nature methods

ARTICLES

https://doi.org/10.1038/s41592-019-0461-4

EPIC: software toolkit for elution profile-based inference of protein complexes

Lucas ZhongMing Hu[1,2,5], Florian Goebels[1,5], June H. Tan[1,2], Eric Wolf[1,2], Uros Kuzmanov[1], Cuihong Wan[1,4], Sadhna Phanse[1], Changjiang Xu[1], Mike Schertzberg[1], Andrew G. Fraser[1,2], Gary D. Bader[1,2*] and Andrew Emili[1,2,3*]

Protein complexes are key macromolecular machines of the cell, but their description remains incomplete. We and others previously reported an experimental strategy for global characterization of native protein assemblies based on chromatographic fractionation of biological extracts coupled to precision mass spectrometry analysis (chromatographic fractionation–mass spectrometry, CF–MS), but the resulting data are challenging to process and interpret. Here, we describe EPIC (elution profile-based inference of complexes), a software toolkit for automated scoring of large-scale CF–MS data to define high-confidence multi-component macromolecules from diverse biological specimens. As a case study, we used EPIC to map the global interactome of *Caenorhabditis elegans*, defining 612 putative worm protein complexes linked to diverse biological processes. These included novel subunits and assemblies unique to nematodes that we validated using orthogonal methods. The open source EPIC software is freely available as a Jupyter notebook packaged in a Docker container (https://hub.docker.com/r/baderlab/bio-epic/).

Systematic mapping of multi-protein complexes formed by protein–protein interactions (PPI) is critical to understand the mechanistic basis of cellular processes. Affinity purification–mass spectrometry (AP–MS)[1] is a powerful method for identifying such assemblies and has been applied widely[2–9], but is difficult to scale up or apply to non-model organisms. Biochemical CF–MS is a more efficient and flexible alternate strategy for examining native macromolecules on a global scale[10,11]. CF–MS is based on biophysical (typically chromatographic) co-purification of stable-associated proteins starting from cell-free mixtures (for example, tissue lysates). However, sophisticated data processing is needed to define genuine interactions, which can be challenging to implement.

To facilitate such studies, we have developed a simplified, standardized and fully automated CF–MS data analysis software toolkit, EPIC, which enables routine scoring and interpretation of large-scale CF–MS data regardless of sample source. Using supervised machine-learning EPIC integrates experimentally derived CF profiles and complementary functional evidence from public databases to create probabilistic PPI networks, which are then clustered to define high-confidence complexes.

We demonstrate the use and performance of EPIC by applying it to the nematode *C. elegans*. By analyzing quantitative mass spectra generated for whole organism soluble protein extracts resolved by ion-exchange chromatography, we identified 612 putative complexes from a network of 16,098 high-confidence PPIs that encompassed 3,855 worm proteins, most of which have never been reported before. The resulting 'WormMap' reveals assemblies with links to disparate lineage-restricted processes, conserved animal systems and human disease. To facilitate community adoption of CF–MS workflows, the EPIC toolkit is freely available as a Jupyter notebook packaged in a Docker container.

Results

Systematic scoring of PPI networks and identification of native multi-protein complexes. CF–MS is based on extensive experimental separation of native macromolecular mixtures under non-denaturing conditions. While there is no universally optimal protocol, ion-exchange–high-performance liquid chromatography (IEX–HPLC) is efficient at resolving stable endogenous complexes (Fig. 1a). To maximize coverage, non-ionic detergents can be added to solubilize hydrophobic complexes[9], chemical cross-linkers can be used to stabilize labile assemblies[12] and organelle compartments can be enriched before HPLC. For example, bead-based pre-fractionation (see Methods) improves detection of less abundant macromolecules (Supplementary Fig. 1), while concomitantly reducing 'chance' co-elution (that is, co-fractionation of functionally unrelated proteins). The results from a CF–MS experiment can be summarized as a matrix of biochemical fractions versus protein identities containing MS-derived protein amounts for each fraction (for example, summed precursor ion intensities or spectral counts).

EPIC software environment. EPIC employs python scripts to score CF–MS data, with modules to (1) process protein co-elution profiles, (2) optionally download supporting functional association information from public databases (CORUM[13], UniProt[14], IntAct[15], Gene Ontology (GO)[16], GeneMANIA[17], STRING[18] and InParanoid[19]), (3) predict and benchmark predicted associations versus curated reference assemblies (CORUM, IntAct and GO, Supplementary Fig. 2) and (4) cluster and visualize the resulting PPI network using Cytoscape[20] (Fig. 1b). Given suitable experimental CF–MS data and a standard taxonomy identifier for the organism under study, the software collects required information from online sources and

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3]Departments of Biochemistry and Biology, Boston University, Boston, MA, USA. [4]Present address: School of Life Science, Central China Normal University, Wuhan, China. [5]These authors contributed equally: Lucas ZhongMing Hu, Florian Goebels.
*e-mail: gary.bader@utoronto.ca; aemili@bu.edu

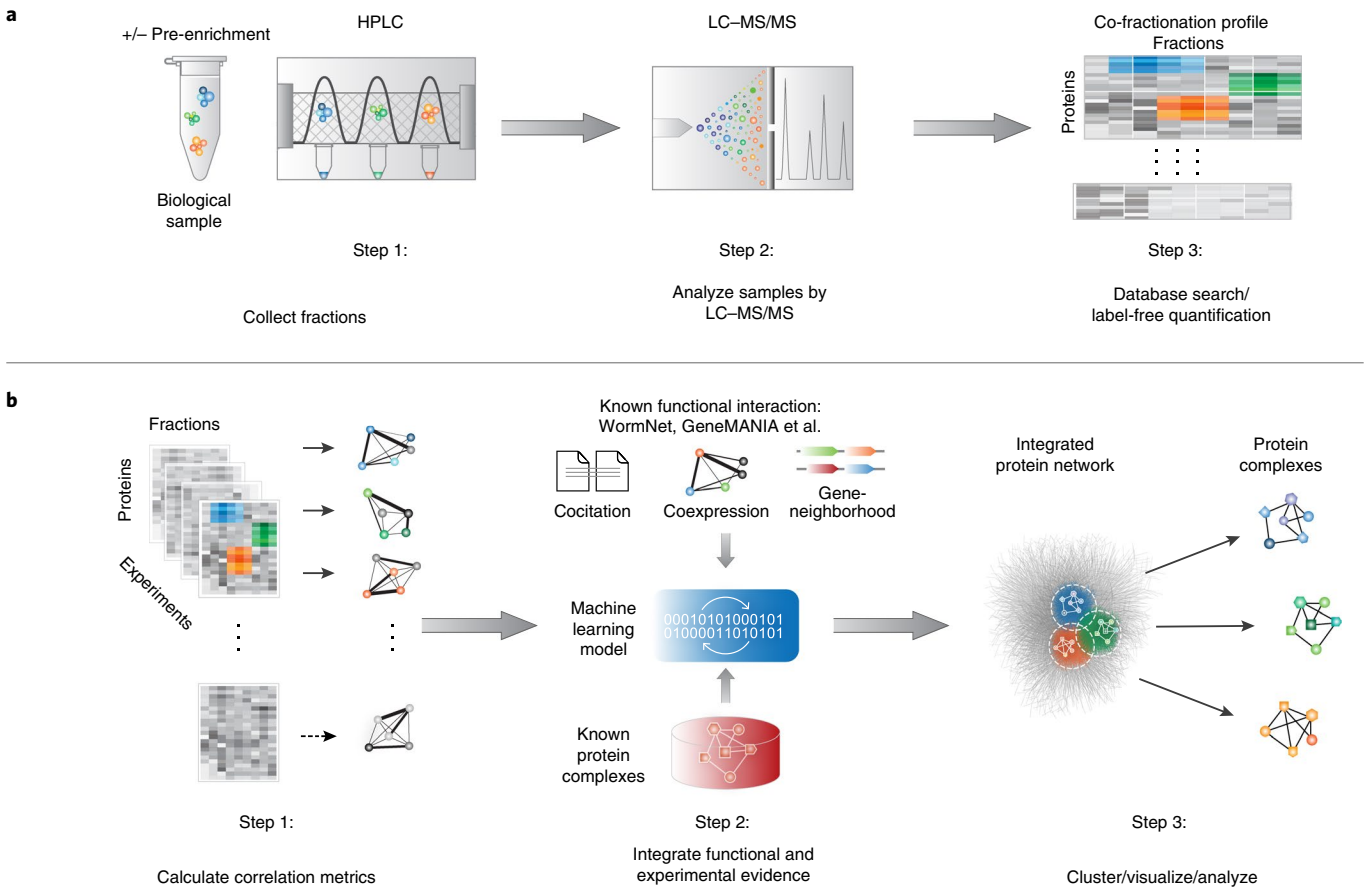

Fig. 1 | EPIC workflow. a, CF–MS experiments have three main steps: biochemical fractionation, MS analysis and protein profile scoring. **b**, Automated computational analysis using EPIC takes CF–MS data as input and consists of three main steps: (1) calculation of co-elution profile similarity using correlation metrics; (2) co-complex PPI scoring using machine-learning-based integration of experimental and functional evidence and (3) prediction, clustering and benchmarking of derived complexes. LC–MS/MS, liquid chromatography–tandem MS.

automates all data processing from raw data scoring to visualizing the results. In addition to convenient automation, EPIC outperforms an existing computational tool[21] both in terms of prediction quality and quantity (Supplementary Table 1, see Methods) and can process both isotope-labeled and label-free CF–MS data as input.

Since stably associated components within a complex are expected to co-fractionate together, EPIC first computes pairwise protein profile similarity using up to eight correlation metrics (Euclidean, Jaccard, Apex, Pearson, Pearson with Poisson noise, weighted cross correlation (WCC), mutual information (MI) and Bayes correlation[22]) that emphasize different profile features (Supplementary Notes). Positive (known) and negative reference co-complex PPIs display distinct correlation distributions (Supplementary Fig. 3). While it is likely not possible to predefine a universally optimal combination of correlation metrics for all possible CF–MS experiments, EPIC provides default parameters tuned on comprehensive CF–MS data (described below), and can optimize settings for any given data set. To reduce computational time, proteins observed in only one fraction and protein pairs with co-fractionation correlation scores less than 0.5 are removed (see Methods and Supplementary Fig. 4) before generating a scored co-complex PPI vector for each input experiment. Multiple correlation vectors are then combined and input into a supervised machine-learning model that is both trained to predict new PPIs and benchmarked against reference positive (annotated) PPIs (that is, co-complex relationships curated in the CORUM, IntAct and GO databases) and negatives (that is, combinations of proteins in distinct complexes).

To generate a comprehensive reference (gold standard) set for both training and benchmarking, EPIC retrieves species-specific complexes from the IntAct and GO complex databases. Since positive examples are limited for certain species, such as *C. elegans*, the benchmark is supplemented by mapping annotated human protein complexes from the CORUM database based on stringent one-to-one orthology (InParanoid). To minimize redundancy and bias, complexes with the majority of subunits in common (overlap score >0.8) are merged, while large assemblies with 50+ members (for example, ribosome) that could dominate learning are eliminated.

EPIC uses support vector machine and random forest classifiers by default, but other algorithms can be substituted programmatically. Since CF–MS data are often incomplete (for example due to proteome under-sampling) or noisy (for example, chance co-elution of unrelated proteins), EPIC can integrate additional supporting evidence (for example functional interactions inferred from co-expression, domain co-occurrence and co-citation) from public sources such as GeneMANIA or STRING, thereby producing richer and more accurate interaction networks. To avoid circularity, functional interactions based on published PPIs are excluded. To ensure all complexes have CF–MS experimental support, those complexes inferred based solely on functional evidence are removed. Prediction performance is evaluated by two-fold cross validation (that is, against an independent 'holdout' set of reference protein complexes, see Methods).

Finally, EPIC applies network-partitioning to define complex membership. ClusterONE[23] is used by default, although other

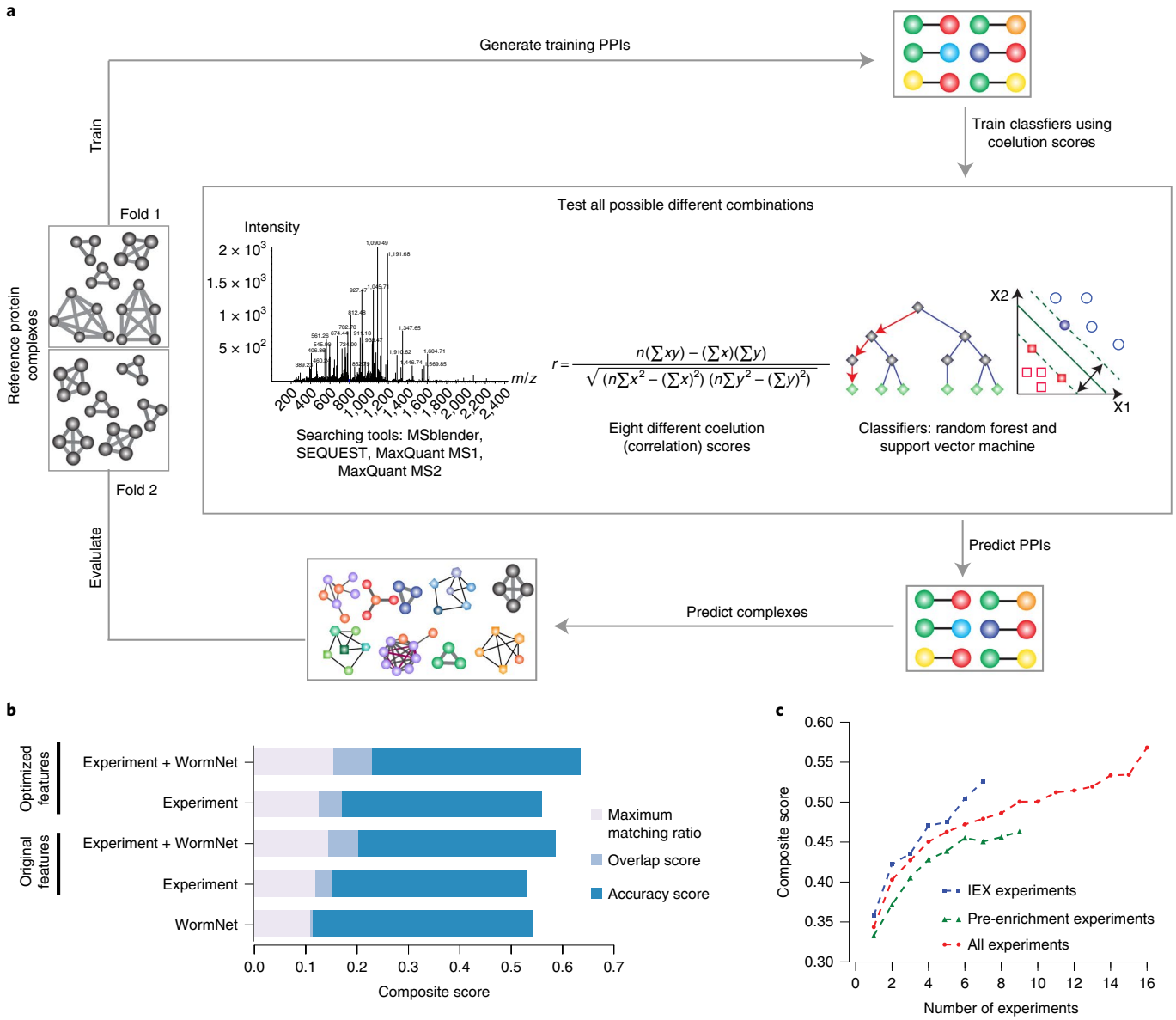

Fig. 2 | EPIC parameter evaluation. a, Computational procedures for protein interaction and co-complex prediction, driven by global optimization of classifier performance. The best combination of features was obtained using a nested cross-validation procedure (see main text). **b**, Bar chart shows predicted worm complex scores (maximal matching ratio, overlap and accuracy, the sum of which forms the composite score) using different combinations of experimental (CF–MS) data, functional evidence (WormNet) and correlation scores. 'Original features' indicates results from the set of correlation metrics (parameters) used in previous publications, and 'optimized features' indicates our newly optimized EPIC parameters. **c**, Average EPIC performance (composite score) based on including different number of co-fractionation experiments from IEX, pre-enrichment or the combined set of IEX and pre-enrichment experiments.

algorithms can be evaluated to optimize complex definition[24]. Each cluster is compared to annotated complexes curated in CORUM, GO and IntAct, and overall performance is measured by three complementary evaluation metrics (maximum matching ratio (MMR), accuracy and overlap score, see Methods), from which a single summary composite score is calculated to assign prediction quality[23].

Optimizing EPIC performance. We evaluated EPIC performance using a novel data set of 1,380 IEX–HPLC fractions generated for soluble worm protein extracts from mixed stage *C. elegans* cultures. co-eluting proteins were acid precipitated, alkylated and trypsin digested, and the resulting peptide mixtures analyzed by precision Orbitrap MS. To optimize major EPIC parameters (MS search tool, set of profile correlation metric and machine-learning classifier), we compared predicted complexes from each parameter setting (2,040 parameter combinations) against an independent benchmark of known complexes compiled from CORUM, IntAct and GO using composite score as the evaluation measure (Fig. 2a, see Methods). Optimized parameters substantially improved the resulting composite score compared to previously used parameters[10,11] (Fig. 2b). We evaluated the performance benefit of integrating functional interactions with the CF–MS data, again based on composite score, and found that including GeneMANIA, STRING or WormNet[25] clearly boosted performance (Fig. 2b and Supplementary Fig. 5). Functional evidence was not effective when used alone as input to predict complexes (Supplementary Fig. 5 and Supplementary Fig. 6). Since CF–MS studies consume considerable resources (for example, liquid chromatography–mass spectrometry run time), we

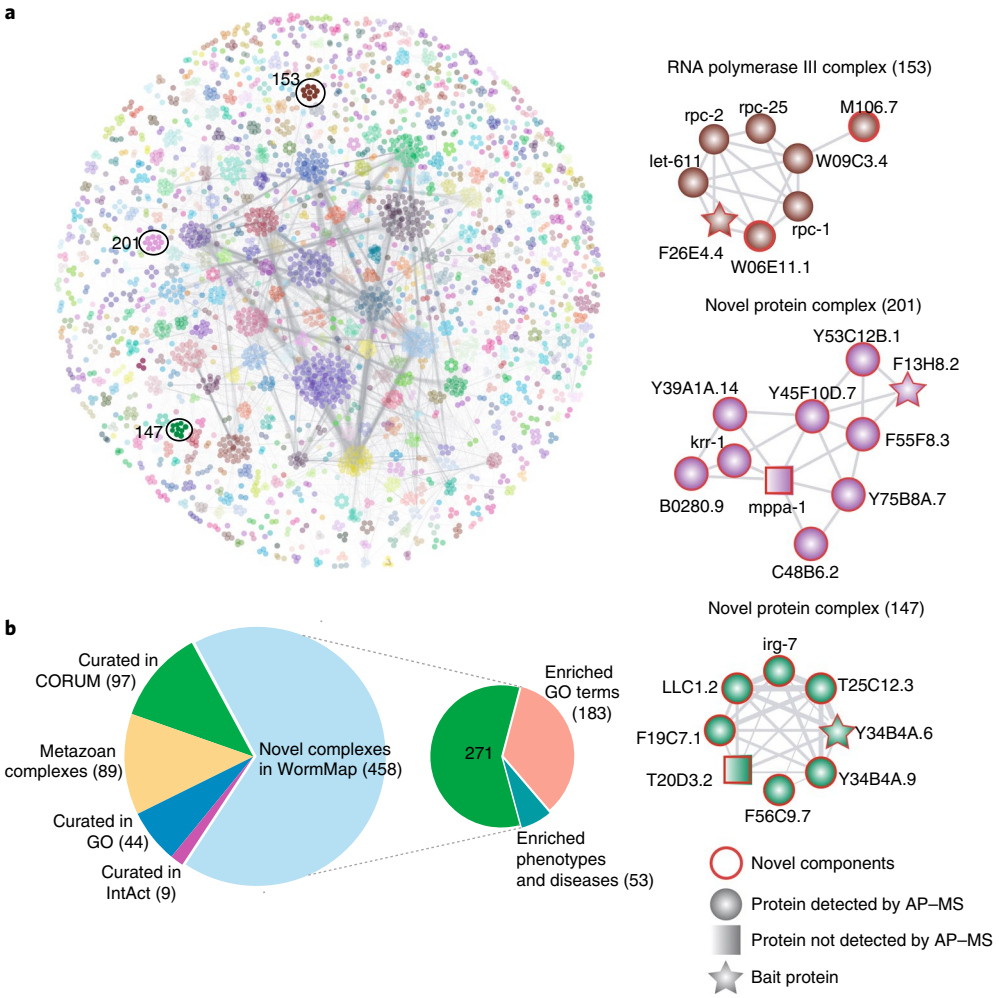

Fig. 3 | Prediction, benchmarking and analysis of *C. elegans* protein complexes. a, EPIC-derived WormMap. The left side shows the global overview of WormMap. Complexes validated using AP–MS are circled and AP–MS results are shown on the right, including novel components of the RNA polymerase III complex, as well as two novel complexes. Protein nodes are colored according to complex assignments, with novel assemblies and components highlighted with red circles. Gray lines between proteins indicate interactions that are supported by strong co-elution evidence. Bait proteins are shown as stars, prey proteins as circles and undetected proteins as squares. Novel components are indicated by a red node outline. The map is drawn using data available in Supplementary Tables 2 and 3; AP–MS spectral counts are summarized in Supplementary Table 4. **b**, Pie charts showing the overlap of predicted worm complexes found by EPIC with previously known macromolecules (from CORUM, GO, IntAct and the metazoan protein complex map[11]) and enrichment of putative novel assemblies for select biological function (GO terms), phenotype and/or disease associations (data available in Supplementary Tables 5, 6 and 7).

used EPIC to explore the 'cost/benefit' ratio of repeat biochemical fractionations by evaluating the relationship between prediction accuracy and the number of experiments performed. We calculated the average composite score by randomly sampling different numbers of co-fractionation experiments. Notably, while performance steadily improved as more data were acquired, prediction performance grew fastest over the first 2–4 separations (Fig. 2c, see Methods for details), suggesting an efficient lower bound (that is, ~4 IEX–HPLC experiments) for study design.

WormMap—a comprehensive map of soluble protein complexes in *C. elegans*. Using all 16 *C. elegans* co-fractionation experiments with optimized parameter settings and including functional interactions, EPIC predicted 16,098 high-confidence co-complex PPIs among 3,855 worm proteins (~25% of the nematode proteome), each directly supported by CF–MS data (at least one co-elution correlation score >0.5). Most (13,547) of these PPIs have not been reported before (compared to iRefWeb[26], BioGRID[27] or our previously generated Metazoan Complex Map[11]) (Supplementary Fig. 7, see Supplementary Table 2 for complete listing). Partitioning the network using ClusterONE predicts 612 complexes (Fig. 3a) of which only 150 map to known assemblies in CORUM, GO and IntAct. Most of the novel complexes appear to be clade-specific as only 89 are also found in the Metazoan Complex Map (see Supplementary Table 3 for complete listing).

We used multiple independent approaches to assess the accuracy of the predicted worm protein complexes. Experimentally, we used an established, orthogonal biochemical approach (AP–MS, see Methods) to validate both entirely novel assemblies as well as previously reported assemblies for which EPIC predicted unexpected new components (Fig. 3a and Supplementary Table 4). For example, we verified three new nematode-specific components (*F26E4.4*, *W06E11.1* and *M106.7*) of the worm RNA polymerase III machinery, one of which (*M106.7*) has DNA and nucleotide binding activity[28] (Fig. 3a). We also validated *unc-15* as part of a large myosin complex, an association not reported in a public database or our training set, but has been observed in previous work[29]. Likewise, we verified a predicted novel ten-member complex (Fig. 3a), for which most

components have limited functional annotation in WormBase[30], suggesting an overlooked biological role. Two of the subunits (*B0280.9* and *krr-1*) are orthologs of human small-subunit processome components involved in ribosomal biogenesis, suggesting a related function in nematodes. Another subunit, *Y45F10D.7*, is an ortholog of human *WDR36*, which is linked to primary open-angle glaucoma type 1G (GLC1G)[31], potentially providing a mechanistic connection. We also confirmed another putative novel complex with eight protein components (Fig. 3a) containing mostly uncharacterized components according to UniProt[14] and WormBase[30]. *Irg-7* is the only annotated subunit, with links to innate immunity and expression in the intestine[32], suggesting a potential role in the host response to pathogens. Some interacting proteins identified by AP–MS with low counts, indicating a weak MS signal, were nonetheless consistent with co-elution evidence (Supplementary Table 4).

To assess the physiological significance of the putative worm assemblies, we analyzed the network of complexes for coherent biological functions (based on GO annotations, Supplementary Table 5), mutant phenotypes (based on information from WormBase[30], Supplementary Table 6) or disease associations (based on orthology to human proteins in genetic disorder databases such as OMIM[33] and HGMD[34], Supplementary Table 7). Almost half of the novel complexes in WormMap were enriched for associations to essential processes, phenotypes or diseases (Fig. 3b). For example, knockdown of components of dozens of complexes either cause embryonic lethality or sterility, and have links to cancer in humans, reinforcing the use of EPIC for gaining fundamental mechanistic insight into large CF–MS data.

Discussion

Current knowledge of the physical networks of cells and tissues remains limited for many species, particularly non-traditional animal models. The majority of known/curated protein assemblies are annotated to mammals, whereas inference based on homology may not be the ideal for more distant organisms. CF–MS is an ideal experimental technology to address this, as it can be applied directly to any biological sample. However, CF–MS data are complex and challenging to process. We have developed the EPIC software to facilitate routine CF–MS analysis of native macromolecular assemblies in diverse contexts. EPIC provides optimized computational workflows, does not require expert computational skills to run, automates the entire data analysis process and is applicable to diverse model systems. We used EPIC to map protein complexes in *C. elegans*, which has classically been studied using genetic methods, thereby revealing nematode-specific biochemical network adaptations. In practice, EPIC enables users to process their own data and supply their own manually curated reference protein complexes to optimize classifier training.

We have shown that EPIC predicts complexes with high accuracy, particularly if four or more biochemical separations are available. While transient or unstable macromolecules may not be efficiently detected by CF–MS, chemical cross-linking can potentially be beneficial[12], while other gentle separation techniques, such as isoelectric focusing (R. Pourhaghighi et al., submitted) and size-exclusion chromatography[35], can provide complementary data. Regardless, to mitigate the false-discovery rate (FDR), EPIC implements customizable data filtering procedures and can optionally integrate supporting independent functional evidence. Integrating functional evidence will reduce false negative PPIs, but may introduce bias toward well-studied proteins[36]. While it is difficult to evaluate this bias, we note that many WormMap complexes, including those validated by AP–MS, contain uncharacterized proteins or proteins with diverse functional annotations, which suggests that EPIC is not strongly affected by this bias.

EPIC is both open source (https://github.com/BaderLab/EPIC) and compatible with disparate proteomic sampling techniques, including 'top-down' analysis of intact proteins[37] and sample multiplexing (isotopic labeling)[38] to map differential networks across conditions[39]. To facilitate broader uptake, we provide an automatically executable Jupyter-based notebook along with a Docker container (https://hub.docker.com/r/baderlab/bio-epic/) encompassing all necessary scripts and packages, enabling easy installation, deployment and optimization on any operating system. The distributed version of EPIC has step-by-step instructions and a user-friendly interface that enables uploading of local user defined CF–MS data files and the graphical display of results.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0461-4.

Received: 16 February 2018; Accepted: 15 May 2019;
Published online: 15 July 2019

References

1. Rigaut, G. et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
2. Krogan, N. J. et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
3. Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
4. Ho, Y. et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
5. Gavin, A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
6. Hu, P. et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96 (2009).
7. Huttlin, E. L. et al. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
8. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
9. Babu, M. et al. Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* **489**, 585–589 (2012).
10. Havugimana, P. C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
11. Wan, C. et al. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).
12. Liu, F., Rijkers, D. T., Post, H. & Heck, A. J. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Meth.* **12**, 1179–1184 (2015).
13. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).
14. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
15. 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).
16. The Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
17. Zuberi, K. et al. GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* **41**, W115–W122 (2013).
18. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
19. Sonnhammer, E. L. & Ostlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–D239 (2015).
20. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
21. Stacey, R. G., Skinnider, M. A., Scott, N. E. & Foster, L. J. A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics* **18**, 457 (2017).
22. Sanchez-Taltavull, D., Ramachandran, P., Lau, N. & Perkins, T. J. Bayesian correlation analysis for sequence count data. *PloS ONE* **11**, e0163595 (2016).
23. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein–protein interaction networks. *Nat. Meth.* **9**, 471–472 (2012).
24. Wiwie, C., Baumbach, J. & Rottger, R. Comparing the performance of biomedical clustering methods. *Nat. Meth.* **12**, 1033–1038 (2015).
25. Cho, A. et al. WormNetv3: a network-assisted hypothesis-generating server for *Caenorhabditis elegans*. *Nucleic Acids Res.* **42**, W76–W82 (2014).

26. Turner, B. et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010**, baq023 (2010).
27. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
28. Mulder, N. J. et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318 (2003).
29. Kagawa, H., Gengyo, K., McLachlan, A. D., Brenner, S. & Karn, J. Paramyosin gene (unc-15) of *Caenorhabditis elegans*. Molecular cloning, nucleotide sequence and models for thick filament structure. *J. Mol. Biol.* **207**, 311–333 (1989).
30. Harris, T. W. et al. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38**, D463–D467 (2010).
31. Monemi, S. et al. Identification of a novel adult-onset primary open-angle glaucoma (POAG) gene on 5q22.1. *Hum. Mol. Genet* **14**, 725–733 (2005).
32. Yunger, E., Safra, M., Levi-Ferber, M., Haviv-Chesner, A. & Henis-Korenblit, S. Innate immunity mediated longevity and longevity induced by germ cell removal converge on the C-type lectin domain protein IRG-7. *PLoS Genet.* **13**, e1006577 (2017).
33. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
34. Stenson, P. D. et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
35. Olinares, P. D., Ponnala, L. & van Wijk, K. J. Megadalton complexes in the chloroplast stroma of *Arabidopsis thaliana* characterized by size exclusion chromatography, mass spectrometry, and hierarchical clustering. *Mol. Cell. Proteomics* **9**, 1594–1615 (2010).
36. Skinnider, M. A., Stacey, R. G. & Foster, L. J. Genomic data integration systematically biases interactome mapping. *PLoS Comput. Biol.* **14**, e1006474 (2018).
37. Tran, J. C. et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).
38. Werner, T. et al. Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal. Chem.* **86**, 3594–3601 (2014).
39. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).

Acknowledgements

This study was supported by a Foundation Grant (FDN no. 148399) from the Canadian Institute of Health Research (CIHR, to A.E.), and US National Institutes of Health grants (nos. P41 GM103504, GM070743 to G.D.B.) L.Z.M.H. was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Mass Spectrometry-Enabled Science and Engineering (MS-ESE) program. *C. elegans* strain access was supported by the NIH Office of Research Infrastructure Programs (P40 OD010440).

Author contributions

A.E. and G.D.B. conceived the project. L.Z.M.H. and F.G. wrote the software, performed computational analysis and wrote the manuscript. L.Z.M.H. and C.W. performed the co-fractionation experiments. J.H.T. performed the protein GFP tagging in *C. elegans* with assistance and guidance from M.S. and A.G.F. The AP–MS experiments were performed by E.W. with assistance from U.K. S.P. and C.X. provided technical support. G.D.B and A.E. supervised the study and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0461-4.

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

Correspondence and requests for materials should be addressed to G.D.B. or A.E.

Peer review information: Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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

Methods

Protein extract preparation. Mixed-staged N2 strain *C. elegans* (strains were obtained from the Caenorhabditis Genetics Center, CGC) were collected in M9 buffer (standard recipe[40]), and re-suspended into lysis buffer (50 mM HEPES pH 7.4, 1 mM $MgCl_2$, 1 mM EGTA, 100 mM KCl) plus protease inhibitor cocktail (Roche). Worms were lysed by three rounds of 10 s sonication on ice (Branson Sonifer 450, output 6.0, duty cycle 60%). Soluble protein lysate (~2 mg ml^{-1}) was collected by filter centrifugation (Ultrafree-MC-HV, 0.45 µm). Bradford assay was used to determine protein concentration.

Pre-enrichment before HPLC fractionation. Differential affinity capture beads (NuGel *PRO*spector; BSG) were used to pre-enrich the worm lysate according to the manufacturer's protocol. After removal of lipids and insoluble biomass, extract incubated with different reagent beads (*PRO*-A, *PRO*-B, *PRO*-C, *PRO*-L, *PRO*-N, *PRO*-R). The suspensions were mixed for 10 min at 4 °C, centrifuged using Spin-X filters, and the filtrate was collected as 'flow-through' fractions. Bound proteins were eluted with 200 µl elution buffer (0.2 M Tris, 0.5 M NaCl, pH 9.0). The buffer was exchanged for HPLC loading buffer by Zeba desalt spin column (Thermo) before HPLC fractionation.

HPLC separations. *C. elegans* lysate and affinity enriched eluates (plus flow-through fractions) were individually fractionated by ion-exchange liquid chromatography using a quaternary pump 1100 HPLC system (Agilent Technologies). Whole proteome lysate was resolved into 120 fractions on a PolyCATWAX mixed-bed ion-exchange column (200 × 4.6 mm internal diameter (i.d.), 12 µm, 1,500 A) over a 240 min salt gradient (0.15 to 1.5 M NaCl). Enriched eluates were separated on a PolyCATWAX mixed-bed ion-exchange column (200 × 4.6 mm i.d., 5 µm, 1,000 A) into 60 fractions using a 120 min salt gradient (0.15–1.5 M NaCl). The detailed protocol has been described previously[10].

Liquid chromatography–tandem MS analysis. Proteins from the HPLC fractions were acid precipitated, re-dissolved and digested by sequencing grade trypsin overnight at 37 °C. The resulting peptides were dried and solubilized in 5% formic acid. Data-dependent liquid chromatography–tandem MS was performed using a nano-flow HPLC System (EASY-nLC, Proxeon) coupled to an LTQ Orbitrap Velos Mass Spectrometer (Thermo Fisher). After loading onto a 2.5 cm C18 trap column (75 mm inner diameter) packed with 100A Luna 5u C18 beads (Phenomenex) using an auto-sampler, peptides were separated on a 10 cm analytical column (75 mm i.d.) packed with 2 mm Zorbax 80XDB C18 reverse phase beads (Agilent). A 60 min gradient consisting of 5–35% ACN in water (with 1% formic acid) was used to elute peptides. Electrospray ionization was performed using at 2.5 kV spray voltage, and the instrument was operated in a data-dependent mode (one full MS1 ion survey scan directing consecutive MS2 acquisition scans on the top ten most prominent precursor ions). Collison induced dissociation directed peptide fragmentation was performed by 35% normalized collision energy.

Protein identification and label-free quantification. Raw spectral files were converted into mzXML format using the ReAdW software. A canonical FASTA file for protein searching was downloaded from the UniProt database and appended with common contaminants and reverse decoy sequences to assess the FDR. The peptide-spectrum matches from three different searching engines (comet, MSGF+ and X!Tandem) were integrated probabilistically using MSblender[41], setting the FDR to less than 1% for peptide and protein identifications. Parameter settings and detailed search protocols are available online (http://www.marcottelab.org/index.php/MSblender). MaxQuant[42] (v.1.6.0.16) search was performed at a fragment ion mass tolerance of 20 parts per million (pp.), maximum missed cleavage of two and a 1% false-discovery level (controlled by target/decoy approach). SEQUEST (v.2.7) search was performed at 20 pp. fragment ion mass tolerance and one missed cleavage allowance. The STATQUEST[43] model was used to assign confidence scores to all putative matches of peptides and proteins and a FDR was controlled at 1% for all identifications.

Generating green fluorescent protein (GFP)-tagged worm strains for AP–MS. To create GFP-tagged proteins for AP–MS experiments, *C. elegans* strains were grown and maintained at 20 °C on nematode growth media plates seeded with *E. coli* strain OP50. Some strains (wild-type N2 and RW1596: *myo-3 (st386) stEx30 (myo-3p::GFP::myo-3+rol-6(su1006)*[44]) were ordered from the CGC (https://cgc.umn.edu/). Extra-chromosomal array strains containing a C-terminal GFP translational fusion construct of *F26E4.4*, *Y34B4A.6* and *F13H8.2* were also generated in this study. For instance, the open reading frame and 617 base pair promoter region of *F26E4.4* (ref. [45]) were amplified and cloned into the pPD95.75 vector (Fire Lab Vector Kit). The construct was then injected at 20 ng $µl^{-1}$ along with pRF4 as a co-injection marker. Roller positive F2 animals were isolated and imaged to confirm the GFP expression (*rol-6* was used as a co-injection marker). Mixed stage worms were harvested for AP–MS validation studies. All other GFP-tagged strains (*Y34B4A.6* and *F13H8.2*) were generated in a similar fashion.

AP–MS validation. AP was performed essentially as described[46] with minor modifications. Briefly, frozen cell pellets were re-suspended in high-salt NP-40 lysis buffer (10 mM Tris-HCl p H 8.0, 420 ml NaCl, 0.1% NP-40) with protease and phosphatase inhibitors (Roche). After three freeze-thaw cycles, each lysate was briefly sonicated, treated with nuclease (Thermo Scientific Cat. no. 88700), followed by centrifugation at 14,000 r.p.m. The resulting soluble protein extract was split for technical replicate purifications. Each lysate was incubated at 4 °C on a rotator with 1 µg of rabbit anti-GFP antibody (Thermo Scientific Cat. no. G10362) for 2 h, followed by incubation with 25 µl of Protein-G Dynabeads slurry for 1 h. The beads were washed twice with low-salt buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl) and bound proteins subsequently eluted (4×) with 1% ammonium hydroxide pH 11. Recovered protein samples were dried, re-suspended in 50 mM ammonium bicarbonate, reduced with 5 mM DTT at 56 °C for 45 min and alkylated with 10 mM iodoacetamide at room temperature for 45 min in the dark. Trypsin digestion was performed overnight at 37 °C. Peptide samples were de-salted and re-suspended in 1% formic acid and then analyzed by data-dependent (top-15 MS2) acquisition on a Q Exactive HF mass spectrometer (Thermo Scientific) using a 90-min gradient on the same HPLC system described above. The resulting MS spectra were searched with MSblender.

EPIC computational workflow. In the following sections, we describe the computational components of the EPIC workflow that use machine-learning methods with the goal of identifying as many interacting proteins as possible, while minimizing the 'chance co-elution' problem, based on MS-based protein profiles (Supplementary Fig. 8). For each fractionation experiment, search results are summarized into a single data matrix, in which, each row represents an identified protein while each column value refers to an estimated relative protein amount (spectral count or summed MS1 ion intensity) for a corresponding fraction. Under the assumption that proteins not detected by mass spectrometry are likely to be low abundant or simply not expressed, the values of missing proteins are set to zero. In the case of multiple co-fractionation experiments, a single unified matrix is created. For classification, EPIC has two major steps: first, a training set of co-complex PPIs is derived from reference protein complex datasets (for example, CORUM) that map onto the experimental data. Second, one of two built-in machine-learning algorithms (support vector machine, random forest) is used to define a probabilistic interaction network from which protein complexes are inferred using a network-partitioning algorithm.

Data processing. Several steps are required to pre-process the raw mass spectrometry co-elution table to improve the quality of the predicted network, similar to those performed in previous work[11]. However, we have added new features to improve prediction quality and to reduce the computational runtime.

Removing 'one-hit-wonders'. The central principle of EPIC is based on the guilt-by-association approach, which posits that proteins that are physically associated tend to elute at the same time. However, to meaningfully evaluate fractionation data, EPIC requires the proteins to be present across multiple biochemical fractions in the same experiment. Thus, proteins measured in exactly only one fraction are deemed 'one-hit-wonders' and removed from further analysis. The reason for discarding such proteins is not because we assume they were falsely measured, but rather that EPIC measures co-elution profile similarities based on correlation metrics that evaluate similarity over the entire elution profile, which is not effective for singletons. Some proteins may be identified in only one fraction in multiple experiments. However, if we predict PPIs in this way, overall performance is markedly decreased (data not shown). Hence, each experiment is processed individually in EPIC, followed by merging or concatenating all the resulting co-elution correlation metric scores into a single unified matrix for machine learning. From the initial raw MS data, we observed that MSblender is highly sensitive and identifies the largest number of peptides of which many are one-hit-wonders. However, even after removing one-hit-wonders, MSblender still has the largest number of identified peptides compared with single search engines, resulting in the highest predicted quality protein complexes (Supplementary Fig. 9, see main text).

Elution data normalization. Before calculating correlation coefficient metrics, the protein elution profile matrix is normalized column-wise to correct for slight sample injection variation. The protein elution profile matrix for each co-fractionation experiment consists of MS1 ion intensity or MS2 spectral counts for M proteins across N fractions. Thus, before calculating protein elution profile similarities, the raw data of each protein in each fraction are normalized by dividing the amount of the particular protein (either MS1 ion intensity or MS2 spectral counts) by the total amount of proteins in corresponding fractions. So, given a protein elution matrix A of the size $M \times N$, where each $A_{i,j}$ denotes the value of MS1 intensity or MS2 spectral counts of a particular protein i in fraction j, the column-wise normalized protein elution profile matrix $B_{i,j}$ is calculated as:

$$B_{i,j} = \frac{A_{i,j}}{\sum_i A_{i,j}}$$

Some similarity score metrics (that is, Euclidean distance score) require row-wise normalization after column-wise normalization to make sure the sum of

each row equal to one. So the final normalized protein elution profile matrix $C_{i,j}$ is calculated as:

$$C_{i,j} = \frac{B_{i,j}}{\sum_j B_{i,j}}$$

Creating candidate protein pairs. In previous work[11], we first created all possible pairs of proteins for each experiment, followed by calculating their corresponding co-elution scores and then removed all protein pairs without co-elution correlation scores equal or more than 0.5. However, this approach is computationally demanding and requires high-performance computational resources to perform all calculations in a reasonable amount of time. Thus, we decided to apply a pre-filtering step: instead of calculating all possible protein pairs for each experiment we first generate a super-set of all possible protein pairs across all experiments and remove those pairs for which the two proteins do not overlap (never occur in same fraction across all experiments). Usually, this filtering step removes a substantial (up to 60%) of possible candidate pairs, significantly reducing computational time. In the subsequent step, we calculate co-elution scores for each candidate protein pair across each experiment and then summarize the results into matrices, and then we remove all protein pairs whose co-elution score is below 0.5 across all experiments.

Similarity metrics. Proteins that belong to the same protein complex should co-elute in the same or adjacent fractions, and thus should have similar elution profiles. In EPIC, we deploy several methods for measuring the similarity of two protein elution profiles. We treat each elution profile as a vector consisting of the observed MS2 spectral counts or MS1 ion intensities for a particular protein across the corresponding biochemical fractions, and a complete co-fractionation experiment is stored as a matrix where rows and columns represent proteins and fractions, respectively. To measure the co-elution profile similarity between two proteins, we employ various correlation metrics that range from simple scores, such as Euclidean distance, to more sophisticated metrics based on information theory. Some co-elution scores use normalized data $B_{i,j}$ while some use raw data $A_{i,j}$. In the following formulas: p_a and p_b denote protein a and protein b in the same co-fractionation experiment, M denotes the total number of proteins and N the total number of fractions.

Euclidean distance. Euclidian distance denotes the distance between two vectors (or two points) in a high-dimensional space (also known as 2-norm). The two points, for which the distance is calculated, represent a protein pair while the number of fractions is the dimension of space that the Euclidean theorem applies to. This Euclidean distance feature uses normalized counts and lies between 0 and 1, where identical elution profiles have a distance of 0 and elution profiles that differ greatly have a distance closer to 1.

Jaccard score. Jaccard score computes the ratio of how often proteins elute in the same fractions and how often proteins are detected in all fractions. Thus, the Jaccard score between two proteins is calculated by counting the number of fractions that contain both proteins and dividing by the number of fractions that have at least one of the two proteins. The formula is as follows:

$$\text{Jaccard}(p_a, p_b) = \frac{|\{\text{no.}p_a > 0\} \cap \{\text{no.}p_b > 0\}|}{|\{\text{no.}p_a > 0\} \cup \{\text{no.}p_b > 0\}|}$$

Bayes correlation. We integrated a novel method that uses a Bayesian probabilistic framework for calculating correlation scores between two MS2 spectral counts-based vectors. Originally, this method was proposed[22] to process RNA-seq gene expression data that are based on sequence counts for various genes under different conditions. Here, we applied the same method for peptide counts for various proteins across the biochemical fractions. The main advantage of Bayesian statistics over Pearson correlation is that it considers both measured signal magnitudes and associated uncertainties in those magnitudes. Thus, Bayesian correlation will return high correlation values if measurement confidence is high and prevents high correlation values when the measurement confidence is low. To integrate Bayesian correlation, we integrated a public R script (https://www.perkinslab.ca/software#h.p_ORAttguSX-rx) into our python pipeline using the rpy python package that allows the import of R code into python. Bayesian correlation calculation scores support three different assumptions of how the priors distributed: uniform, Dirichlet-marginalized and zero count-motivated. We used zero count for this work, as it performed best (Supplementary Fig. 11).

Apex score. Most proteins tend to elute with a specific retention time, and thus the fraction that contains the largest amount of a particular protein is typically also the most critical fraction for that protein. Thus, two proteins are considered to be more likely to interact with each other if the fractions having the largest recorded amount across all fractions are the same. Based on this premise, previous co-fractionation experiments introduced the apex score[10], which scores protein co-elution profiles highly if their respective peak fractions are the same (apex score = 1) or else penalizes them (apex score = 0).

Pearson correlation coefficient (PCC). The Pearson correlation is used to measure the similarity of two protein co-elution profiles. To calculate PCC, we used the scipy package in python. PCC was calculated by using the vector of raw peptide counts or intensities obtained for each protein. From experience, PCC works well for proteins with high signal but not well for proteins with low peptide counts. Nevertheless, we decided to integrate this correlation metric into EPIC as it is a frequently used similarity metric, thus is also useful for benchmarking and evaluating other correlation metrics.

PCC plus noise (PCCN). The PCC is relatively good at determining protein co-elution based on normalized protein elution profiles. However, proteins with low signals (low MS2 values) are more likely to co-elute by chance. To avoid this issue, the PCCN metric introduces a low level of random artificial signal on the raw co-elution data in the form of Poisson noise to each protein across all fractions, followed by co-elution matrix normalization and co-elution score calculation via Pearson correlation. This process is repeated n times, and the resulting PCCN score is the average of those n runs. The same strategy has been used in creating previous co-elution networks[10,11], but here we systematically investigated the iteration parameter n.

WCC. One of the issues of detecting eluting protein complexes from a liquid chromatography-based system is that the component subunits might show some residual retention time shifts. Unlike PCC, WCC considers this small variance between otherwise similar co-elution profiles. To avoid promiscuity, stringent parameters are used to tolerate a small shift of roughly only one fraction when comparing two proteins. The WCC calculation is performed using the wccsom R package[47], which we integrated into our python pipeline using the rpy2 python R interface package. WCC similarity is measured between 0 and 1.

MI. MI considers both linear and nonlinear dependencies between vectors. The initial step in calculating MI is to binarize the spectral count vector elements into 'with protein' and 'without protein', since MI measures statistical dependence between the two given proteins based on their relative co-elution frequency (percentage co-eluted fractions) and each protein's individual relative frequency (percentage fractions containing the respective protein). The elution matrix was binarized by temporarily changing each protein spectral count to 1 (if there were spectral counts observed in the fraction) or to 0 (if not present). Thus, $P(p_a = 1)$ denotes the individual relative frequency of p_a, which is calculated by dividing the total number of fractions with value 1 for protein p_a by the total number of fractions in the corresponding co-fractionation experiment, whereas the joint relative co-elution frequency of protein p_a and p_b named $P(p_a = 1, p_b = 1)$ is calculated by counting the total number of fractions that contain both p_a and p_b and dividing this number by the total number of fractions. MI is calculated as follows:

$$\text{MI}(p_a, p_b) = H(p_a, p_b) - H(p_a) - H(p_b)$$

In the formula above, $H(p_a)$ denotes the entropy of protein a and $H(p_a, p_b)$ the joint entropy with following formulas:

$$H(p_a) = -\sum_i^{\{0,1\}} P(p_a = i) \times \log_2(P(p_a = i))$$

$$H(p_a, p_b) = -\sum_j^{\{0,1\}} \sum_i^{\{0,1\}} P(p_a = i, p_b = j) \times \log_2 (P(p_a = i, p_b = j))$$

Protein interaction prediction metrics. We use different measurements to evaluate EPIC performance based on its capabilities of predicting both PPIs and multi-protein complexes. Most of the evaluation metrics that we applied for measuring how well EPIC can predict PPIs are commonly used throughout the machine-learning field and are briefly mentioned in this section.

One first needs to define criteria of what is true for a predicted interaction (Supplementary Table 8). With EPIC, this is done by comparing the predicted interactions to the above-mentioned generated reference data set of positive and negative protein interactions. Based on this concept, one defines precision, recall and F-measure (also known as $F1$ score) using true positives (TP), false positives (FP) and false negatives (FN) as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F\text{-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Additionally, we evaluate performance using the precision-recall (PR) and the receiver-operating characteristic (ROC) curve. We use the area under the PR curve (auPR), and the area under the ROC curve (auROC) to give single value performance, which can be used to compare different methods or parameter settings.

PR curve. The PR curve is created by first sorting the list of predicted protein interactions by their confidence scores and then iteratively removing the top element from that list while calculating the resulting precision and recall value for the updated list. The PR curve is the line that results by plotting those generated PR values. This line shows the trade-off between precision and recall, and area under the PR curve measures the average precision of the classifier. It can be used to compare multiple models, since a better classifier will lead to a higher PR curve and thus results in a larger auPR value.

ROC curve. The ROC is generated analogously to the PR curve, but instead of plotting the resulting precision and recall values, the ROC plots true-positive rate against the false-positive rate. The auROC curve describes the probability of the classifier of scoring a positive interaction higher than a negative interaction, which means it shows how well the classifier can separate positive and negative PPIs. Thus, in a two-class problem, an auROC score of 0.5 means the classifier cannot differentiate between a positive interaction and a negative interaction, whereas a score of 1 means the classifier can perfectly predict the class labels.

Cluster prediction evaluation metrics. Training a classifier on a single instance object such as PPIs to determine whether or not a prediction is true is straightforward, as it only involves comparing the set of predicted PPIs against a set of pre-defined positive and negative protein interactions (see previous sections). However, in the case of predicting protein complexes that typically consist of three or more members, this comparison is more difficult. First, we describe a simple measurement for determining the precision of the predicted protein complexes based on the overlap of the predicted complexes to a given set of reference complexes. However, an important issue here is when one should consider two protein complexes as a match. Several protein complex prediction studies have investigated how to evaluate cluster overlap, and essentially all their measurements are based on how to evaluate the overlap between the set of proteins within complex A and the set of proteins within complex B. The overlap score between protein complexes are calculated as below (note that $|A|$ denotes the number of proteins in complex A):

$$\text{overlap}(A, B) = \frac{|A \cap B|^2}{|A| \times |B|}$$

It is suggested to consider two protein complexes to be matching when the overlap score between them is greater than 0.25, since two clusters of the same size would have this score if the intersection set is half of the complex size.

Additionally, we calculate prediction sensitivity, accuracy, positive predictive value and cluster separation[48]. For the following scores we consider $a_1,\ldots,a_i,\ldots,a_m$ predicted complexes thatwe compare to a set of $b_1,\ldots,b_j,\ldots,b_n$ reference complexes, and $T_{i,j}$ denotes the number of proteins that are found in both complex i and j.

Sensitivity (Sn) is the fraction of proteins in predicted complexes that are found in reference complexes.

$$\text{Sn} = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} t_{i,j}}{\sum_{i=1}^{n} |b_i|}$$

Positive predictive value (PPV) indicates how specific and complete the predicted complexes match the reference complexes. A score of 1 indicates that each predicted complex only overlaps exactly one reference complex, and a low score indicates low or redundant overlap with the reference.

$$\text{PPV} = \frac{\sum_{j=1}^{m} \max_{i=1}^{n} T_{i,j}}{\sum_{j=1}^{m} \sum_{i=1}^{n} T_{i,j}}$$

Accuracy (Acc) shows the trade-off between PPV and Sn.

$$\text{Acc} = \sqrt{\text{Sn} \times \text{PPV}}$$

MMR. The MMR was developed to cope with some of the limitations of the PPV. PPV tends to be lower if there is substantial overlap in the reference data[23], but those overlaps are common in biological data sets such as CORUM. Our merging step only removes highly overlapping clusters, but smaller overlaps are still present. Thus, even EPIC perfectly predicts the reference complexes it will not achieve a score of 1 for PPV and Sep (clustering-wise separation score suggested by Brohée and Van Helden[48]). MMR can cope with this problem:

$$\text{MMR} = \frac{\sum_{i=1}^{n} \max_{i=1}^{m} O(n_i, m_j)}{|\max_{i=1}^{n} O(n_i, m) > 0|}$$

As established by others[23], we summarize MMR, overlap score, and accuracy to create the composite score, and we consider the parameter combination with the highest composite score to be the best combination.

Machine-learning prediction. At the end of the data processing, EPIC generates a co-elution matrix, which contains rows for each protein pair and columns for each co-elution score across all co-fractionation experiments. In cases where a protein pair was not present in one of the experiments, we set all of its co-elution scores for the given experiment to zero. In the subsequent section, we describe how EPIC creates co-elution PPI network and the set of protein complexes:

Reference data set. Our goal is to make EPIC a generic tool for surveying protein complexes in different species. To facilitate standardization, we decided to use CORUM database[13] as the source of the gold standard set, as it is the largest manually curated protein complex database available. EPIC uses human protein complexes for generating the necessary reference data, since protein complex information is typically sparse for the majority of species and as CORUM itself mainly curates human protein complex information. EPIC automatically downloads the current CORUM version and retains only those complexes that are annotated for human or mammals. Further, only protein complexes defined based on biochemical approaches are retained in the reference dataset, as protein complexes defined based on non-biochemical methods might not be expected to co-elute by chromatographic separation.

As an added set, EPIC downloads all human protein complexes from the IntAct database, for which again only complexes detected by biochemical methods are retained.

Additionally, EPIC automatically downloads a set of curated protein complexes in the GO database, annotated based on biochemical evidence for relevant target species (for example, *C. elegans*).

We then generate an extracted set of positive and negative PPIs for both the training and holdout protein complexes, respectively. PPIs are defined as positive if they are observed in the same protein complex. If proteins exist in the protein complex dataset but never appear in the same protein complex, then these two proteins are defined as negative PPIs.

For mapping human proteins to the input species (test sample), we integrated the InParanoid database, which is also automatically downloaded for each EPIC run. We only consider one-to-one orthologous protein mappings between human and the test species with an InParanoid confidence score of 100%. In this manner, curated human protein complexes are projected on to corresponding orthologous protein complexes in a target species of interest. To avoid bias, protein complexes with fewer than three members and large assemblies with more than 50 proteins are removed, because these would dominate the machine-learning process. Further, to remove redundancy in our data set, highly overlapping protein complexes (high fraction of shared components) are merged. We evaluate the overlap of two complexes A and B as follows, where $|A|$ denotes the number of proteins in A:

$$\text{overlap}(A, B) = \frac{|A \cap B|^2}{|A| \times |B|}$$

Protein complexes are merged if they have an overlap score of at least 0.8. This automatic process for generation of reference data set currently only supports UniProt identifiers because they are used by GO, IntAct, InParanoid and CORUM.

Train machine-learning classifier and predict PPIs. The machine-learning classifier is trained on the sets of positive and negative PPIs as we defined before based on CORUM, IntAct and GO. We created the union of training set by merging the training set obtained from the above three databases, in which only the protein pairs have at least one elution profile similarity score larger than 0.5 (among all co-fractionation experiment and among all correlation metrics) are retained. We then trained the classifier on this reduced set of negative and positive interactions with correlation metrics scores from different co-fractionation experiments as input features. Because the classifier is trained to distinguish true-positive co-complex membership with high co-elution score from non-interacting protein pairs including false-positive chance co-elution associations that also have high co-elution scores, we decided to additionally integrate functional evidence data (that is, GeneMANIA, STRING and WormNet) into the machine-learning method. However, to reduce circular reasoning in the machine-learning step, functional evidence derived from 'physical interaction', 'protein complexes' and 'predicted interactions' was excluded from input features.

EPIC generates a set of PPIs using the classifier trained on experimental data with an option to include functional evidence. Then a set of PPIs are predicted by the classifier trained on experimental data or optionally experimental data integrated with functional data. A protein elution profile correlation score cut-off was applied here to ensure all PPIs have experimental evidence support. EPIC then applies a clustering algorithm (see below) to predict protein complexes based on the PPIs from the combined set of data. Novel protein complexes are identified by comparing the predicted set of complexes from above and the curated protein complexes from the major databases (CORUM, IntAct and GO) by setting a liberal overlap score cut-off at 0.25.

Predict protein complexes from the PPI network. In the final step, EPIC generates a set of putative complexes from the predicted protein interaction network. As with our previous work, we use the ClusterONE clustering method[23], since it has been shown to provide excellent performance among several different clustering algorithms for predicting protein complexes from PPI networks, thus we do not investigate clustering methods here.

Benchmarking. We extensively benchmarked EPIC and optimized parameters for each step of the EPIC pipeline on WormMap data. In an ideal scenario, we would evaluate the complete space of all possible parameters, however, the space for searching the optimal parameter configuration grows exponentially ($2^{|parameters|}$) with the number of parameters we want to configure. Thus, to make the benchmarking of EPIC feasible, we investigated only one parameter at a time while keeping the remaining parameters fixed. First, we will describe benchmarking statistics and evaluation criteria followed by the results of benchmarking.

Feature parameters. In this part, we evaluate the optimal parameter settings for co-elution scores (if any parameter setting is involved). From the total of eight correlation features, two of them have parameters to optimize: the prior used for the Bayes correlation and the number of noise iterations for PCCN. We evaluated those parameters based on how well they can predict PPIs (that is, precision, recall, F1, auROC, auPR). To be consistent, all the evaluations were performed using elution data generated by the MSblender search engine, as it is the search engine used in our previous publication that generated the largest data set with the most identified proteins. The results for number of noise iterations can be found in Supplementary Fig. 10 and we observed optimal scores obtained for five noise iterations. After analyzing the three possible Bayes priors, we observed no significant differences between the three different priors based on ROC and PR curves (Supplementary Fig. 11). However, if we analyze the evaluation metrics for predicted protein complexes, we see the best composite score for the zero-count prior (Bayes3) (Supplementary Table 9). Thus, we use the zero-count prior for EPIC.

EPIC parameter optimization by nested cross validation. It is not possible to provide globally optimal parameters for all data sets. In EPIC, we developed a nested cross-validation strategy to optimize parameters for our worm data and used the optimized set of parameters to generate our WormMap. As described in Fig. 2a, we first collected and merged all worm protein complexes from CORUM, GO and IntAct. We initially used *k*-means clustering and an overlap score as the measurement metric to divide the whole set of reference protein complexes set into two distinct sets of complexes. We then balanced the two sets while minimizing the overlap by pruning. The first half was used for training (based on our co-fractionation data) while the second half was used as the 'holdout' set for evaluation (two-fold cross validation at the protein complex level). In our study and in the EPIC software, we implemented two machine-learning classifiers, support for four protein searching/quantification tools and eight different correlation scores, which gave us 2,040 total parameter combinations. We trained machine-learning classifiers with our worm co-fractionation data to predict PPIs and protein complexes for each of the 2,040 different parameters combinations. The resulting 2,040 predicted protein complex sets were then benchmarked with the held out 'test' half of the curated protein complexes using composite score (see main text and above) as the evaluation metric. Random forest, in general, outperformed support vector machine for predicting protein complexes (Supplementary Fig. 12a). MSblender gives the best composite score compared with other protein search/quantification tools (Supplementary Fig. 12b). To get a relatively good prediction, at least three different correlation scores are required (Supplementary Fig. 12c). The optimized set of parameters (machine-learning classifier: random forest, protein searching/quantification tool: MSblender, correlation scores: MI, Bayes correlation, Euclidean distance, WCC and apex score) for generating WormMap is the combination that gives the highest composite score. Functional evidence data were then added to the matrix formed by the optimal set of correlation scores for predicting PPIs. Since extensive computational resources are required for this optimization, we performed this analysis on the SciNet supercomputing platform (https://www.scinethpc.ca/). We provide a parameter optimization function in the EPIC software and encourage users to optimize their parameters using their own data if a super computing resource is available, but otherwise, we recommend using the default EPIC parameters, which are the ones that were found optimal for WormMap using the above procedure.

Exploring the value of the additional experiments. After nested cross validation, the selected optimal correlation score combination and random forest machine-learning classifier was used for evaluating whether a pre-enrichment step improves protein complex prediction and what is the most economic way to perform experiments. We performed the analysis using data collected from pre-enrichment, non-pre-enrichment (IEX) and the combination of both (all experiments), individually. Similar to the step of nested cross validation, we benchmarked the predicted protein complexes using composite score, based on our two-fold cross-validation strategy. For each specific number of experiments, we considered all combinations and reported the average of the evaluating metrics. For example, for the first point in the plot indicating use of one experiment, we analyzed each of our seven IEX experiments individually to predict complexes, evaluated the composite score and then calculated the average of number predicted complexes and composite scores over the seven experiments. We observed a positive correlation between composite score and the number of experiments (Supplementary Fig. 13a). After five experiments, using IEX alone performed much better than using all experiments. Similarly, a sharp increase was observed for the last point of the 'all experiments' line (red line). We then asked whether the sharp increase of IEX performance was the result of sacrificing the coverage of predicted protein complexes. To balance the coverage of predicted protein complexes and composite score, we then plotted 'composite score × the number of predicted complexes' versus 'number of experiments' (Supplementary Fig. 13b). In this plot, we noticed the 'all experiments' line reached its stationary phase at nine experiments. We also noticed a dramatic decrease of the 'IEX' line at seven experiments, which shows that the sharp increase of composite score for 'IEX' is due to a decrease in the number of predicted protein complexes. Also, when using all 16 experiments, the composite score is maximized. Thus, the general guideline would be to use as many experiments as possible and that pre-enrichment will help protein complex prediction in terms of both composite score and coverage; however, if mass spectrometry time is limited, a reasonable lower bound is to run four IEX experiments.

Cut-off for correlation coefficient. We plotted the histogram of maximal correlation scores for all positive PPIs among all seven different correlation coefficients (apex score is not included, since it is either 0 or 1) across all experiments performed (Supplementary Fig. 13a). We noticed there is a clear cut-off at 0.5, which suggests we could retain protein pairs with a coelution correlation score over 0.5 for machine-learning prediction, as pairs without any coelution score over 0.5 are not likely to be positive interactions.

Comparison of EPIC with PrInCE. To objectively compare the performance of the two tools, we downloaded the example SILAC co-fractionation data available from the PrInCE website (condition1.csv and condition2.csv) and used this as input data to predict protein complexes using both PrInCE and EPIC. We then compared the results (predicted complexes) with a benchmark set of reference assemblies (that is CORUM) using the multifactor composite score as the stringent evaluation metric. The resulting set of protein complexes predicted by EPIC with the SILAC data alone produced a substantially higher composite score than PrInCE achieved (Supplementary Table 1) and that EPIC also predicted up to five times as many complexes (with comparable or higher reliability) than PrInCE (Supplementary Table 1).

Disease and phenotype enrichment analysis. Since there is a lack of information available for Worm gene-disease associations, we combined several human resources and mapped human gene names to worm gene names via 1/1 orthology using InParanoid. Gene-disease associations were retrieved from the Online Mendelian Inheritance in Man (OMIM), UniProt and ClinVar databases. However, OMIM only provides gene-disease associations, and thus we retrieved a mapping from gene name to UniProt identifier via the UniProt identifier mapping web service. Moreover, OMIM does not provide a classification system for their diseases and different OMIM IDs might describe the same disease (for example, Alzheimer's has multiple identifiers depending on the types). Thus, we mapped each OMIM disease identifier to their corresponding disease ontology identifier (DOID). In the final step, we combined the resulting data set with a set of DOID annotations for Worm genes from the WormBase database. For phenotype analysis, we annotated our protein complexes with phenotype information taken from WormBase. Statistical enrichment for both phenotype and disease was determined by the Fisher exact test, and the Benjamini–Hochberg procedure was applied for multiple testing correction.

GO enrichment. The GO is a controlled vocabulary that describes genes by using three categories: molecular function, cellular component and biological process. We inferred enriched GO terms using the g:Profiler R package[49]. To ensure we only get significant hits we only considered GO terms with fewer than 500 proteins annotated to it, and the *P* value was corrected by the conservative Bonferroni correction procedure.

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

Data availability

The supporting co-fractionation data are available via ProteomeXchange with the identifier PXD011182. The entire WormMap network (Cytoscape format) is available on GitHub (https://github.com/BaderLab/EPIC/tree/master/WormMap) and has been submitted to the BioGRID database. Source Data for Fig. 2 are available online.

Code availability

EPIC is available via a Docker container (https://hub.docker.com/r/baderlab/bio-epic/). The EPIC software code is publicly available on GitHub (https://github.com/BaderLab/EPIC).

References

40. Stiernagle, T. in *WormBook: The Online Review of C. elegans Biology* (ed. The *C. elegans* Research Community) (WormBook).
41. Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A. I. & Marcotte, E. M. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. *J. Proteome Res.* **10**, 2949–2958 (2011).
42. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
43. Kislinger, T. et al. PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2**, 96–106 (2003).
44. Campagnola, P. J. et al. Three-dimensional high-resolution second-harmonic generation imaging of endogenous structural proteins in biological tissues. *Biophys. J.* **82**, 493–508 (2002).
45. Dupuy, D. et al. A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* **14**, 2169–2175 (2004).
46. Kwan, J. et al. DLG5 connects cell polarity and Hippo signaling protein networks by linking PAR-1 with MST1/2. *Genes Dev.* **30**, 2696–2709 (2016).
47. Wehrens, R., Melssen, W., Buydens, L. & de Gelder, R. Representing structural databases in a self-organizing map. *Acta Crystallogr. B* **61**, 548–557 (2005).
48. Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).
49. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).

nature research

Corresponding author(s): Andrew Emili, Gary D. Bader

Last updated by author(s): Mar 16, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the proteomics (mass spec) data analysis, we used MaxQuant42 (Version 1.6.0.16), Sequest (Version 2.7) and MSblender (https://github.com/marcottelab/MSblender) database search software tools.
Data analysis	The code of EPIC is available on Github (https://github.com/BaderLab/EPIC). The docker version of EPIC is available at (https://hub.docker.com/r/baderlab/bio-epic/). ReAdw is used to convert raw mass spectrometry files to the mzXML format, the code can be obtained at (https://github.com/PedrioliLab/ReAdW) PrInCE can be downloaded at (https://github.com/fosterlab/PrInCE) STATQUEST is a statistical algorithm to validate putative protein identifications, which is described in our previous paper (https://www.mcponline.org/content/2/2/96/tab-article-info). SciPy version 0.19.1 and NumPy 1.13.3 are used in this study, both can be download from (https://www.scipy.org/) or through Anaconda. Python version 2.7 and R version 3.5 + are used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

ProteomeXchange accession code: PXD011182 (reserved until publication)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We performed 16 replicate co-fractionation experiments (varying column conditions, not sample source). In each experiment, we collected multiple column fractions to make sure all chromatographic peaks are covered. In total, we have performed thousands of fractionation experiments, and used eight different correlation scores to extract features. Machine learning classifier used these features for PPIs prediction. Machine learning algorithm is intrinsically performs statistical learning/classifying and give confidence score to each PPIs, we retained the highly confidently ones in our final network.
Data exclusions	No data were excluded from the analyses.
Replication	We only considered proteins identified in more than two fractions by removing one-hit wonders. In PPIs prediction, we applied machine learning algorithm to assign confidence score to each predicted PPIs, and all low confidence PPIs were filtered out. For our validation (AP/MS) work, we performed the experiments in technical duplicate, and report the spectral counts obtained relative to two other unrelated protein baits to contrast the specificity of a given protein of interest. The interacting proteins have interaction patterns that are consistent with our initial co-elution (EPIC) results, reinforcing the reliability of our data and conclusions.
Randomization	Randomization was not performed as in this study as we mainly focused on using co-fractionation experiments to predict protein-protein interactions.
Blinding	Blinding is not relevant to our study, since all the predictions were done by machine learning classifier.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	ThermoFisher Scientific; GFP Antibody, ABfinity™ Rabbit Monoclonal; Catalog number: G10362; clone name: ; lot number: 1965886
Validation	The antibody was validated by ELISA assay (https://assets.thermofisher.com/TFS-Assets/BID/certificate/Certificates-of-Analysis/G10362%20Lot%201965886%20CofA.pdf?fbclid=IwAR32dJ4QpTluzmdiYPEa7JGLrL7kvsE_CFXZFzdwj04QGLElQrgugP3-XR0)

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Caenorhabditis elegans: wild type N2, RW1596 and GFP tagged strains (F26E4.4, Y34B4A.6 and F13H8.2); Samples were collected from mixed-stage population. Male individual are very rare, so majority of the population collected in this work were female.
Wild animals	The study did not involve wild animals.

Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	No ethical approval or guidance was required, as only the invertebrate worm Caenorhabditis elegans was used in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.