Article

A proximity-dependent biotinylation map of a human cell

https://doi.org/10.1038/s41586-021-03592-2

Received: 25 September 2019

Accepted: 29 April 2021

Published online: 02 June 2021

Check for updates

Christopher D. Go[1,2,11], James D. R. Knight[1,11], Archita Rajasekharan[3], Bhavisha Rathod[1], Geoffrey G. Hesketh[1], Kento T. Abe[1,2], Ji-Young Youn[1,2,7], Payman Samavarchi-Tehrani[1], Hui Zhang[4], Lucie Y. Zhu[4], Evelyn Popiel[2], Jean-Philippe Lambert[1,8,9], Étienne Coyaud[5,10], Sally W. T. Cheung[1], Dushyandi Rajendran[1], Cassandra J. Wong[1], Hana Antonicka[3], Laurence Pelletier[1,2], Alexander F. Palazzo[4], Eric A. Shoubridge[3], Brian Raught[5,6] & Anne-Claude Gingras[1,2] ✉

Compartmentalization is a defining characteristic of eukaryotic cells, and partitions distinct biochemical processes into discrete subcellular locations. Microscopy[1] and biochemical fractionation coupled with mass spectrometry[2–4] have defined the proteomes of a variety of different organelles, but many intracellular compartments have remained refractory to such approaches. Proximity-dependent biotinylation techniques such as BioID provide an alternative approach to define the composition of cellular compartments in living cells[5–7]. Here we present a BioID-based map of a human cell on the basis of 192 subcellular markers, and define the intracellular locations of 4,145 unique proteins in HEK293 cells. Our localization predictions exceed the specificity of previous approaches, and enabled the discovery of proteins at the interface between the mitochondrial outer membrane and the endoplasmic reticulum that are crucial for mitochondrial homeostasis. On the basis of this dataset, we created humancellmap.org as a community resource that provides online tools for localization analysis of user BioID data, and demonstrate how this resource can be used to understand BioID results better.

Proximity-dependent biotinylation approaches can be used to characterize the intracellular environment occupied by a protein in living cells[5,8]. In BioID, a mutant *Escherichia coli* biotin ligase (BirA*, R118G) is fused to a 'bait' polypeptide of interest, and the resulting fusion protein expressed in cultured cells or organisms[9]. The abortive BirA* enzyme releases biotinoyl-AMP into the local environment, where it covalently labels lysine residues[5] within approximately 10 nm of the bait protein[10]. Covalent biotinylation enables harsh lysis conditions to be used to solubilize proteins from even poorly soluble intracellular compartments (for example, membranes, chromatin or the nuclear lamina). Biotinylated proteins are then captured by streptavidin affinity, and identified by mass spectrometry. Because the average globular protein is 5–10 nm in diameter, the labelling radius of this technique favours the biotinylation of direct binding partners, other components of protein complexes in which the bait resides, and proteins in the immediate intracellular 'neighbourhood'. BioID has been successfully used to define the composition of many different protein complexes and the spatial organization of several membrane-bound and membraneless organelles (see, for example, refs. [5–8]).

To characterize the organization of the proteome in living human cells further, we used BioID to profile 234 intracellular protein markers (baits) for 32 different cellular compartments (Extended Data Fig. 1a, Supplementary Table 1). Each bait was tagged with BirA*, stably expressed in HEK293 Flp-In T-REx cells, and processed for BioID (Methods). SAINTexpress was used to identify high-confidence proximity interactors (referred to here as 'prey' proteins) by scoring spectral counts against a large set of negative controls (Supplementary Table 2). Reproducibility was high across replicates ($R^2 = 0.95$) (Supplementary Table 2), and quality control for each marker included immunofluorescence microscopy to confirm the expected localization (Fig. 1a, Supplementary Table 1) and Gene Ontology (GO) enrichment analysis of preys (Supplementary Table 3). With the exception of the Golgi lumen (for which all tested baits remained trapped in the endoplasmic reticulum (ER)), all selected compartments were successfully characterized with several baits.

In total, 192 candidate markers passed quality control, and 35,902 interactions were established with 4,424 unique high-confidence proximity interactors (Fig. 1b). A clear correlation was observed between prey abundance detected by mass spectrometry and the probability that the prey was previously reported as an interactor (Extended Data Fig. 2a, b, Supplementary Table 4). Prey detection was not correlated with protein turnover rate or the number of lysine residues per protein, but was correlated with protein expression level (Extended Data Fig. 2c–f, Supplementary Table 5). Notably, although

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health System, Toronto, Ontario, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3]Montreal Neurological Institute and Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [4]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada. [5]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [6]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [7]Present address: Peter Gilgan Centre for Research and Learning, Hospital for Sick Children, Toronto, Ontario, Canada. [8]Present address: Department of Molecular Medicine, Cancer Research Centre, Big Data Research Centre, Université Laval, Quebec City, Quebec, Canada. [9]Present address: CHU de Québec-Université Laval Research Center (CHUL), Quebec City, Quebec, Canada. [10]Present address: PRISM INSERM U1192, Université de Lille, Villeneuve d'Ascq, France. [11]These authors contributed equally: Christopher D. Go, James D. R. Knight. ✉e-mail: gingras@lunenfeld.ca

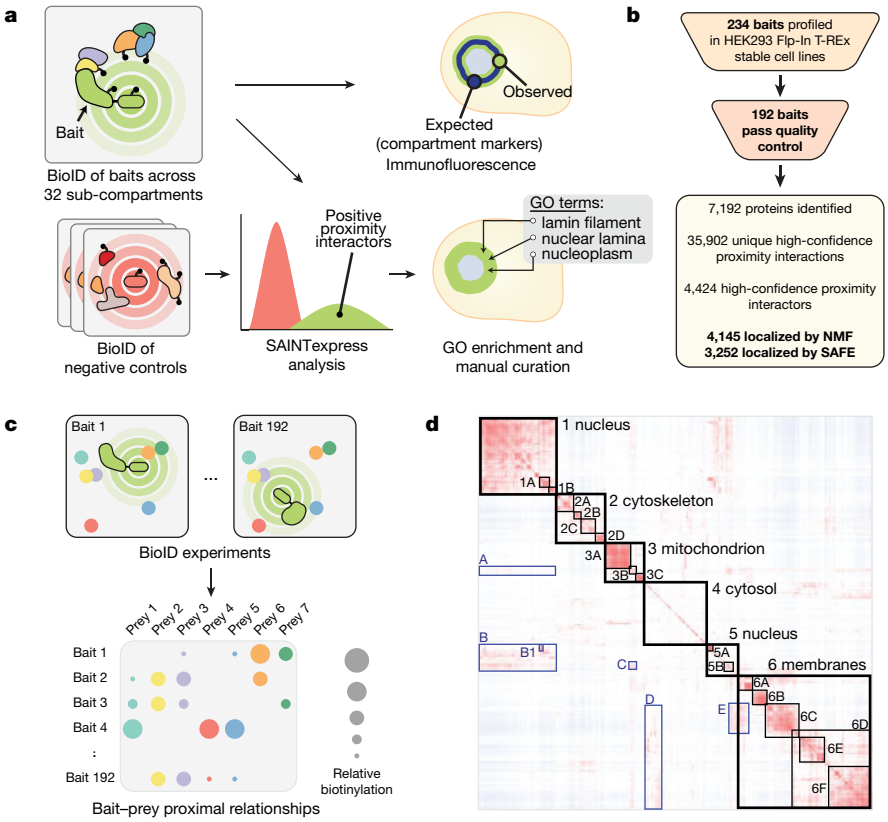


Fig. 1 | Generation and analysis of BioID dataset, and validation strategy. **a**, A bait protein (lime green) is fused to an abortive mutant of the *E. coli* biotin ligase BirA (R118G or BirA*), and expressed in live cells. Highly reactive biotinoyl-AMP is generated by BirA*, and released into the immediate intracellular neighbourhood. Biotinoyl-AMP covalently reacts with the epsilon amine groups of lysine residues in proteins (yellow, orange, blue), with an effective labelling radius of around 10 nm. Affinity purification is used to isolate biotinylated proteins, and mass spectrometry is used to identify them. Interactomes for each bait protein are compared against a large set of control BioID analyses (using SAINTexpress) to identify high-confidence proximity interactors. For quality control, bait localization was assessed by immunofluorescence and GO term enrichment of significant proximity interactors. **b**, Dataset summary. Quality control excluded 42 baits from the original set of 234, and SAINT analysis of the final bait set yielded 35,902 interactions with 4,424 unique high-confidence proximity interactors. **c**, Rationale for prey-association based localization. The relative labelling of preys across baits is dependent on the proximity of prey proteins to each other in situ. This correlation, which is on the basis of shared labelling patterns, can be used to reveal compartment locations. **d**, 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 (annotations and GO enrichments of the highlighted clusters are in Supplementary Table 6). Many subclusters within large organelle clusters are much larger than traditional protein complexes, which suggests that there is a layer of organization between complexes and organelles, possibly akin to protein communities[26].

baits that localized to the same compartment shared similarity in prey profiles (Extended Data Fig. 1b), they also exhibit highly specific interactions with unique prey subsets. For example, although the interactomes of the mitochondrial matrix baits AARS2 (a tRNA synthetase) and PDHA1 (pyruvate dehydrogenase) yielded a Jaccard index of 0.66, AARS2 preys were also enriched for components of the mitochondrial ribosome and translation factors, and PDHA1 preferentially recovered components of the pyruvate dehydrogenase complex (Extended Data Fig. 2g, Supplementary Table 2). As previously discussed[6,7,11,12], BioID can thus provide both compartment and sub-compartmental resolution.

To localize preys to discrete intracellular subcompartments, we exploited the fact that prey proteins with correlated behaviour—that is, those that interact with the same baits—probably reside in the same multiprotein complex, organelle or subcellular region[7] (Fig. 1c). Pearson correlation of prey profiles highlighted several clearly defined cellular (sub-)compartments (Fig. 1d, Supplementary Table 6). Spatial analysis of functional enrichment (SAFE) (Extended Data Fig. 2h) localized 3,252 of the 4,424 high-confidence prey proteins to 23 intracellular compartments (Extended Data Fig. 3, Supplementary Table 7), and the application of non-negative matrix factorization (NMF) localized 4,145 preys to 20 compartments (Fig. 2a, Extended Data Fig. 4, Supplementary Table 8). We found that 54% of the localizations assigned by SAFE were previously reported, and 50% of those assigned by NMF. When both SAFE and NMF made a prediction (3,252 proteins), they were consistent in 88% (2,855) of cases (Supplementary Tables 9, 10).

Compartments identified by the NMF and SAFE pipelines displayed enrichment for expected protein domains and motifs (Supplementary Table 11). For example, the plasma membrane compartment is significantly enriched for pleckstrin homology (PH), immunoglobulin, RhoGAP, RhoGEF and tyrosine kinase domains, whereas the related cell junction compartment is enriched for PDZ and FERM domains. Nuclear sub-compartments were similarly distinguished, with the chromatin compartment enriched for the KRAB domain, C2H2 zinc-fingers, bromodomains and the PWWP domain; the nucleolus enriched for the DEAD and helicase domains; and the nuclear body compartment enriched for the RNA recognition motif (RRM) and G-patch domains. Compartments also exhibited clear enrichment for specific protein sequence motifs or characteristics (coiled-coiled, disordered, low complexity, signal peptide and transmembrane), but in contrast to domains, sequence motifs were often shared between compartments (Extended Data Figs. 3b, 4b).

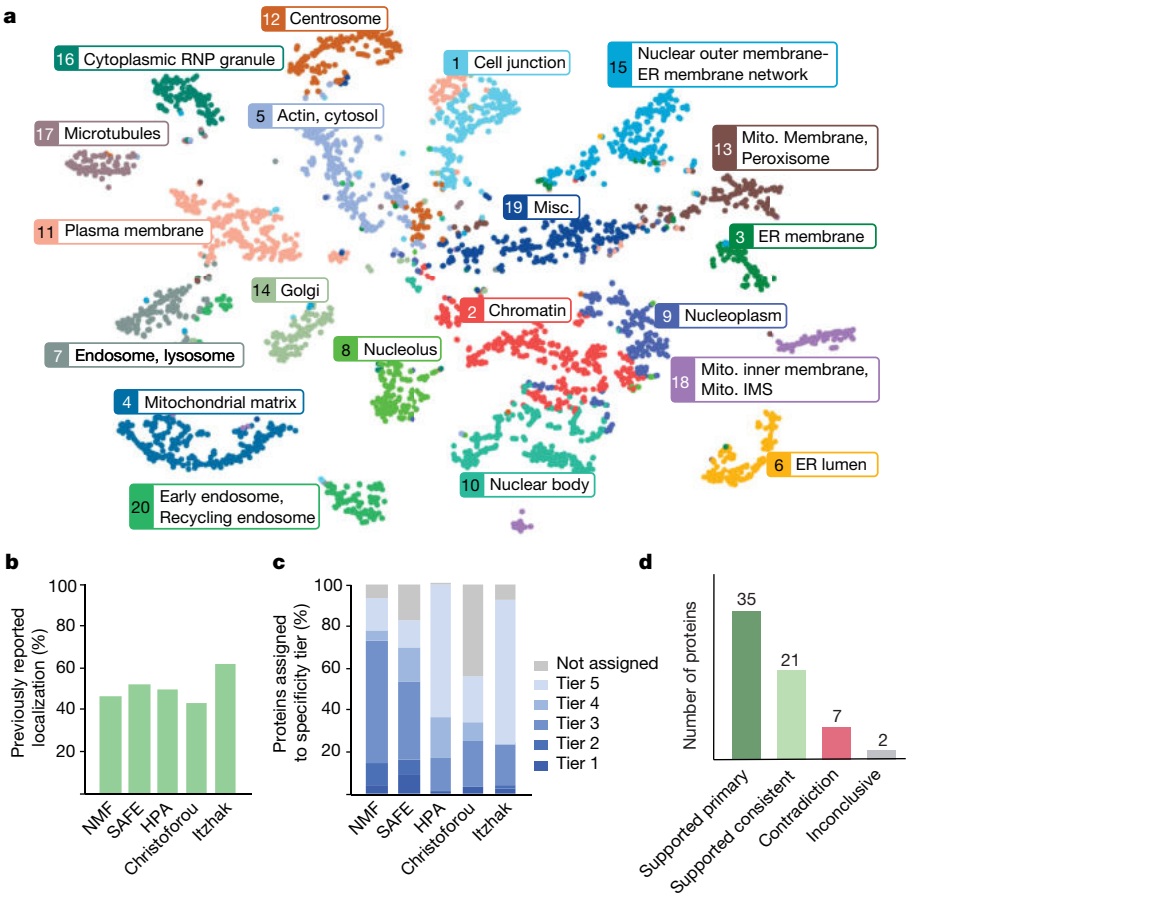


Fig. 2 | Localization of proteins using prey-centric analysis. a, NMF-based map of the cell generated by *t*-distributed stochastic neighbour embedding (*t*-SNE). Prey colour indicates primary localization. IMS, intermembrane space. **b**, Performance of NMF and SAFE compared with the immunofluorescence-based HPA[1] (www.proteinatlas.org) and the fractionation studies of Christoforou et al.[2] and Itzhak et al.[3]. The *y* axis indicates the number of proteins assigned to a previously reported localization (Gene Ontology: cellular component (GO:CC) term). **c**, Specificity of localization assignments. GO:CC terms were binned into specificity tiers (1: most specific; 5: least specific), and the number of proteins assigned to each tier was quantified for our pipelines and the indicated published studies. **d**, Summary of experimental validations for predicted localizations of proteins by immunofluorescence microscopy. 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; 'contradiction' indicates proteins that failed to localize to the predicted localization made by NMF and SAFE; 'inconclusive' indicates proteins with no clear subcellular compartment localization.

We next compared our localization predictions with those made by large-scale microscopy and fractionation studies. After removing Human Protein Atlas (HPA) annotations from GO to prevent self-validation, our recovery of known protein localizations for NMF and SAFE analyses was similar to HPA and the fractionation approaches (Fig. 2b, Supplementary Table 12, Extended Data Fig. 5a–c). However, after binning GO localization annotations into 'precision tiers', with tier 1 containing the most specific localizations (for example, terms such as 'peroxisome' or 'spliceosome') and tier 5 the least specific (for example, 'cytoplasm' or 'nucleus'), we found that our predictions are more specific than those of other approaches. For example, 73% of proteins were localized to the tier 3 bin or better in our NMF analysis (54% for SAFE), versus 17–25% for the other datasets (Fig. 2c, Supplementary Table 12). High recall of known localizations with increased specificity is thus a marked advantage of this methodology.

To further assess the accuracy of our predictions, immunofluorescence microscopy was conducted, focusing on poorly studied proteins (for example, annotated as an open reading frame (ORF), or simply as 'family with sequence similarity' (FAM)) and protein families. A total of 65 green fluorescent protein (GFP)-tagged prey proteins assigned the same annotations by NMF and SAFE were transiently expressed in HEK293 cells, and colocalization with well characterized markers was assessed. We found that 86% (56 out of 65) of the predictions tested were supported by this method, and 17 out of 20 proteins re-tested after stable expression recapitulated this localization (Fig. 2d, Extended Data Fig. 5d, Supplementary Table 13). The cell map analysis pipeline can thus correctly assign subcellular localization to poorly characterized proteins.

The topology of membrane-associated prey proteins may also be predicted from our data, on the basis of the exposure of non-transmembrane stretches of protein sequence to the cytosol or lumen (Extended Data Fig. 6a). For example, all transmembrane-domain-containing prey proteins localized to the ER membrane were assigned a cytosol/lumen ratio (CLR) score, on the basis of their NMF localization scores for the cytosolic or lumenal faces of the ER. For proteins with known sequence orientation information, this metric showed clear correlation ($R^2 = 0.42$) with existing annotations (Extended Data Fig. 6b, c, Supplementary Table 14), which suggests that this approach could be used to predict the orientation of uncharacterized preys.

Although many proteins (684) yielded data that were consistent with exchange between contiguous intracellular compartments (for example, between cell junctions and plasma membrane; early and late endosomes; and across nuclear substructures), only 305 (7%) prey proteins exhibited a clear signature in at least two non-contiguous cell

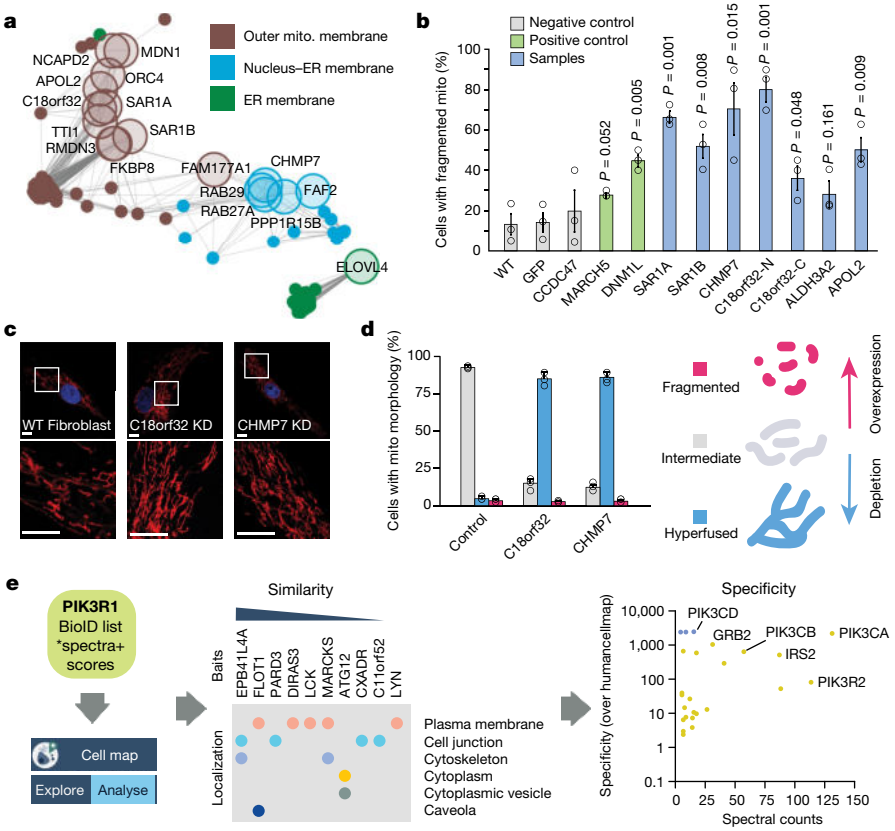


Fig. 3 | Discovery of new connections using the humancellmap. **a**, Sub-network from Extended Data Fig. 4a of the ER–mitochondrial proteins selected for follow-up, and their first degree neighbours. ER–mitochondrial proteins are shown as large transparent nodes with labels. First-degree neighbours are shown as small filled nodes. Edges connect proteins that share a correlation score of ≥0.9 across all NMF compartments. **b**, Quantification of HeLa cells with fragmented mitochondrial morphology after overexpression of GFP-tagged proteins. Negative controls are in grey; positive controls MARCH5 and DNM1L are in green; candidates are in blue. WT, wild type. $n = 3$ biologically independent experiments were performed. *P* values were determined by a two-sided Student's *t*-test. Data are mean and s.d. **c**, Immunofluorescence of mitochondrial morphology in human primary fibroblasts after siRNA-mediated knockdown (KD). White box indicates magnified area in bottom panels. The mitochondrial marker is an antibody against cytochrome *c* (Methods). Scale bars, 10 µm. Images are representative of $n = 3$ biologically independent experiments. **d**, Quantification of primary fibroblast mitochondrial morphology after siRNA-mediated knockdown. The fraction of cells with hyperfused, fragmented and intermediate mitochondrial morphology is indicated in blue, red and grey, respectively. $n = 3$ biologically independent experiments were performed. Data are mean and s.d. Right, phenotype associated with overexpression and depletion of C18orf32 and CHMP7. **e**, Illustration of the humancellmap.org 'Analyze' function: bait similarity and specificity enrichment, as applied to the BioID data of test bait PIK3R1, are depicted (screenshots and further examples are in Extended Data Fig. 9).

compartments (Supplementary Tables 8, 15)—a phenomenon that could be due to moonlighting[1,13] or localization to membrane contact sites[14]. Very little inter-compartmental crosstalk indicative of moonlighting was detected in our dataset between non-contiguous compartments (Extended Data Fig. 6d, e, Supplementary Table 15), perhaps because of the relatively long biotinylation times that could disfavour more dynamic and/or condition-dependent localizations. Future experiments using faster BioID enzymes[15] (which provide results compatible with the original BirA* enzyme used here) (Extended Data Fig. 7, Supplementary Table 16) may help to define moonlighting activities better.

Contacts between the mitochondria and ER are crucial for lipid and calcium exchange, and mitochondrial dynamics[16–18]. Inter-compartmental crosstalk analyses identified 17 proteins that were linked to both the mitochondria or peroxisome and ER membrane compartments (Fig. 3a, Extended Data Fig. 8a, Supplementary Table 17), including SAR1A and SAR1B, two GTPases that regulate mitochondria–ER contact sites[19]. Reciprocal BioID conducted on six of these proteins revealed strong enrichment of ER components, but SAR1A, SAR1B, C18orf32 and CHMP7 (and to a lesser extent APOL2 and PPP1R15B) also recovered components of the outer mitochondrial membrane and ER–mitochondrial membrane contact sites (Extended Data Fig. 8b, Supplementary Table 17). This set of baits also detected two proteins linked to mitochondrial fission—DNM1L (orthologous to yeast Drp1[20]) and INF2 (a formin that mediates actin-dependent fission[21]). To test whether these proteins were involved in mitochondrial dynamics, we overexpressed them as GFP fusions and quantified fragmented mitochondrial morphology. The expression of APOL2 (apolipoprotein L2), C18orf32 (a protein that traffics to lipid droplets[22]) and CHMP7 (an ESCRT-III component) induced mitochondrial fragmentation similar to, or to a higher degree than, the positive controls MARCH5 (also known as MITOL)[23] and DNM1L (Fig. 3b, Extended Data Fig. 8c). By contrast, short interfering RNA (siRNA)-mediated knockdown of C18orf32 and CHMP7 in fibroblasts induced a notable hyperfused mitochondrial phenotype, which suggests that these proteins are important for mitochondrial homeostasis (Fig. 3c, d). Although the specific roles of these proteins in mitochondrial dynamics remain to be defined, these results demonstrate how our dataset can be mined to reveal new protein functions within and between different subcellular compartments.

To facilitate exploration of our dataset, we created humancellmap.org, which enables searching and viewing data on all profiled baits, identified preys and organelles. 'Help' documentation that describes all available features for the site can be found at humancellmap.org/help.

A key feature of the website is the ability to compare user BioID data to the humancellmap database (Extended Data Fig. 9a), to localize a query bait to a specific subcellular compartment(s) on the basis of similarity with other bait interactomes, to identify previously queried baits with similar interactomes, and to identify those preys that are most specific to the queried bait. To illustrate these functions, we performed BioID on PIK3R1, an SH2 domain-containing adaptor protein that recruits PI3K to activated receptor complexes at the plasma membrane. Analysis of its BioID data using humancellmap.org revealed that the baits in our dataset with interactomes similar to PIK3R1 localize primarily to the plasma membrane and cell junction. Highly specific proximity interactions (as compared with all humancellmap baits) with PI3K catalytic subunits, insulin receptor substrate proteins (IRS2 and IRS4) and the scaffold protein GRB2 were also detected, as expected[24] (Fig. 3e). Other examples that confirm previous knowledge (RNGTT[7]), or highlight potential new associations for poorly characterized proteins (FAM171A1, FAM171B and MTFR2 (also known as FAM54A)) (Supplementary Table 18) further illustrate the usefulness of this analysis module (Extended Data Fig. 9b–d). Finally, we re-analysed BioID data on the bromodomain-containing protein BRD3 in cells treated with the BET inhibitor JQ1, which leads to a relocalization of BRD3 to the nucleolus[25]. This relocalization was apparent when analysed against the humancellmap, with JQ1-treated BRD3 more similar to nucleolar baits (Extended Data Fig. 9e, Supplementary Table 19), which attests to the applicability of the humancellmap for the exploration of condition-dependent BioID datasets.

Future versions of the humancellmap will feature higher density coverage of baits (for example, by merging published organelle-specific datasets[6,7,11] into humancellmap.org), include other cell types, and highlight dynamic interactomes profiled with faster enzymes, supplementing other proteomics and cell biological resources.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03592-2.

1. Thul, P. J. et al. A subcellular map of the human proteome. *Science* **356**, eaal3321 (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**, e16950 (2016).
4. Orre, L. M. et al. SubCellBarCode: Proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73**, 166–182 (2019).
5. 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).
6. Gupta, G. D. et al. A dynamic protein interaction landscape of the human centrosome-cilium interface. *Cell* **163**, 1484–1499 (2015).
7. Youn, J. Y. et al. High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Mol. Cell* **69**, 517–532 (2018).
8. Rhee, H. W. et al. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013).
9. Gingras, A. C., Abe, K. T. & Raught, B. Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. *Curr. Opin. Chem. Biol.* **48**, 44–54 (2019).
10. Kim, D. I. et al. Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl Acad. Sci. USA* **111**, E2453–E2461 (2014).
11. Antonicka, H. et al. A high-density human mitochondrial proximity interaction network. *Cell Metab.* **32**, 479–497 (2020).
12. Botham, A. et al. Global interactome mapping of mitochondrial intermembrane space proteases identifies a novel function for HTRA2. *Proteomics* **19**, e1900139 (2019).
13. Chapple, C. E. et al. Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* **6**, 7412 (2015).
14. 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).
15. Branon, T. C. et al. Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* **36**, 880–887 (2018).
16. Lee, S. & Min, K. T. The interface between ER and mitochondria: molecular compositions and functions. *Mol. Cells* **41**, 1000–1007 (2018).
17. 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).
18. Rowland, A. A. & Voeltz, G. K. Endoplasmic reticulum-mitochondria contacts: function of the junction. *Nat. Rev. Mol. Cell Biol.* **13**, 607–625 (2012).
19. Ackema, K. B. et al. Sar1, a novel regulator of ER-mitochondrial contact sites. *PLoS ONE* **11**, e0154280 (2016).
20. Kalia, R. et al. Structural basis of mitochondrial receptor binding and constriction by DRP1. *Nature* **558**, 401–405 (2018).
21. 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).
22. Bersuker, K. et al. A proximity labeling strategy provides insights into the composition and dynamics of lipid droplet proteomes. *Dev. Cell* **44**, 97–112 (2018).
23. 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).
24. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45** (D1), D369–D379 (2017).
25. Lambert, J. P. et al. Interactome rewiring following pharmacological targeting of BET bromodomains. *Mol. Cell* **73**, 621–638 (2019).
26. Kastritis, P. L. et al. Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol. Syst. Biol.* **13**, 936 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Methods

Data reporting

No statistical methods were used to predetermine sample size. Mass spectrometry sample acquisition was randomized but no further experimental randomization was done. Investigators were blinded to allocation during experiments and outcome assessment for quantifying defects in mitochondrial morphology in fibroblasts. Investigators were not blinded in any other experiments or outcome assessments.

Selection of compartment markers

We aimed to select at least three independent baits (referred to here 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 sub-compartments (for example, matrix, inner membrane and outer membrane for the mitochondria). These markers were selected by manual literature curation (for example, 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 HPA[1], commercially used as compartment markers for immunofluorescence (for example, Cell Signaling Technology), or following advice from experts in cell biology. The list of the constructs used is 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, the same tag location was used for BioID. For proteins without such information available (such as those used as endogenous markers), the structural organization of the protein was taken into consideration (for example, if a crucial domain or motif such as a mitochondrial localization sequence or prenylation motif, was present at one of the termini, the other terminus was used for tagging). For transmembrane domain-containing proteins, membrane topology was analysed from both the literature and using the Protter tool[27], and the tag integrated on the side of the membrane where compartment labelling was desired. In six 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[28] (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 for N-terminal fusions were first 'closed' by PCR amplification and re-cloned as closed entries to prevent cloning scars[29]. Sequence tags[30] were PCR-amplified from relevant cDNA or Gateway ORF clones of the full-length proteins, or from oligonucleotide annealing, and inserted into the pcDNA5-Flag-BirA* backbone. All constructs generated by PCR amplification were validated by Sanger sequencing.

Cell lines

HEK293 Flp-In T-REx cells were from Invitrogen and were authenticated by short-tandem repeat analysis with The Center for Applied Genomics Genetic Analysis Facility (Sick Kids Hospital, Toronto). HeLa cells were from the ATCC (CCL-2) and were not independently validated. Primary fibroblasts were from the Cell Bank at Montreal Children's Hospital and were not independently validated. Parental cell lines were routinely monitored for mycoplasma contamination as assessed by a commercial kit (MycoAlert, Lonza).

Cell line generation for BioID

For BioID, the parental cell line, HEK293 Flp-In T-REx 293, was grown at 37 °C in DMEM high-glucose medium supplemented with 5% fetal bovine serum, 5% cosmic calf serum and 100 U ml^{-1} penicillin/streptomycin (growth medium).

For the generation of stable cell lines, HEK293 Flp-In T-REx cells were transfected using the jetPRIME transfection reagent (Polyplus CA89129-924). Cells were seeded at 250,000 cells per well in a 6-well plate in 2 ml growth medium (day 0). The next 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 medium (final concentration of 200 µg ml^{-1}). This selection medium 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 cm^2 plates. Cells were grown to 70% confluence before the induction of protein expression using 1 µg ml^{-1} tetracycline, and the medium was supplemented with 50 µM biotin for protein labelling. Cells were collected 24 h later as follows: cell medium was decanted, cells were washed once with 5 ml PBS, then collected by scraping in 1 ml PBS. Cells from one or two 150 cm^2 plates were pelleted at 233*g* for 5 min, the supernatant was aspirated, and pellets were frozen on dry ice. Cell pellets were stored at −80 °C until further processing.

BioID

The BirA* enzyme used in this study for profiling compartments was the original BioID enzyme previously described[5]. 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 grams:lysis buffer volume in millilitres). After the addition of lysis buffer, 1 µl of benzonase (EMD, CA80601-766, 250 U) was added to each sample, and cell pellets were incubated on a nutator at 4 °C for 20 min. Lysates were sonicated (3 × 10 s bursts with 2 s rest) on ice at 65% amplitude using a Qsonica with a CL-18 probe. Lysates were centrifuged for 30 min at 20,817*g* at 4 °C. After centrifugation, lysate supernatants were added to pre-washed streptavidin-sepharose beads (GE 17-5113-01; 30 µl bed volume of pre-washed beads per sample), and biotinylated proteins were affinity-purified at 4 °C on a nutator for 3 h. After affinity purification, streptavidin sepharose beads were pelleted (400*g*, 1 min), and the supernatant was 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*g* for 1 min. Beads were washed twice with modRIPA and $MgCl_2$ (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 V5071) in 60 µl of 50 mM ammonium bicarbonate was added to each sample. Proteins were digested on beads overnight at 37 °C on a rotator. The next day, an additional 0.5 µg trypsin/Lys-C mix was added to samples that were further digested at 37 °C on a rotator for 2 h. Each sample was spun down at 400*g* 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 7732-18-5), centrifuged at 400*g* for 1 min to pellet beads, and the supernatant was pooled with digested peptides collected previously (this step was repeated once). Samples were centrifuged at 16,000*g* 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 °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 grams:lysis buffer volume in millilitres). 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*g* 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 °C on a rotator. After the addition of an additional 0.5 µg of trypsin and incubation for 2 h, 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 7732-18-5), centrifuged at 400*g* for 1 min to pellet the beads, and the supernatant was 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*g* 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 °C.

Mass spectrometry analysis

Dried peptides were resuspended in 20 µl of 5% formic acid and centrifuged at 16,000*g* for 1 min. Then, 5 µl was 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, 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). 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^{-1} with 5% buffer B for 14 min followed by a drop in flow rates from 400 µl min^{-1} to 200 µl min^{-1} 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^{-1} 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^{-1} 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^{-1} 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 2,000) 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 1,000, 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 bovine serum albumin (BSA) tryptic peptide standard (Michrom Bioresources) 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^{-1} with 5% buffer B, 1 min going from 400 µl min^{-1} to 200 µl min^{-1} using a linear gradient from 5 to 2% buffer B, 30 min at 200 µl min^{-1} using a linear gradient from 2 to 35% buffer B, 5 min at 200 µl min^{-1} using a linear gradient from 35 to 80% buffer B, 5 min at 200 µl min^{-1} with 80% buffer B, 5 min at 200 µl min^{-1} using a linear gradient from 80 to 2% buffer B and 5 min at 200 µl min^{-1} with 2% buffer B.

Mass spectrometry data analysis

Mass spectrometer raw files were converted to mzML using ProteoWizard (3.0.4468)[31] and analysed using the iProphet[32] pipeline implemented within ProHits[33] 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://www.coxdocs.org/doku.php?id=maxquant:start_downloads.htm 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 (labelled 'gi|9999' or 'DECOY'); in total, 72,226 entries (including decoys) were searched. Spectra were analysed separately using Mascot (2.3.02; Matrix Science) and Comet (2012.01 rev.3)[34] 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[35], and peptides were assembled into proteins using parsimony rules first described in ProteinProphet[36] 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[37] 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 analysed alongside 70 negative-control runs, consisting of purifications from untransfected cells or cells that express BirA*-Flag or BirA*-Flag-GFP. For BioID protocol 2, 52 bait protein samples (26 baits) were analysed alongside 16 negative-control runs, consisting of purifications from untransfected cells or cells that express BirA*-Flag or BirA*-Flag-GFP. No compression of the controls was performed and default parameters for SAINTexpress were used. A 1% Bayesian FDR cut-off was used to select confident proximity interactors relative to the expected spectral counts distribution seen in control samples. All prey proteins detected in controls samples and enriched GO terms for the top preys in these samples can be found in Supplementary Table 20. 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 five significant 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.

Immunofluorescence 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, 354085). The next day, cells were treated with 1 μg ml^{-1} tetracycline and medium was supplemented with 50 μM biotin for 24 h. Medium was aspirated, and cells were washed with PBS supplemented with 200 μM $CaCl_2$, 100 μM $MgCl_2$, before fixation with 4% formaldehyde in PBS for 10 min, and washing three times in TBS-T (Tris-buffered saline and 0.1% Tween 20 (v/v)). The cells were then treated for 10 min in permeabilization buffer (0.1% Triton X-100 in TBS-T), followed by three 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 (1:2,000 dilution, Sigma Aldrich, F3165) and an endogenous compartment marker antibody from rabbit (see Supplementary Table 21 for list of antibodies), or anti-Flag from rabbit (1:500 dilution, Sigma Aldrich, F7425) and an endogenous compartment marker antibody from mouse. All samples were then washed three 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 (1:1,000, Invitrogen, A11034), anti-mouse coupled to Alexa Fluor 555 (1:1,000; Invitrogen, A21422), streptavidin-coupled to Alexa Fluor 647 (1:2,500, Invitrogen, S32357); (2) anti-mouse coupled to Alexa Fluor 488 (1:1,000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1,000; Invitrogen, A21428), streptavidin-coupled to Alexa Fluor 647 (1:2,500, Invitrogen, S32357); (3) anti-mouse coupled to Alexa Fluor 555 (1:1,000; Invitrogen, A21422), DAPI (1:2,000), streptavidin-coupled to Alexa Fluor 647 (1:2,500, Invitrogen, S32357); or (4) anti-mouse coupled to Alexa Fluor 555 (1:1,000, Invitrogen, A21422), phalloidin-coupled to Alexa Fluor 488 (1:1,000, Invitrogen, A12379), streptavidin-coupled to Alexa Fluor 647 (1:2,500, Invitrogen, S32357). After incubation, samples were washed three times with TBS-T. Each coverslip was mounted on a glass slide using 4 μl of ProLong Gold Antifade Mountant (Thermo Fisher Scientific, 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 60× objective lens magnification and 3× field zoom.

In some instances, ice-cold methanol was used as a fixative to visualize microtubules better and facilitate the use of specific antibodies only amenable to methanol fixation conditions. Ice-cold methanol 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 three times with TBS-T, the protocol continued as described above with the addition of blocking buffer. When wheat germ agglutinin (WGA)-coupled to Alexa Fluor 488 (1:250, Invitrogen, W11261) was used as a counterstain, all steps were performed with samples chilled on ice, using ice-cold buffers and in the dark. After the initial wash, cells were incubated with a solution containing WGA coupled to Alexa Fluor 488 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.

The localization of negative controls (BirA*-Flag and GFP-BirA*-Flag) can be found in Extended Data Fig. 10.

GO enrichment analysis

GO enrichments were performed using g:Profiler[38]. Enrichments were performed considering gene lists as unordered, allowing only genes with annotations, using all significant proximal interactors as background, a maximum *P* value of 0.01 and the g:SCS multiple test correction method. For bait quality control, 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 maximum *P* value was relaxed from 0.01 to 0.05.

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 $1 - J(A,B)$.

Control-subtracted length-adjusted spectral counts

The prey proximity order for each bait was determined from the prey's control-subtracted length-adjusted spectral counts. For bait *i* and prey *j* this value was calculated by first subtracting the average spectral counts of the prey 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*.

$$\mathrm{NormSpec}_{i,j} = \frac{\widetilde{\mathrm{length}}}{\mathrm{length}_j}(\mathrm{AvgSpec}_{i,j} - \mathrm{AvgSpec}_{\mathrm{control},j})$$

Prey specificity

The specificity of a prey for a particular bait at the humancellmap is calculated as the spectral counts detected with the bait, divided by the mean spectral counts across all other baits or against the top ten most similar baits as indicated on the analysis report. Whenever available, the average spectral counts detected in control samples is subtracted from the prey counts before calculating the specificity, as was done for all specificity metrics reported in this study.

Prey–prey correlation

The SAINTexpress file was processed using the correlation tool at ProHits-viz[39] with an FDR score filter of 0.01 and an abundance cut-off value 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 (0.2-18) in R (version 3.3.3).

SAFE

The Matlab (version 9.4) implementation of SAFE (version 1.5) was used[40]. A network was built from prey–prey correlation data using ProHits-viz as described above. Networks were built in Cytoscape[41] (version 3.6.1) using a spring embedded layout. All preys that passed an FDR cut-off 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 as we could not know a priori what an ideal cut off would be although manual assessment suggested that 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'), in which 60.2% (2,351 out of 3,903) 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

NMF is an approach to create a compressed and simplified version of an $n \times m$ dimension matrix **V**, such that $\mathbf{V} \approx \mathbf{WH}$, in which **W** has dimensions $n \times r$ and **H** has dimensions $r \times m$, and both matrices consist entirely of non-negative entries[42]. Given an interaction matrix of n preys and m baits, where $V_{i,j}$ is the spectral counts 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, in which 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 4,424 × 192 for the 192 baits in the dataset and 4,424 preys passing an FDR cut-off of 1%. Prey spectral counts had their average value in controls subtracted and were then rescaled from 0 to 1 across baits as we wanted each prey to be considered of equal weight. NMF, as implemented optimizing the squared Frobenius norm initialized by Non-negative Double Singular Value Decomposition (NNDSVD) with L1 regularization in the 'scikit-learn' Python package[43], 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[38] 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). Although 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. Because 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 (on the basis of GO:CC annotations) (Supplementary Table 8), 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*-SNE[44]. *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 1,000 iterations.

Information content

The information content of each GO cellular component term was calculated as $-\log(P)$, in which P is the probability a gene has an annotation, that is, 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 out of 18,858 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 HPA subcellular localization data was downloaded on 15 March 2019 from www.proteinatlas.org/about/download, and is based on the Human Protein Atlas[45] 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 Data 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 Data 1, tab 3, columns H, K and N ('compartment prediction', 'sub compartment prediction' and 'global classifier'). Columns H and K were first merged. Missing predictions, 'no prediction' and 'large protein complex' assignments were ignored and instead the classification from column N was used. Localization terms were mapped to the closest GO term. Predictions from canonical isoforms were used when possible or alternative isoforms if the canonical isoform had no prediction. In the case of only non-canonical isoforms, the most specifically localized isoform was used. All genes from these datasets with their assigned and corresponding GO IDs are listed in Supplementary Table 12. Localization tiers were defined using the information content of each GO term as defined in the 'information content' section. When genes were assigned several localizations, the lowest information content term (that is, 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 4,424 genes in our NMF analysis, 4,368 had domain information available and 4,301 had motif information available in Pfam. Of the 3,903 genes in our SAFE analysis, 3,855 had domain information available and 3,809 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 localization predictions by immunofluorescence microscopy and BioID

For prediction validation, we prioritized proteins without a clear annotation, such as proteins annotated as an ORF, or as FAM. We also focused on families that share domains or structural features for which multiple

members were present on our map, but with different predicted localization. These included proteins annotated as solute carriers, transmembrane proteins, and proteins that contain a Rab small GTPase domain. We only selected proteins in which the NMF and SAFE predictions were in agreement and for which we could readily access a full-length cDNA or ORF clone locally. Selected targets for validation were cloned in Gateway compatible pcDNA5-GFP and pcDNA5-Flag-BirA* backbones (with tags at either the 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 (a list of tested baits is in Supplementary Table 13).

GFP-tagged constructs were transiently transfected into HeLa cells using the jetPRIME transfection reagent (Polyplus CA89129-924). Cells were seeded at 250,000 cells per well in a 6-well plate in 2 ml growth medium. The next 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 was used as described above 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, 11814460001) and an endogenous compartment marker antibody from rabbit (refer to Supplementary Table 21 for list of antibodies used), or anti-GFP from rabbit (1:2,000 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 (1:2,000), anti-rabbit coupled to Alexa Fluor 488 (1:1,000, Invitrogen, A11034), anti-mouse coupled to Alexa Fluor 555 (1:1,000; Invitrogen, A21422); (2) DAPI (1:2,000), anti-mouse coupled to Alexa Fluor 488 (1:1,000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1,000; Invitrogen, A21428); or (3) DAPI (1:2,000), anti-rabbit coupled to Alexa Fluor 488 (1:1,000, Invitrogen, A11034), phalloidin-coupled to Alexa Fluor 647 (1:1,000, Invitrogen, A22287). Images were acquired on a Nikon C1Si Confocal Microscope using a 60× objective lens magnification and 1× or 2× 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 analysed alongside 74 negative-control runs, consisting of purifications from untransfected cells or cells that express BirA*-Flag or BirA*-Flag-GFP. GO enrichments were performed using g:Profiler[38]. 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.

Cell culture for mitochondrial fragmentation assays

Primary fibroblasts and HeLa cells were grown in high-glucose DMEM medium supplemented with 10% fetal bovine serum, at 37 °C in an atmosphere of 5% CO_2. Stealth RNA interference duplex constructs (Invitrogen) were used for transient knockdown of C18orf32 and CHMP7 in primary fibroblasts. 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 immunofluorescence experiments for assaying mitochondrial fragmentation, candidate proteins were tagged with GFP and the constructs were transiently transfected into HeLa cells. HeLa cells were transfected using the jetPRIME transfection reagent (Polyplus, CA89129-924). Cells were seeded at 250,000 cells per well in a 6-well plate in 2 ml growth medium. The next 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 immunofluorescence protocol with formaldehyde fixation was used as described above 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, 11814460001) and anti-COXIV from rabbit (1:250, Cell Signaling Technology, 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 (1:1,000, Invitrogen, A11001), anti-rabbit coupled to Alexa Fluor 555 (1:1,000; 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 60× 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 wild-type mitochondrial staining compared to controls (HeLa cells untransfected or with GFP alone). Statistical confidence of mitochondrial fragmentation was calculated using a Student's *t*-test.

Primary fibroblasts were fixed in warm 4% formaldehyde 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% BSA in PBS, followed by incubation with primary antibodies (rat anti-KDEL and mouse anti-cytochrome *c*) (Supplementary Table 21) in 3% BSA in PBS for 1 h 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 (1:2,000, Invitrogen) (Supplementary Table 21) 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 100× 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, 175 cells were analysed, and experiments were done three times independently. Error bars represent mean ± standard deviation.

Statistics and reproducibility

Each experiment for mitochondrial morphology was performed in $n = 3$ biological independent experiments. Immunofluorescence images shown for these experiments are representative of $n = 3$ biological independent experiments.

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

BioID was performed on selected mito–ER candidates 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 analysed 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[38]. Enrichments were performed considering gene lists as unordered, allowing only genes with annotations, using a maximum *P* value of 0.05 and the g:SCS multiple test correction method.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mass spectrometry datasets consisting of raw files and associated peak lists and results files have been deposited in ProteomeXchange through partner Mass spectrometry Interactive Virtual Environment MassIVE (http://proteomics.ucsd.edu/ProteoSAFe/datasets.jsp) as complete submissions. Other 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 (Supplementary Table 2): Go_BioID_humancellmap_HEK293_lowSDS_core_data set_2019 MassIVE ID MSV000084359 and PXD015530. Dataset 2 (Supplementary Table 2): Go_BioID_humancellmap_HEK293_highSDS_core_data set_2019 MassIVE ID MSV000084360 and PXD015531. Dataset 3 (Supplementary Table 18): Go_BioID_humancellmap_HEK293_prediction_2019 MassIVE ID MSV000084369 and PXD015554. Dataset 4 (Supplementary Table 17): 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[46] (CRAPome.org) and assigned samples numbers CC1100 to CC1185 (Supplementary Table 2); this will be part of the next release of the database.

The BioGRID[47] human database v3.5.169 was downloaded on 13 February 2019 (https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.169/). Human gene annotations were downloaded from the GO on 15 February 2019 (GO version date 1 February 2019, http://release.geneontology.org/2019-02-01/annotations/goa_human.gaf.gz). The GO hierarchy (release date 1 February 2019) was downloaded from GO[48,49] on 15 February 2019 (http://release.geneontology.org/2019-02-01/ontology/go-basic.obo). The UniProt database[50] release 2019_2 was downloaded on 21 February 2019 (ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2019_02/knowledgebase/uniprot_sprot-only2019_02.tar.gz). The IntAct[51] human database release 2018_11_30 was downloaded on 13 February 2019 (ftp.ebi.ac.uk/pub/databases/intact/2018-11-30/psimitab/intact.txt). Human protein domain annotations and motifs were retrieved from Pfam[52] (version 32) on 21 February 2019 (ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/proteomes/9606.tsv.gz). ProteomicsDB[53] was queried for protein expression information on 14 January 2020 using their API. Text mining data was downloaded from the Compartments database[54] on 21 January 20 (https://compartments.jensenlab.org/Downloads). Source data are provided with this paper.

Code availability

Source code used for analysis can be accessed from https://github.com/knightjdr/cellmap-scripts.

27. Omasits, U., Ahrens, C. H., Müller, S. & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**, 884–886 (2014).
28. Couzens, A. L. et al. Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci. Signal.* **6**, rs15 (2013).
29. 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).
30. 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).
31. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
32. 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).
33. Liu, G. et al. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat. Biotechnol.* **28**, 1015–1017 (2010).
34. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
35. 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).
36. 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).
37. Teo, G. et al. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J. Proteomics* **100**, 37–43 (2014).
38. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47** (W1), W191–W198 (2019).
39. Knight, J. D. R. et al. ProHits-viz: a suite of web tools for visualizing interaction proteomics data. *Nat. Methods* **14**, 645–646 (2017).
40. Baryshnikova, A. Spatial Analysis of Functional Enrichment (SAFE) in large biological networks. *Methods Mol. Biol.* **1819**, 249–268 (2018).
41. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
42. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
43. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. van der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
45. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
46. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
47. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
48. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
49. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45** (D1), D331–D338 (2017).
50. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47** (D1), D506–D515 (2019).
51. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
52. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44** (D1), D279–D285 (2016).
53. Samaras, P. et al. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res.* **48** (D1), D1153–D1163 (2020).
54. Binder, J. X. et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* **2014**, bau012 (2014).
55. Zecha, J. et al. Peptide level turnover measurements enable the study of proteoform dynamics. *Mol. Cell. Proteomics* **17**, 974–992 (2018).
56. Burkhardt, J. K. In search of membrane receptors for microtubule-based motors - is kinectin a kinesin receptor? *Trends Cell Biol.* **6**, 127–131 (1996).
57. St-Denis, N. et al. Phenotypic and interaction profiling of the human phosphatases identifies diverse mitotic regulators. *Cell Rep.* **17**, 2488–2501 (2016).
58. Li, X. et al. Defining the protein-protein interaction network of the human protein tyrosine phosphatase family. *Mol. Cell. Proteomics* **15**, 3030–3044 (2016).
59. Rasila, T. et al. Astroprincin (FAM171A1, C10orf38): a regulator of human cell shape and invasive growth. *Am. J. Pathol.* **189**, 177–189 (2019).
60. Monticone, M. et al. The nuclear genes *Mtfr1* and *Dufd1* regulate mitochondrial dynamic and cellular respiration. *J. Cell. Physiol.* **225**, 767–776 (2010).

Acknowledgements We thank Z.-Y. Lin for profiling PIK3R1 and members of the Gingras laboratory for discussion and advice throughout the project, Q. Morris and S. W. M. Eng for help with NMF, and J. Zhang and many cell biologists for suggestions about bait selection. Work in the Gingras laboratory was supported by a Canadian Institutes of Health Research (CIHR) Foundation Grant (FDN 143301). E.A.S. 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., J.D.R.K. and B.R. wrote the paper with input from 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.T.A. contributed the cell model and illustrations. C.J.W. helped with mass spectrometry data acquisition. A.R., H.A. and E.A.S. performed mitochondrial morphology experiments and analysed results. B.R. generated constructs and cell lines for BioID and testing predictions. G.G.H., K.T.A., J.-Y.Y., P.S.-T., H.Z., L.Y.Z., E.P., J.-P.L., D.R., E.C., S.W.T.C., L.P., B.R. and A.F.P. contributed constructs and cell lines. A.-C.G. supervised the project.

Competing interests The authors declare no competing interests.

Additional information

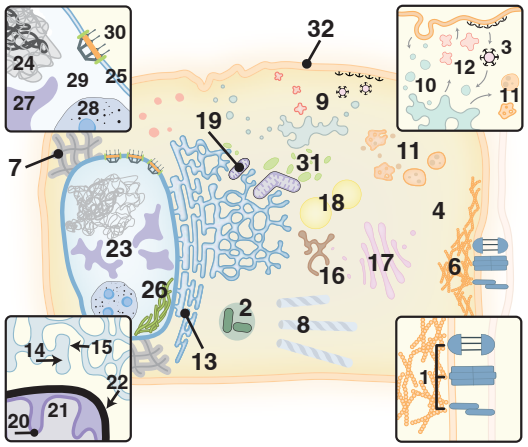
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03592-2.

Correspondence and requests for materials should be addressed to A.-C.G.

Peer review information *Nature* thanks Luca Scorrano and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

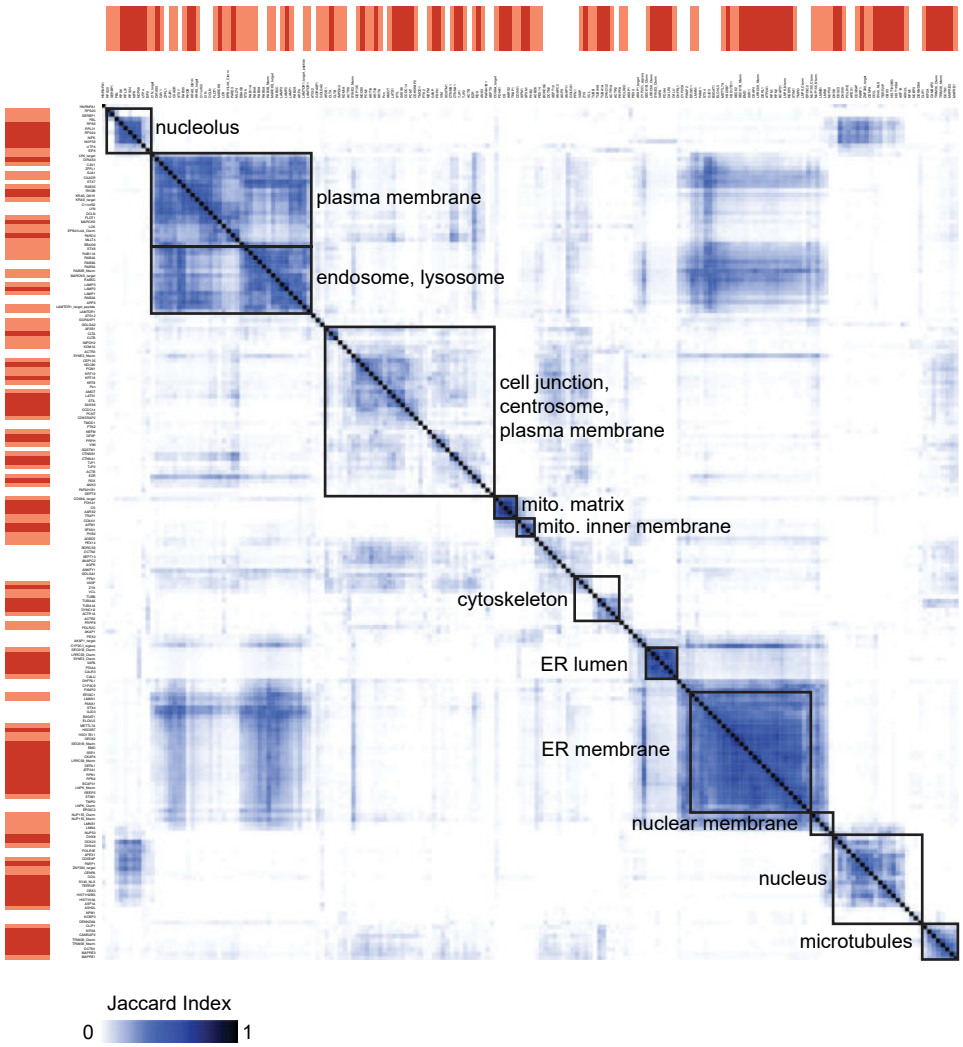
a



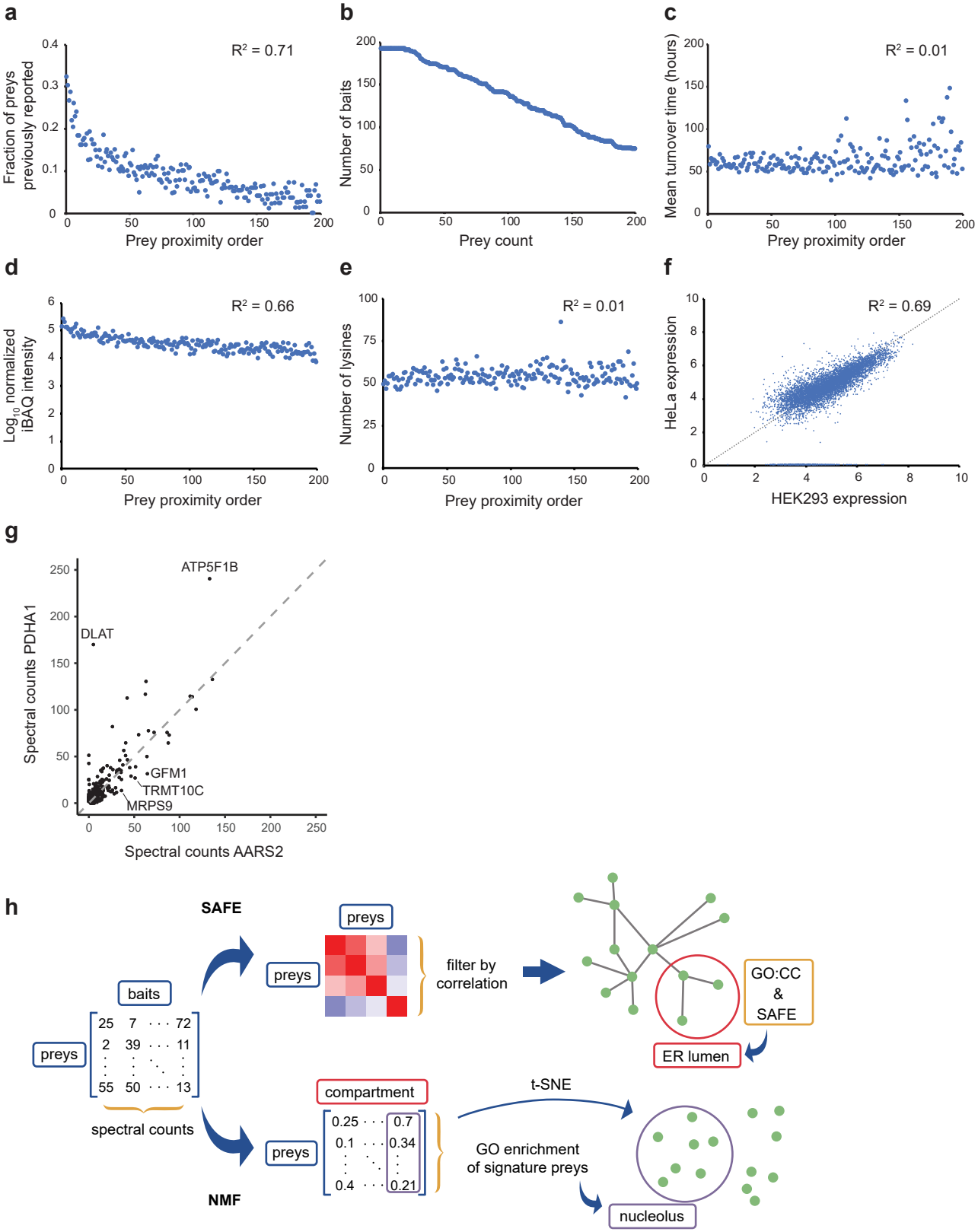
Legend *(# of baits)*

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



Extended Data Fig. 1 | Overview of the dataset. a, Cellular compartments targeted for profiling by BioID. Bold numbers on the schematic correspond to the indices in the legend. Italicized numbers in brackets next to the compartment name indicate the number of baits used to profile that compartment after quality control. **b**, Bait similarity and localization. The Jaccard index was calculated between each pair of baits in the core dataset using the list of high confidence (1% FDR) interactors. Baits were clustered using the Euclidean distance and complete linkage method, and clusters optimized using the CBA package in R. The colour 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 on the basis of the expected localization of the components.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Factors affecting prey labelling and rationale for prey-wise analysis. a, After sorting preys by proximity order and grouping by order across baits, the proportion of previously reported preys was calculated for the *n*th proximity order for $n = 1, 2, \ldots 200$. **b**–**f**, For each bait, the relative proximity of every prey (proximity order) was calculated from the control-subtracted length-adjusted spectral counts (CLSC) (Methods), such that the prey with the highest CLSC value was considered to be the 'interactor' most proximal to the bait and the lowest CLSC value the most distal. **b**, Number of baits with a minimum of *n* preys at a 1% FDR, for $n = 1, 2, \ldots 200$. **c**, Proximity order versus protein turnover rate (hours) in HeLa cells (turnover data are from ref. [55]). **d**, Proximity order versus protein expression as represented by the $\log_{10}$-normalized MS1 iBAQ intensity from ProteomicsDB[53]. **e**, Proximity order versus the number of lysine residues per protein. **f**, The $\log_{10}$-normalized MS1 iBAQ intensity of proteins expressed in HEK293 versus HeLa cells from ProteomicsDB[53]. The similarity in proteomes supports the usage of HeLa data in **c** as suitable HEK293 data was not available. Values along the *x* axis could reflect zero expression or missing data in HeLa cells. These were ignored when calculating the R^2 value. **g**, Bait comparisons for a pair of mitochondrial matrix proteins. Control-subtracted spectral counts are plotted for all high confidence preys (1% FDR) detected with either bait pair under comparison. AARS2 preferentially enriches components of the mitochondrial ribosome and proteins involved in translation, such as GFM1, MRPS9 and TRMT10C, whereas PDHA1 preferentially interacts with the pyruvate dehydrogenase complex component DLAT and the mitochondrial membrane ATP synthase ATP5F1B. **h**, Pipelines for localizing prey proteins using SAFE[40] and NMF[42]. In our SAFE pipeline, preys with a correlation across baits ≥ 0.65 are considered interactors and these pairs are used to generate a network that is annotated for GO:CC terms (Methods). In NMF, the bait–prey spectral counts matrix is reduced to a compartment-prey matrix and compartments are then defined using GO:CC for the compartment's most abundant preys. A 2D network is generated in parallel from the compartment–prey matrix using *t*-SNE[44].