used for transient knockdown of C18orf32 and CHMP7 in primary fibroblasts or HeLa cells. Stealth siRNA duplexes at 12 nM were

transiently transfected into cells using Lipofectamine RNAiMAX (Invitrogen 13778-150), according to the manufacturer's specifications. The transfection was repeated on day 3 and the cells were imaged for mitochondrial morphology analysis on day 6.

Mitochondrial fragmentation assays

For IF experiments for assaying mitochondrial fragmentation, candidate proteins were GFP-tagged and the constructs were transiently transfected into HeLa cells. HeLa cells were transfected using the jetPRIME transfection reagent (Polyplus Cat# CA89129-924). Cells were seeded at 250,000 cells/well in a 6-well plate in 2 ml growth media. The following day, cells were transfected with 400 ng of pcDNA5-GFP-tagged construct and 40 µl of jetPRIME buffer mixed with 0.8 µl of jetPrime reagent. The next day an IF protocol with FA fixation, described above, was used with the following alterations. Samples were incubated with primary antibodies in blocking buffer in a humidified chamber for 1 h. The primary antibodies used were anti-GFP from mouse (1:500 dilution, Roche, CAT# 11814460001) and anti-COXIV from rabbit (1:250, Cell Signaling Technology, Product# 4850). Samples were then incubated with blocking buffer containing secondary antibodies in a dark, humidified chamber for 1 h with anti-mouse coupled to Alexa Fluor 488 (1:1000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1000; Invitrogen, A21428) and Concanavalin A coupled to Alexa Fluor 647 (1:200, Invitrogen, C21421). Images were acquired on a Nikon C1Si Confocal Microscope using a 60x objective lens magnification. Experiments and image acquisition were separate independent experiments done in triplicate, with an average of n=149 cells per GFP-tagged protein. Mitochondrial fragmentation was quantified manually as deviations from WT mitochondrial staining compared to controls (HeLa cells untransfected or with GFP-alone). Statistical confidence of mitochondrial fragmentation was calculated using the Student's t-test.

Primary fibroblasts were fixed in warm 4% formaldehyde (FA) in PBS at room temperature for 20 min, then washed three times with PBS before cells were permeabilized in 0.1% Triton X-100 in PBS, followed by three washes in PBS. The cells were then blocked with 3% bovine serum albumin (BSA) in PBS, followed by incubation with primary antibodies (rat anti-KDEL and mouse anti-Cytochrome C, refer to **Supplementary Table 14**) in 3% BSA in PBS for 1 hr at room temperature. After three washes with 3% BSA in PBS, cells were incubated with the appropriate anti-species secondary antibodies coupled to Alexa fluorochromes (1:2000, Invitrogen, **Supplementary Table 14**) for 30 min at room temperature. After three washes in PBS, coverslips were mounted onto slides using fluorescence mounting medium (Agilent Dako). Stained cells were imaged using a 100x objective lenses (NA1.4) on an Olympus IX81 inverted microscope with appropriate lasers using an Andor/Yokogawa spinning disk system (CSU-X), with a sCMOS camera. Mitochondrial network morphology was manually classified, in a blinded manner, as fused, intermediate, or fragmented. For every knockdown condition and controls, 100 - 150 cells were analyzed, and experiments were done three times independently. Error bars represent mean ± standard deviation.

BioID, mass spectrometry analysis and SAINT file processing for mitochondria-ER contact sites

BioID was performed on selected targets as described above for cell line generation, BioID Protocol 1, mass spectrometry data analysis and SAINT file processing. For the baits used with BioID Protocol 1, 20 bait protein samples (10 baits) were analyzed alongside 74 negative control runs, consisting of purifications from untransfected cells or cells expressing BirA*-FLAG, or BirA*-FLAG-GFP. GO enrichments were performed using g:Profiler[45]. Enrichments were performed considering gene lists as unordered, allowing only genes with annotations, using a max p-value of 0.05 and the g:SCS multiple test correction method.

References

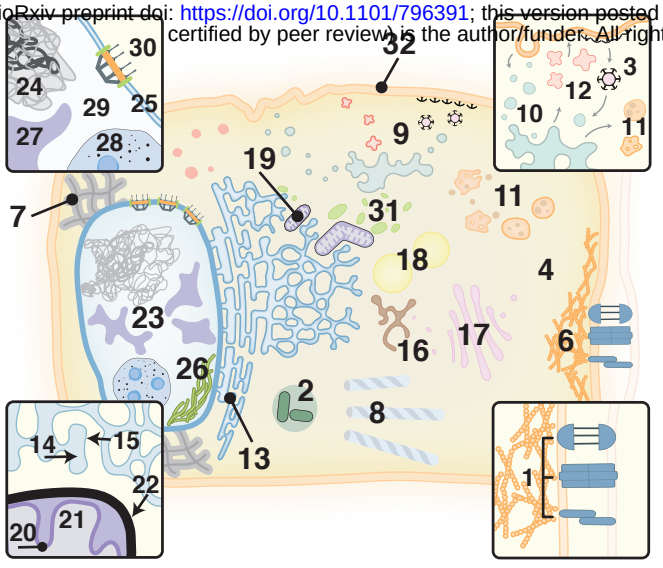
1. Thul, P.J. et al. A subcellular map of the human proteome. *Science* **356** (2017).
2. Christoforou, A. et al. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* **7**, 8992 (2016).
3. Itzhak, D.N., Tyanova, S., Cox, J. & Borner, G.H. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife* **5** (2016).
4. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187-199 (2006).
5. Kislinger, T. et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125**, 173-186 (2006).
6. Orre, L.M. et al. SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol Cell* **73**, 166-182 e167 (2019).
7. Roux, K.J., Kim, D.I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* **196**, 801-810 (2012).
8. Rhee, H.W. et al. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328-1331 (2013).
9. Gupta, G.D. et al. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. *Cell* **163**, 1484-1499 (2015).
10. Youn, J.Y. et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532 e511 (2018).
11. Teo, G. et al. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics* **100**, 37-43 (2014).
12. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-539 (2006).
13. Orchard, S. et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363 (2014).
14. Hung, V. et al. Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Mol Cell* **55**, 332-341 (2014).
15. Baryshnikova, A. Spatial Analysis of Functional Enrichment (SAFE) in Large Biological Networks. *Methods Mol Biol* **1819**, 249-268 (2018).
16. Frigyesi, A. & Hoglund, M. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer Inform* **6**, 275-292 (2008).
17. Chapple, C.E. et al. Extreme multifunctional proteins identified from a human protein interaction network. *Nat Commun* **6**, 7412 (2015).
18. Eisenberg-Bord, M., Shai, N., Schuldiner, M. & Bohnert, M. A Tether Is a Tether Is a Tether: Tethering at Membrane Contact Sites. *Dev Cell* **39**, 395-409 (2016).

19. Rowland, A.A. & Voeltz, G.K. Endoplasmic reticulum-mitochondria contacts: function of the junction. *Nat Rev Mol Cell Biol* **13**, 607-625 (2012).
20. Prudent, J. & McBride, H.M. The mitochondria-endoplasmic reticulum contact sites: a signalling platform for cell death. *Curr Opin Cell Biol* **47**, 52-63 (2017).
21. Lee, S. & Min, K.T. The Interface Between ER and Mitochondria: Molecular Compositions and Functions. *Mol Cells* **41**, 1000-1007 (2018).
22. Ackema, K.B. et al. Sar1, a Novel Regulator of ER-Mitochondrial Contact Sites. *PLoS One* **11**, e0154280 (2016).
23. Gomez-Suaga, P. et al. The ER-Mitochondria Tethering Complex VAPB-PTPIP51 Regulates Autophagy. *Curr Biol* **27**, 371-385 (2017).
24. Gomez-Suaga, P., Paillusson, S. & Miller, C.C.J. ER-mitochondria signaling regulates autophagy. *Autophagy* **13**, 1250-1251 (2017).
25. Kalia, R. et al. Structural basis of mitochondrial receptor binding and constriction by DRP1. *Nature* **558**, 401-405 (2018).
26. Korobova, F., Ramabhadran, V. & Higgs, H.N. An actin-dependent step in mitochondrial fission mediated by the ER-associated formin INF2. *Science* **339**, 464-467 (2013).
27. Xu, S. et al. Mitochondrial E3 ubiquitin ligase MARCH5 controls mitochondrial fission and cell sensitivity to stress-induced apoptosis through regulation of MiD49 protein. *Mol Biol Cell* **27**, 349-359 (2016).
28. Bersuker, K. et al. A Proximity Labeling Strategy Provides Insights into the Composition and Dynamics of Lipid Droplet Proteomes. *Dev Cell* **44**, 97-112 e117 (2018).
29. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379 (2017).
30. St-Denis, N. et al. Phenotypic and Interaction Profiling of the Human Phosphatases Identifies Diverse Mitotic Regulators. *Cell Rep* **17**, 2488-2501 (2016).
31. Li, X. et al. Defining the Protein-Protein Interaction Network of the Human Protein Tyrosine Phosphatase Family. *Mol Cell Proteomics* **15**, 3030-3044 (2016).
32. Rasila, T. et al. Astroprincin (FAM171A1, C10orf38): A Regulator of Human Cell Shape and Invasive Growth. *Am J Pathol* **189**, 177-189 (2019).
33. Monticone, M. et al. The nuclear genes Mtfr1 and Dufd1 regulate mitochondrial dynamic and cellular respiration. *J Cell Physiol* **225**, 767-776 (2010).
34. van der Maaten , L.J.P. & Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
35. Omasits, U., Ahrens, C.H., Muller, S. & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**, 884-886 (2014).
36. Banks, C.A., Boanca, G., Lee, Z.T., Florens, L. & Washburn, M.P. Proteins interacting with cloning scars: a source of false positive protein-protein interactions. *Sci Rep* **5**, 8530 (2015).
37. Allen, M.D. & Zhang, J. Subcellular dynamics of protein kinase A activity visualized by FRET-based reporters. *Biochem Biophys Res Commun* **348**, 716-721 (2006).
38. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536 (2008).
39. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **10**, M111 007690 (2011).
40. Liu, G. et al. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat Biotechnol* **28**, 1015-1017 (2010).

41. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22-24 (2013).
42. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002).
43. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003).
44. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* **10**, 730-736 (2013).
45. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198 (2019).
46. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
47. The Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**, D331-D338 (2017).
48. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515 (2019).
49. Finn, R.D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285 (2016).
50. Knight, J.D.R. et al. ProHits-viz: a suite of web tools for visualizing interaction proteomics data. *Nat Methods* **14**, 645-646 (2017).
51. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
52. Baryshnikova, A. Systematic Functional Annotation and Visualization of Biological Networks. *Cell Syst* **2**, 412-421 (2016).
53. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).
54. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).
55. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* **44**, W83-89 (2016).
56. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

Figure 1

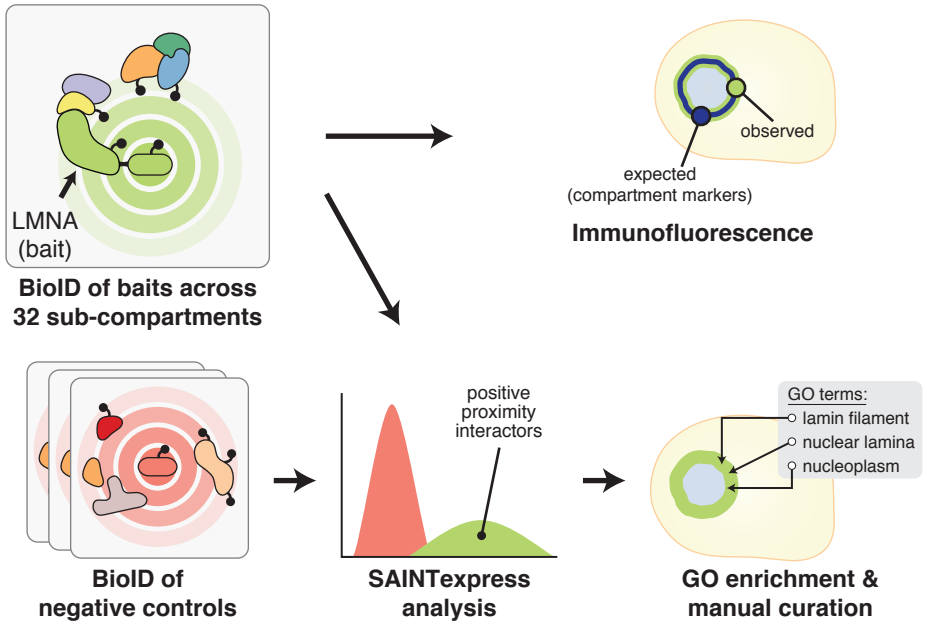
a



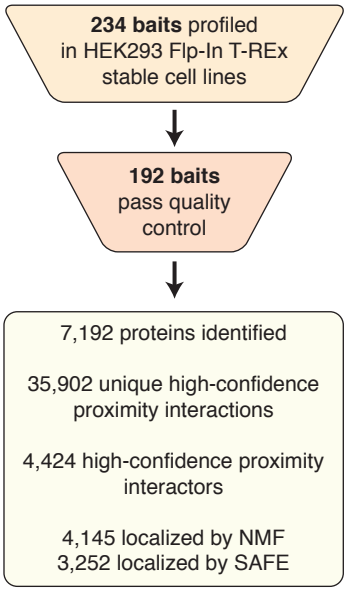
Legend (# of baits)

1. cell junction *(17)*
2. centrosome *(11)*
3. clathrin coat *(3)*
4. cytoplasm *(10)*
5. cytoskeleton *(34)*
   6. actin *(7)*
   7. intermediate filament *(7)*
   8. microtubules *(13)*
9. endosomal/lysosomal system *(19)*
   10. early *(5)*
   11. late endosome/lysosome *(9)*
   12. recycling *(6)*
13. endoplasmic reticulum *(30)*
   14. lumen *(8)*
   15. membrane *(19)*
16. ERGIC *(3)*
17. Golgi apparatus *(10)*
18. lipid droplet *(3)*
19. mitochondrion *(12)*
   20. inner membrane *(4)*
   21. matrix *(6)*
   22. outer membrane *(2)*
23. nucleus *(42)*
   24. chromatin *(4)*
   25. membrane *(3)*
   26. lamina *(2)*
   27. nuclear body *(5)*
   28. nucleolus *(15)*
   29. nucleoplasm *(15)*
   30. pore *(3)*
31. peroxisome *(5)*
32. plasma membrane *(22)*

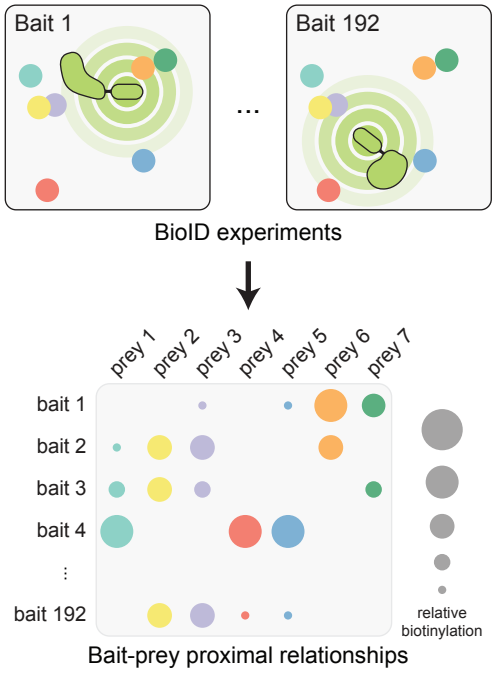
b



c



d



e

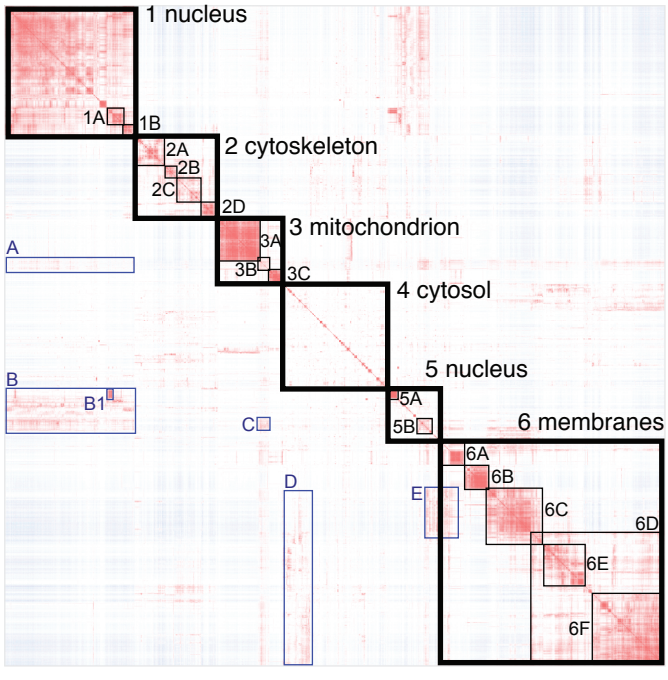
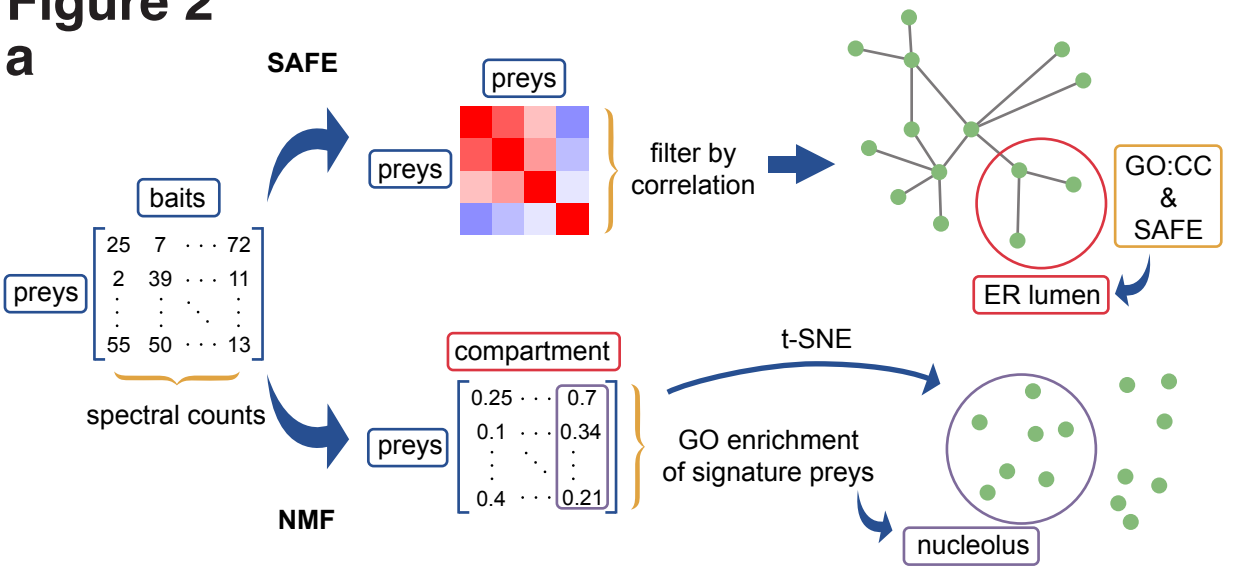
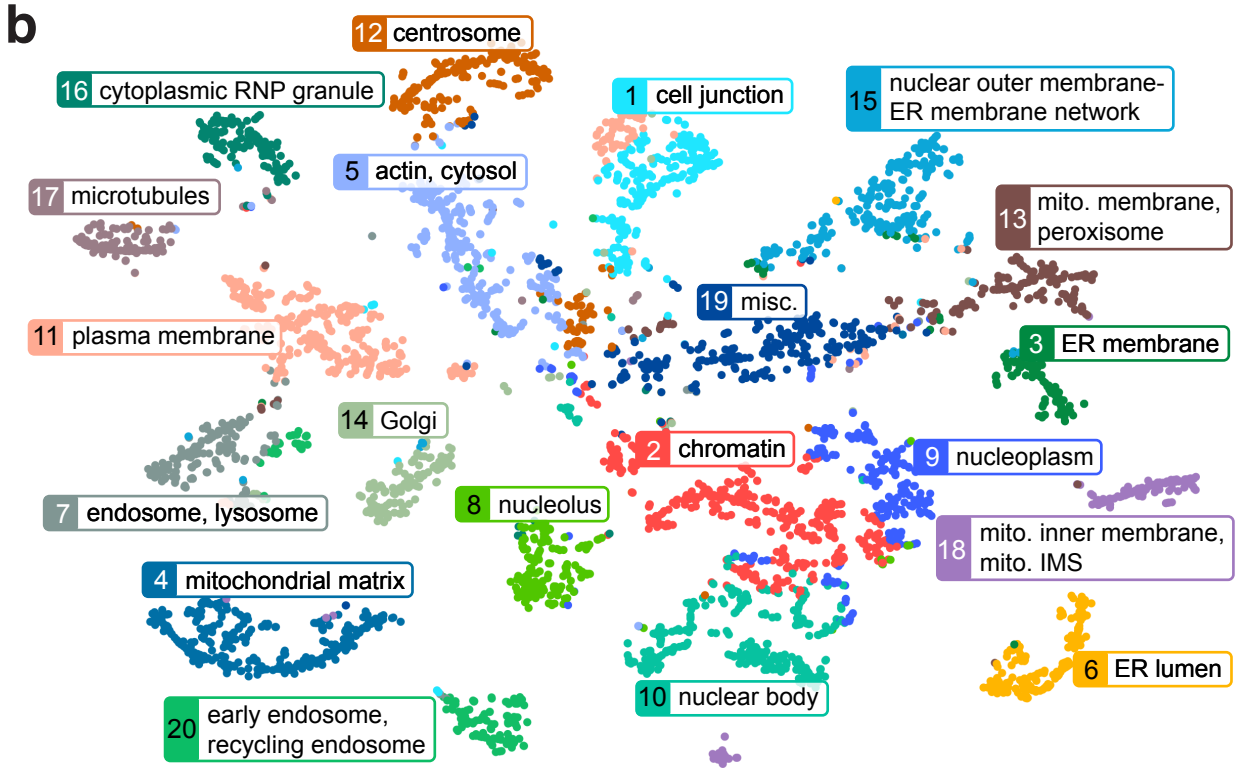
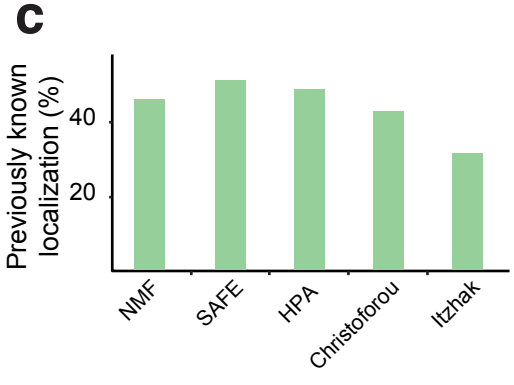


Figure 2


a
SAFE
preys
baits
preys
25 7 ··· 72
2 39 ··· 11
55 50 ··· 13
spectral counts
preys
preys
filter by correlation
GO:CC & SAFE
ER lumen
compartment
0.25 ··· 0.7
0.1 ··· 0.34
0.4 ··· 0.21
NMF
t-SNE
GO enrichment of signature preys
nucleolus


b
12 centrosome
16 cytoplasmic RNP granule
1 cell junction
15 nuclear outer membrane-ER membrane network
5 actin, cytosol
17 microtubules
13 mito. membrane, peroxisome
19 misc.
11 plasma membrane
3 ER membrane
14 Golgi
2 chromatin
9 nucleoplasm
7 endosome, lysosome
8 nucleolus
18 mito. inner membrane, mito. IMS
4 mitochondrial matrix
6 ER lumen
10 nuclear body
20 early endosome, recycling endosome


c
Previously known localization (%)
40
20
NMF
SAFE
HPA
Christoforou
Itzhak

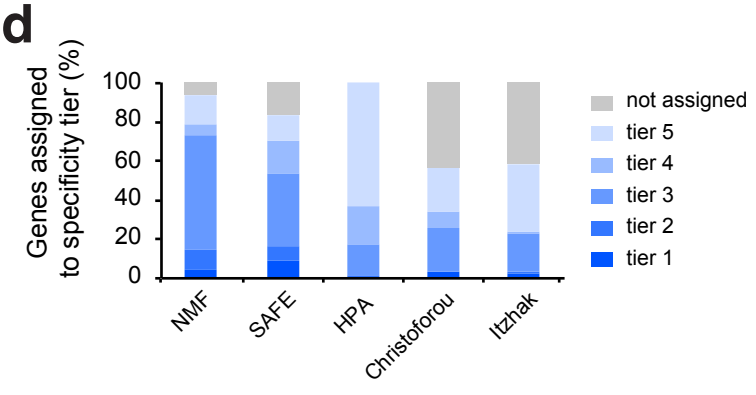
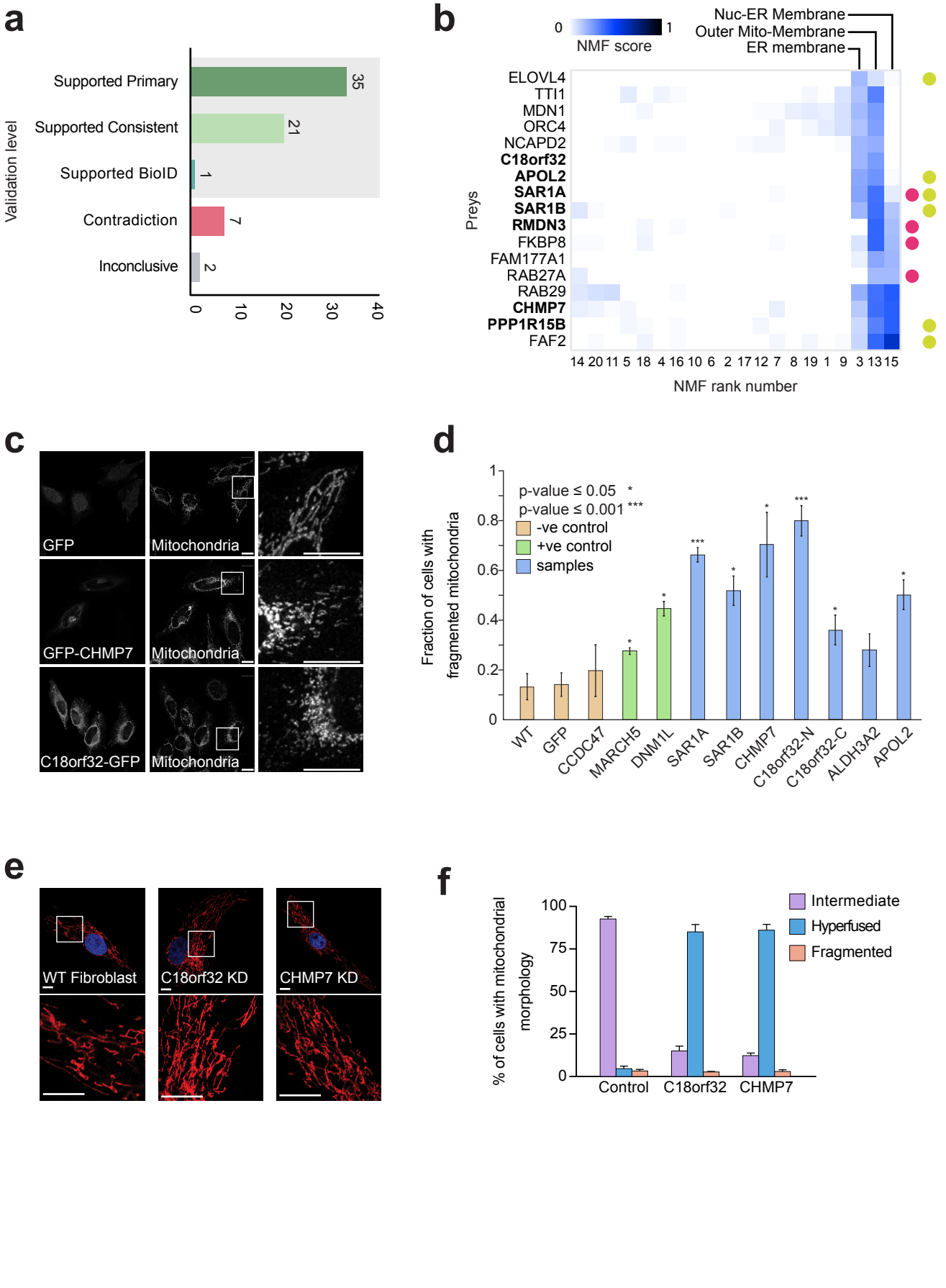

d
Genes assigned to specificity tier (%)
100
80
60
40
20
0
NMF
SAFE
HPA
Christoforou
Itzhak
not assigned
tier 5
tier 4
tier 3
tier 2
tier 1

Figure 3


a
Validation level
Supported Primary
Supported Consistent
Supported BioID
Contradiction
Inconclusive
35
21
1
7
2
0
10
20
30
40
b
0
1
NMF score
Nuc-ER Membrane
Outer Mito-Membrane
ER membrane
Preys
ELOVL4
TTI1
MDN1
ORC4
NCAPD2
C18orf32
APOL2
SAR1A
SAR1B
RMDN3
FKBP8
FAM177A1
RAB27A
RAB29
CHMP7
PPP1R15B
FAF2
14 20 11 5 18 4 16 10 6 2 17 12 7 8 19 1 9 3 13 15
NMF rank number
c
GFP
Mitochondria
GFP-CHMP7
Mitochondria
C18orf32-GFP
Mitochondria
d
Fraction of cells with fragmented mitochondria
p-value ≤ 0.05 *
p-value ≤ 0.001 ***
-ve control
+ve control
samples
0
0.2
0.4
0.6
0.8
1
WT
GFP
CCDC47
MARCH5
DNM1L
SAR1A
SAR1B
CHMP7
C18orf32-N
C18orf32-C
ALDH3A2
APOL2
e
WT Fibroblast
C18orf32 KD
CHMP7 KD
f
% of cells with mitochondrial morphology
Intermediate
Hyperfused
Fragmented
0
25
50
75
100
Control
C18orf32
CHMP7