Published in final edited form as:
Nature. 2015 September 17; 525(7569): 339–344. doi:10.1038/nature14877.

Panorama of ancient metazoan macromolecular complexes

Cuihong Wan[1,2,#], **Blake Borgeson**[2,#], **Sadhna Phanse**[1], **Fan Tu**[2], **Kevin Drew**[2], **Greg Clark**[3], **Xuejian Xiong**[4,5], **Olga Kagan**[1], **Julian Kwan**[1,4], **Alexandr Berzginov**[3], **Kyle Chessman**[4,5], **Swati Pal**[4,5], **Graham Cromar**[4,5], **Ophelia Papoulas**[2], **Zuyao Ni**[1], **Daniel R. Boutz**[2], **Snejana Stoilova**[1], **Pierre C. Havugimana**[1], **Xinghua Guo**[1], **Ramy H. Malty**[7], **Mihail Sarov**[8], **Jack Greenblatt**[1,4], **Mohan Babu**[7], **Brent Derry**[4,5], **Elisabeth Tillier**[3], **John B. Wallingford**[2,6], **John Parkinson**[4,5], **Edward M. Marcotte**[2,6,*], and **Andrew Emili**[1,4,*]

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

[2]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA

[3]Department of Medical Biophysics, Toronto, Ontario, Canada

[4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

[5]Hospital for Sick Children, Toronto, Ontario, Canada

[6]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA

[7]University of Regina, Regina, Saskatchewan, Canada

[8]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Abstract

Macromolecular complexes are essential to conserved biological processes, but their prevalence across animals is unclear. By combining extensive biochemical fractionation with quantitative mass spectrometry, we directly examined the composition of soluble multiprotein complexes among diverse metazoan models. Using an integrative approach, we then generated a draft conservation map consisting of >1 million putative high-confidence co-complex interactions for species with fully sequenced genomes that encompasses functional modules present broadly

[*]Communicating authors – contact information: [AE] – CCBR Rm 914, 160 College Street, Toronto, Ontario, Canada M5S 3E1 Phone: 617-610-4042; Fax: 416-978-8528; andrew.emili@utoronto.ca. [EMM] – MBB 3.148, 2500 Speedway, Austin, Texas, USA 78712 Phone: 512-471-5435; Fax: 512-232-3472; marcotte@icmb.utexas.edu.

[#]These authors contributed equally to this work

Correspondence and requests for materials should be addressed to E.M.M (marcotte@icmb.utexas.edu) or A.E. (andrew.emili@utoronto.ca).

AUTHOR CONTRIBUTIONS

A.E. and E.M.M. designed and co-supervised the project. C.W. performed proteomic experiments, aided by P.H. B.B. coordinated data analysis, aided by S.P., K.D., and S.S. and guided by E.M.M. E.T., G.C. (UofT), A.B., J.P., X.X., K.C., G.C. (Sick Kids), C.W. and S.P. analysed network and conservation data. C.W., F.T., O.K., J.K., S.P. (Sick Kids), O.P., Z.N., D.R.B., X.G., R.H.M, M.S., J.G., M.B., B.D., and J.B.W. contributed validation experiments. S.P. designed the web portal. C.W., B.B., E.M.M. and A.E. drafted the manuscript. All authors discussed results and contributed edits.

The authors declare no competing financial interests.

across all extant animals. Clustering revealed a spectrum of conservation, ranging from ancient Eukaryal assemblies likely serving cellular housekeeping roles for at least 1 billion years, ancestral complexes that have accrued contemporary components, and rarer metazoan innovations linked to multicellularity. We validated these projections by independent co-fractionation experiments in evolutionarily distant species, by affinity-purification and by functional analyses. The comprehensiveness, centrality and modularity of these reconstructed interactomes reflect their fundamental mechanistic significance and adaptive value to animal cell systems.

Elucidating the components, conservation and functions of multiprotein complexes is essential to understand cellular processes[1,2], but mapping physical association networks on a proteome-wide scale is challenging. The development of high-throughput methods for systematically determining protein-protein interactions (PPI) has led to global molecular interaction maps for model organisms including *E. coli*, yeast, worm, fly and human[3–10]. In turn, comparative analyses have shown that PPI networks tend to be conserved[11,12], evolve more slowly than regulatory networks[13], and closely mirror function retention across orthologous groups[11,14,15]. Yet fundamental questions arise[16,17]: To what extent are physical interactions preserved between phyla? Which protein complexes are evolutionarily stable across animals? What is unique about their composition, phylogenetic distribution and phenotypic significance?

Since previous cross-species interactome comparisons, based on experimental data from different sources and methods, show limited overlap[12,18], we sought to produce a more comprehensive and accurate map of protein complexes common to metazoa by applying a standardized approach to multiple species. We employed biochemical fractionation of native macromolecular assemblies followed by tandem mass spectrometry to elucidate protein complex membership (Fig. 1; see Extended Methods). Previous application of this co-fractionation strategy to human cell lines preferentially identified Vertebrate specific protein complexes[6], so we selected eight additional species for study based on their relevance as model organisms, spanning roughly a billion years of evolutionary divergence (Fig. 1a). The resulting co-fractionation data (Fig. 1b) acquired for *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly), *Mus musculus* (mouse), *Strongylocentrotus purpuratus* (sea urchin), and human was used to discover conserved interactions (Fig. 1c), while the data obtained for *Xenopus laevis* (frog), *Nematostella vectensis* (sea anemone), *Dictyostelium discoideum* (amoeba), and *Saccharomyces cerevisiae* (yeast) was used for independent validation. Details on the cell types, developmental stages, and fractionation procedures used are provided in Supplementary Table 1.

We identified and quantified (see Extended Methods) 13,386 protein orthologs across 6,387 fractions obtained from 69 different experiments (Fig. 2a), an order of magnitude expansion in data coverage relative to our original (*H. sapiens* only) study[6]. Individual pair-wise protein associations were scored based on the fractionation profile similarity measured in each species. Next, we used an integrative computational scoring procedure (Fig. 1c; see Extended Methods) to derive conserved interactions for human proteins and their orthologs in worm, fly, mouse and sea urchin, defined as high pair-wise protein co-fractionation in at least two of the five input species. The support vector machine learning classifier used was trained (using 5-fold cross validation) on correlation scores obtained for conserved reference annotated protein complexes (see Extended Methods), and combined all of the input species co-fractionation data together with previously published human[6,19] and fly interactions[5] and additional supporting functional association evidence[20] (HumanNet). Notably, measurements of overall performance showed high precision with reasonable recall by the co-fractionation data alone (Fig. 2b), with external datasets serving only to increase precision and recall as we required all derived interactions to have significant biochemical support (see Extended Methods). Co-fractionation data of each input species impacted overall performance, in each case increasing precision and recall (Extended Data Fig. 1a).

The final filtered interaction network consists of 16,655 high-confidence co-complex interactions in human (Supplementary Table 2). All of the interactions were supported by direct biochemical evidence in at least two input species, with half (8,121) detected in 3 or more (Extended Data Fig. 1b), enabling cross-species modeling and functional inference.

Multiple lines of evidence support the quality of the network: Reference complexes withheld during training were reconstructed with higher precision and recall (Fig. 2b; see Extended Data Fig. 1c) relative to our human-only map[6]. The interacting proteins were also 6-fold enriched (hypergeometric p-value < 10^{-24}) for shared subcellular localization annotations in the Human Protein Atlas Database[21], 21-fold enriched (p-value < 10^{-56}) for shared disease associations in OMIM[22], and showed highly correlated human tissue proteome abundance profiles[23] (Extended Data Fig. 2a).

To independently verify the reliability of these projections, we examined the co-fractionation profiles of putatively interacting orthologs (*i.e.,* interologs) in the four holdout species, as obtained by protein quantification across 1,127 biochemical fractions (see Extended Methods). Strikingly, whereas sequence divergence changed absolute chromatographic retention times (Extended Data Fig. 2b), most of the predicted interactors showed highly correlated co-fractionation profiles among the holdout test species to a degree comparable to the input species used for learning (Fig. 2c). The biochemical data obtained for frog and sea anemone showed slightly better agreement than for *Dictyostelium* and yeast in proportion to evolutionary distance[24].

Besides indicating stably-associated proteins, our multi-species biochemical profiles faithfully recapitulated the architecture of multiprotein complexes of known 3D structure, with a general trend for most correlated protein pairs to be spatially closer (Extended Data Fig. 2c). For example, hierarchical clustering of 30S proteasome subunits according to chromatographic elution profiles of all five input species correctly separated the 20S and 19S particles and the regulatory lid from the base complex (Fig. 2d), reflecting known hierarchies of complex formation and disassembly.

Since most of the interacting components were phylogenetically conserved across vast evolutionary timescales, we were able to predict over 1 million high-confidence co-complex interactions among orthologous protein pairs for 122 extant Eukaryotes with sequenced genomes (Supplementary Table 3). The number of interactions ranged from 8,000 to 15,000 interactions per species depending on phyla (Fig. 2e), with more projected among

Deuterostomes, Protostomes and Cnidaria, which show high component retention, and fewer in Fungi, Plants, and, especially, Protists, where the relative paucity of co-complex conservation likely reflects inherent clade diversity, especially in parasite genomes (*e.g.,* gene loss among Apicomplexa). While largely congruent with previous smaller-scale studies of PPI conservation[25], the majority of conserved co-complex interactions are novel (*i.e.,* <1/3 curated in CORUM, STRING and GeneMania databases; Fig. 2e). This markedly increases the number of metazoan protein interactions reported to date (Supplementary Table 3), covering roughly 10–25% of the estimated conserved animal cell interactome[26,27], opening up many new avenues of inquiry.

To systematically define evolutionarily conserved functional modules, we partitioned the interaction network using a two-stage clustering procedure (Fig. 1c; see Extended Methods) that allowed proteins to participate in multiple complexes (*i.e.,* moonlighting) as merited (Extended Data Fig. 3a). The 981 putative multiprotein groupings (Fig. 3a; see Supplementary Table 4) includes both many well-known and novel complexes linked to diverse biological processes (Extended Data Fig. 3b). The complexes have estimated component ages spanning from ~500 million (*i.e*., metazoan-specific, or *new*) to over 1 billion years (*i.e.,* ancient, or *old*) of evolutionary divergence. Details of species, orthologs, taxonomic groups, protein ages and evolutionary distances are provided in Supplementary Tables 3 and 5 and Supplementary Material.

Strikingly, although proteins arising in metazoa (*i.e.* by gene duplication or other means) account for ~3/4 of all human gene products, they form only ~1/3 (39%; 147) of the clusters (Fig. 3a). These ‘new’ complexes tend to be smaller (*i.e.,* ≤3 components; Fig. 3b) and specific (*i.e*., components not present in ‘mixed’ complexes). This indicates that although protein number and diversity greatly increased with the rise of animals[25], most stable protein complexes were inherited from the unicellular ancestor and subsequently modified slightly over time (Fig. 3c and Supplementary Table 5). Indeed, the dominant phylogenetic profile of complexes across Eukarya (Fig. 3d) is composed either entirely (344 ‘old’ complexes) or predominantly (490 ‘mixed’ complexes) of ancient subunits ubiquitous among eukaryotes (Extended Data Fig. 4a; see Supplementary Table 5 for details), the latter presumably reflecting preferential accretion of new components to pre-existing macromolecules (Fig. 3c)[28].

These primordial complexes are present throughout the Opisthokonta supergroup (animals and fungi), estimated to be >1 billion years old[29], and Plants (and presumably lost/ significantly diverged among parasitic Protists). Reflecting this central importance, these complexes tend strongly to be ubiquitously expressed throughout all cell types and tissues (Extended Data Fig. 5a), are abundant (Extended Data Fig. 5b), and are enriched for associations to human disease and perturbation phenotypes in *C. elegans* (Supplementary Table 6). In comparison with other proteins in the 16,655 interactions, the older, conserved proteins present in these stable complexes have lower average domain complexity ($p < 0.02$; see Extended Methods), suggesting multi-domain architectures underlie more transient or tissue-specific interactions. Notably, whereas ‘mixed’ and ‘old’ complexes are enriched for functional associations with core cellular processes, such as metabolism (Extended Data Fig. 4c), the strictly metazoan complexes were far more likely to be linked to cell adhesion, organization and differentiation, consistent with roles in multicellularity. Reflecting these different evolutionary trajectories, ‘new’ clusters are substantially more enriched for cancer-related proteins (42%; 62/147; hypergeometric $p \leq 10^{-5}$) compared to strictly ‘old’ (15%; 53/344; $p \leq 10^{-3}$) clusters (Z-test < 0.0001) (Supplementary Table 7), have generally lower annotation rates (Extended Data Fig. 4b), and show different preponderances of protein domains (Extended Data Fig. 4c and Supplementary Table 6).

We used multiple approaches to assess the accuracy (Fig. 4) and functional significance (Fig. 5) of the predicted complexes. First, we performed affinity purification-mass spectrometry (AP/MS) experiments on select novel complexes from the ‘new’, ‘old’ and ‘mixed’ age clusters, validating most associations in both worm and human (Fig. 4a, Extended Data Fig. 6a). We next performed a global validation by comparing our derived complexes to a newly reported large-scale AP/MS study of 23,756 putative human protein interactions detected in cell culture (BioGrid pre-publication 166968, Huttlin EL *et al.*, downloaded Feb. 10, 2015), and observed a partial, but exceptionally significant, overlap to a degree comparable to literature-derived complexes (Fig. 4b, Extended Data Fig. 6b).

We also observed broad agreement between the derived complexes’ inferred molecular weights (assuming 1:1 stiochiometries) and migration by size exclusion chromatography (Fig. 4c; Extended Data Fig. 7a) and density gradient centrifugation (Extended Data Fig. 7b). A prime example is the coherent profiles of a large (~500 kDa) ‘mixed’ complex with several unannotated components (Fig. 4d; Extended Data Fig. 8), dubbed Commander because most subunits share COMM (copper metabolism MURR1) domains[30] implicated in copper toxicosis[31], among other roles[30,32]. Commander contains coiled-coil domain proteins CCDC22 and CCDC93 (Figs. 4a, d) in addition to ten COMM domain proteins, broadly supported by co-fractionation in human, fly and sea urchin (Extended Data Fig. 9a–c and Supporting Web Site).

We found an unexpected role in embryonic development for Commander, whose subunits are strongly co-expressed in developing frog (Extended Data Fig. 9d, e). Strikingly, COMMD2/3 knockdown (morpholino) tadpoles showed impaired head and eye development (Fig. 5a; Extended Data Fig. 9f, h), and defective neural patterning and expression changes in brain markers PAX6, EN2 and KROX20/EGR1 (Fig. 5b; Extended Data Fig. 9g, h). Given CCDC22’s recent link[33,34] to human syndromes of intellectual disability, malformed cerebellum and craniofacial abnormalities, the deep conservation of the Commander complex suggests COMMD2/3 as strong candidates in the etiology of these heterogeneous disorders.

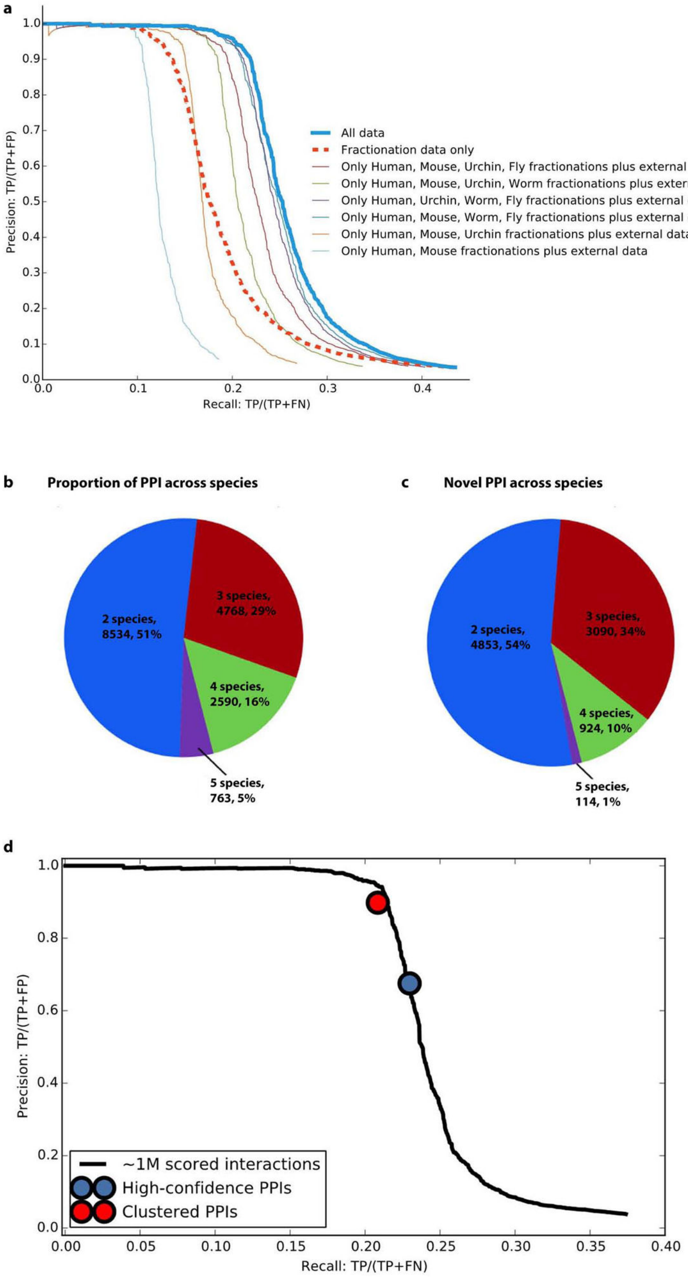
Among metazoan-specific protein complexes, we confirmed physical and functional associations of spindle checkpoint protein BUB3 with ZNF207, a zinc finger protein conspicuously lacking orthologs in cnidarians and fungi. ZNF207 binds Bub3 *via* a Gle2-binding-sequence (GLEBS) motif[35] restricted to deuterostomes and protostomes (Extended Data Fig. 10a). As in human, knockdown of ZNF207 in *C. elegans* enhanced lethality due to impaired Bub3-mediated checkpoint arrest (Fig. 5c).

Among ‘mixed’ complexes, we confirmed metazoan-specific coiled-coil domain protein CCDC97 as a sub-stoichiometric component of human and worm SF3B spliceosomal complex involved in branch site recognition (Fig. 4a). Consistent with a possible role in pre-mRNA splicing, CRISPR-based CCDC97 knockout human cells were slower growing than control lines (Extended Data Fig. 10b, c) and hypersensitive to pladienolide B (Fig. 5d), a macrolide inhibitor of SF3b[36].

Knowledge of conserved macromolecular associations provides a roadmap for additional functional inferences. For instance, fractionation profiles can be compared for any pair of proteins in our dataset to search for evidence of interactions. Notably, we found significant enrichment for interactions among pairs of human proteins acting sequentially in annotated pathways[37] (Fig. 5e), especially G protein and MAP kinase cascades (Supplementary Table 8). Enzymes acting consecutively in core metabolic reactions (Fig. 5f) also showed a higher tendency to interact (Supplementary Table 8), whose significance decayed with more intervening steps (Fig. 5e). For example, strong consecutive interactions were apparent within the widely conserved purine biosynthetic pathway, with enzymes (e.g. PAICS, GART) eluting in two peaks (Fig. 5g), one coincident with the prior enzyme and the second with the downstream enzyme, suggestive of substrate channeling[38].

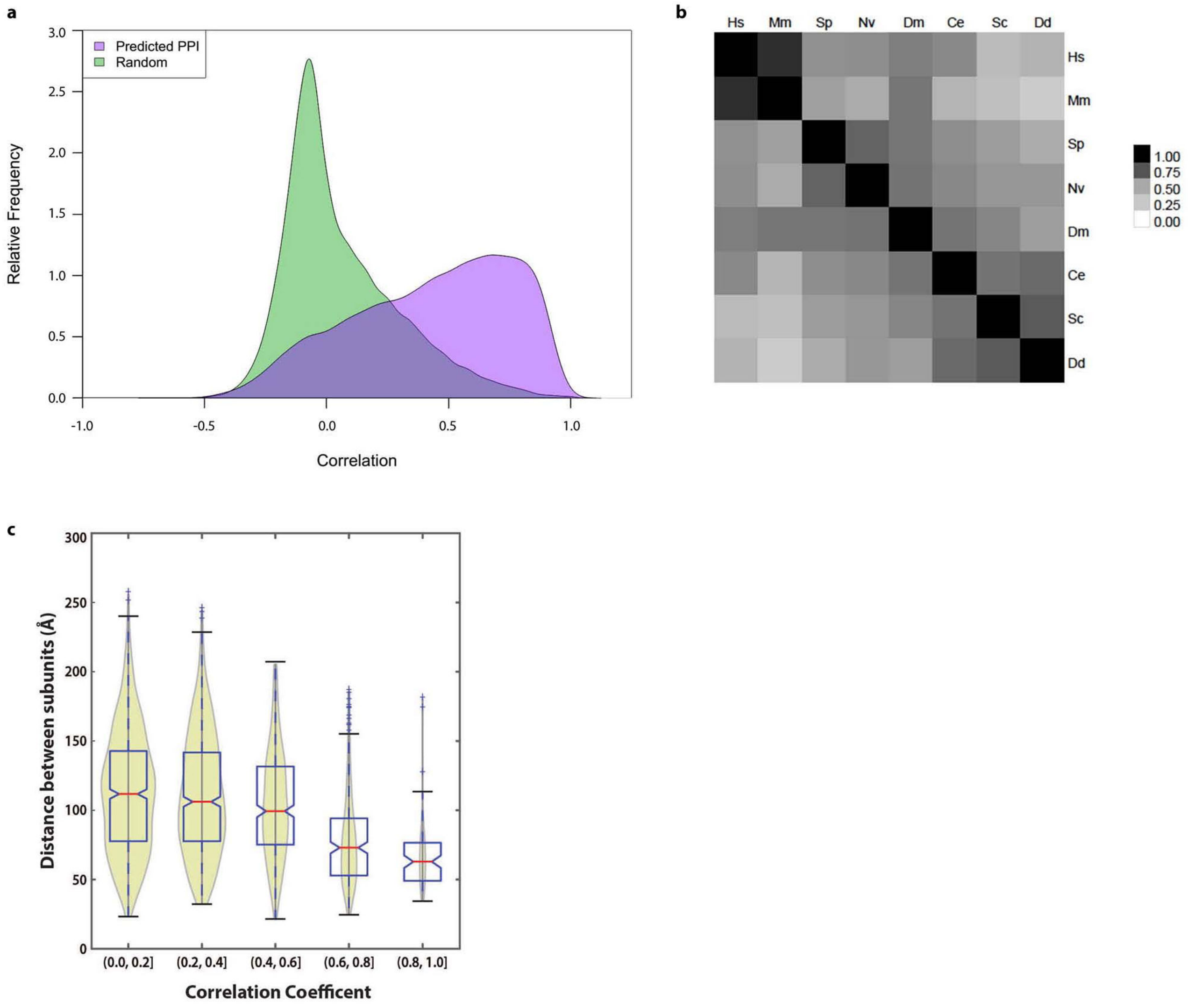
Despite the diversity of multicellular organisms, our study reveals fundamental attributes of the macromolecular machinery of animal cells with near universal pertinence to metazoan biology, development and evolution. Our massive set of supporting biochemical fractionation data (*via* ProteomeXchange with identifiers PXD002319-PXD002328), PPIs (*via* BioGRID; http://thebiogrid.org/185267/publication/panorama-of-ancient-metazoan-macromolecular-complexes.html) and interaction network projections are fully accessible (http://metazoa.med.utoronto.ca) to facilitate in-depth exploration. Although we focused on global conservation properties, these data can be analyzed at the individual animal species or complex levels to assess the variety and functional adaptations of particular protein assemblies across phyla.

Extended Data



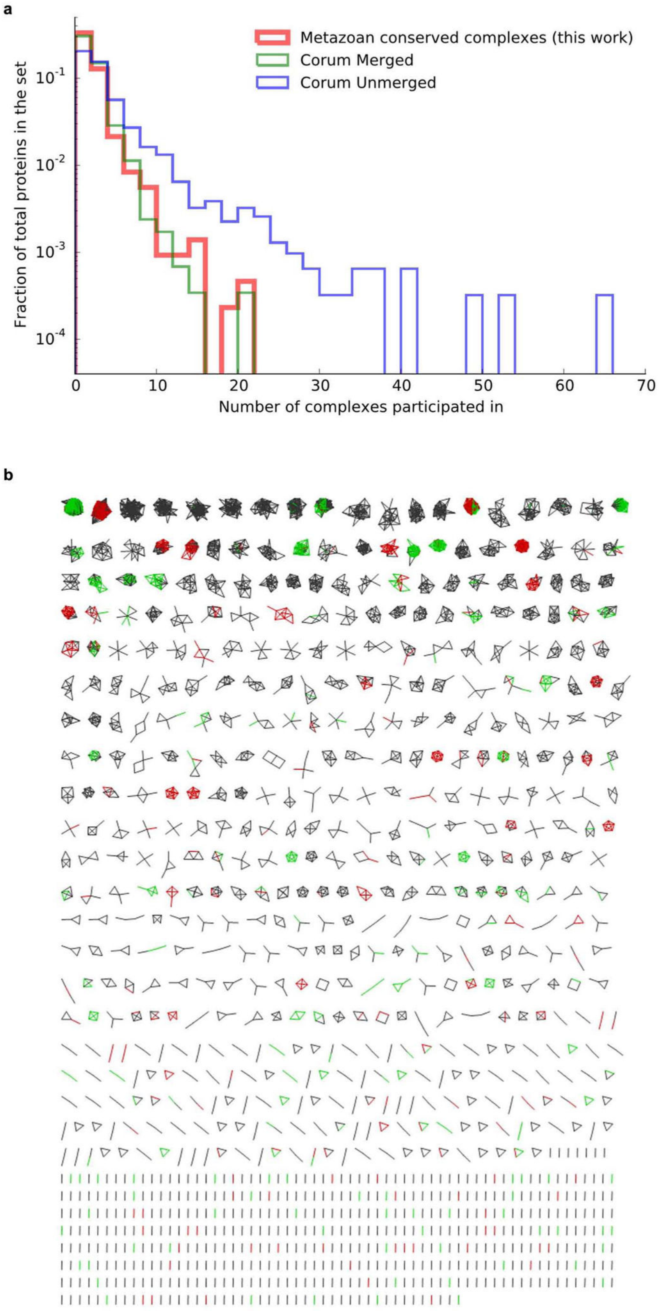
Extended Data Figure 1. Performance measures

a, Performance benchmarks, measuring the precision and recall of our method and data in identifying known co-complex interactions from a withheld reference set of annotated human complexes (from CORUM[39]; as in Fig. 2b). 5-fold cross-validation against this withheld set shows strong performance gains, beyond a baseline achieved using only human and mouse co-fractionation data along with additional evidence from independent protein interaction screens[5,19] and a functional gene network[20] (far-left curve), made by integrating co-fractionation data from the additional non-human animal species (as indicated). “All data" and "Fractionation data only" curves include biochemical fractionation data from all 5 input species: human, mouse, urchin, fly and worm; the latter curve omits all external data. In all cases, at least 2 species were required to show supporting biochemical evidence. Recall is shown fraction of 4,528 total positive interactions derived from the withheld human CORUM complexes. **b**, All 16,655 interactions were identified at least in two species, half (49%, 8,121) found in three or more species. **c**, Among these high-confidence co-complex interactions, 8,981 (54%) were not reported in iRef[44] (v13.0), Biogrid[45] (v3.2.119) or CORUM reference (Supplementary Table 2) for any of the five input species or in yeast; half (46%, 4,128) of these novel co-complex interactions have co-fractionation evidences in 3 or more species. **d**, Final precision/recall performance on withheld interaction test set. An SVM classifier was trained using interactions derived from our training set of CORUM complexes, then ~1M protein pairs co-eluting in at least 2 of the 5 input species were scored by the classifier. Black curve shows precision and recall for ranked list of co-eluting pairs, with recall representing fraction recovered of 4,528 total positive interactions derived from the withheld set of merged human CORUM complexes, and precision measured using co-eluting pairs where both members of the pair are contained in the set of proteins represented in the CORUM withheld set. The top 16,655 pairs, giving a cumulative precision of 67.5% and recall of 23.0% on this withheld test set, form the high-confidence set of co-complex protein-protein interactions (blue circle). The highest-scoring interactions were clustered using the two-stage approach described in the Extended Methods, yielding a final set of 7,669 interactions which form the 981 identified complexes (red circle; precision=90.0%, recall=20.8%).



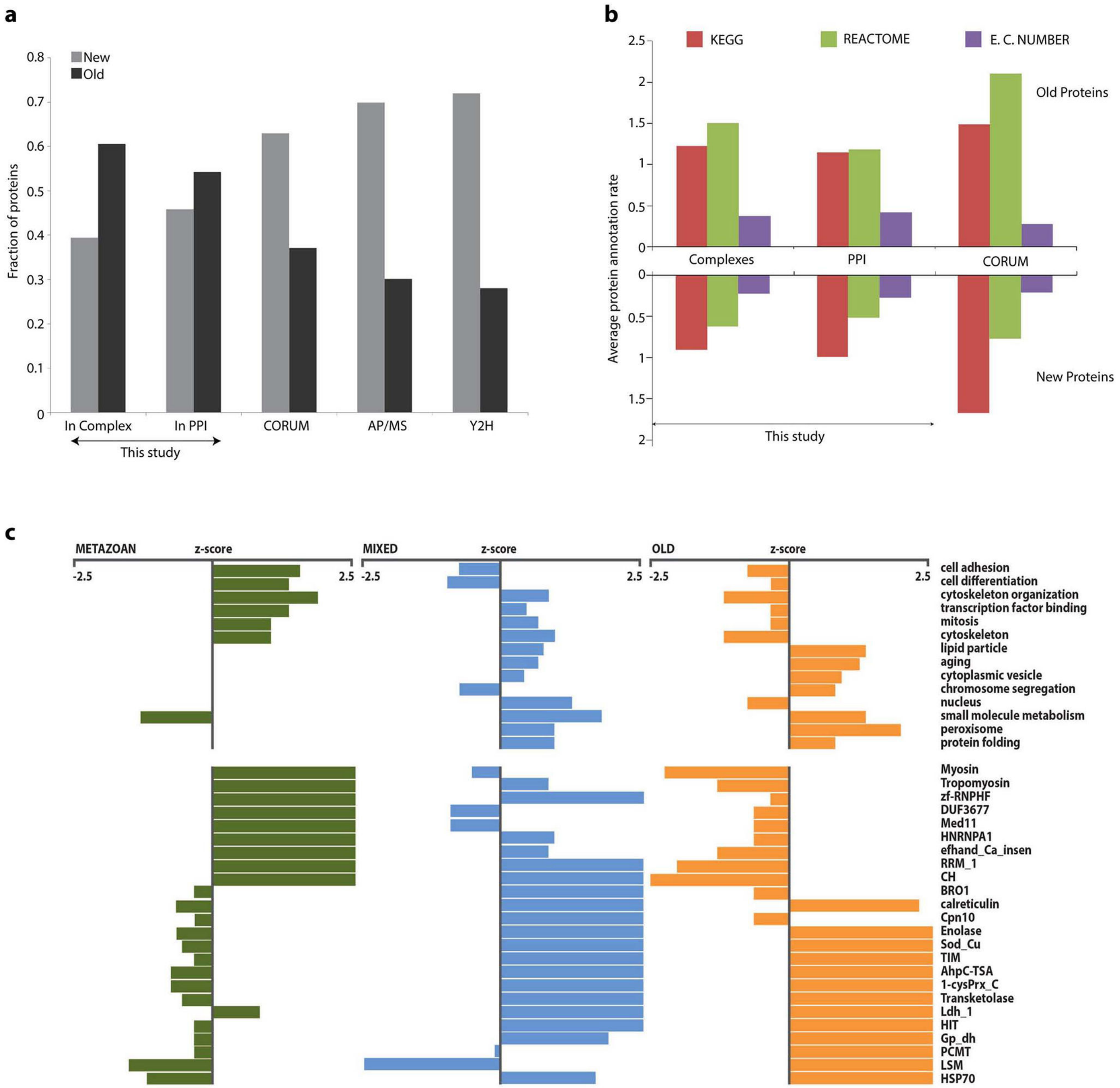
Extended Data Figure 2. Properties of protein elution profiles

a, Distribution of global protein tissue expression pattern similarity, measured as the Pearson correlation coefficient of protein abundance across 30 human tissues[23], showing markedly higher correlations for 16,468 protein-protein pairs of putative co-complex interaction partners compared to the same number of randomized pairs of proteins in the network which were not predicted to interact. **b**, Heatmap illustrating the low to moderate cross-species Spearman's rank correlation coefficients in the elution profiles observed between orthologous proteins during mixed-bed ion exchange chromatography (IEX-HPLC) under standardized conditions, highlighting the shift in absolute chromatographic retention times in different species. This variation indicates that the conservation of co-fractionation by putatively interacting proteins is not merely a trivial result stemming from fixed column retention times. **c**, The degree of co-fractionation is measured as the correlation coefficient between elution profiles. Spatial proximity is calculated from the mean of residue pair distances between components of multisubunit complexes with known 3D structures (see Extended Methods).



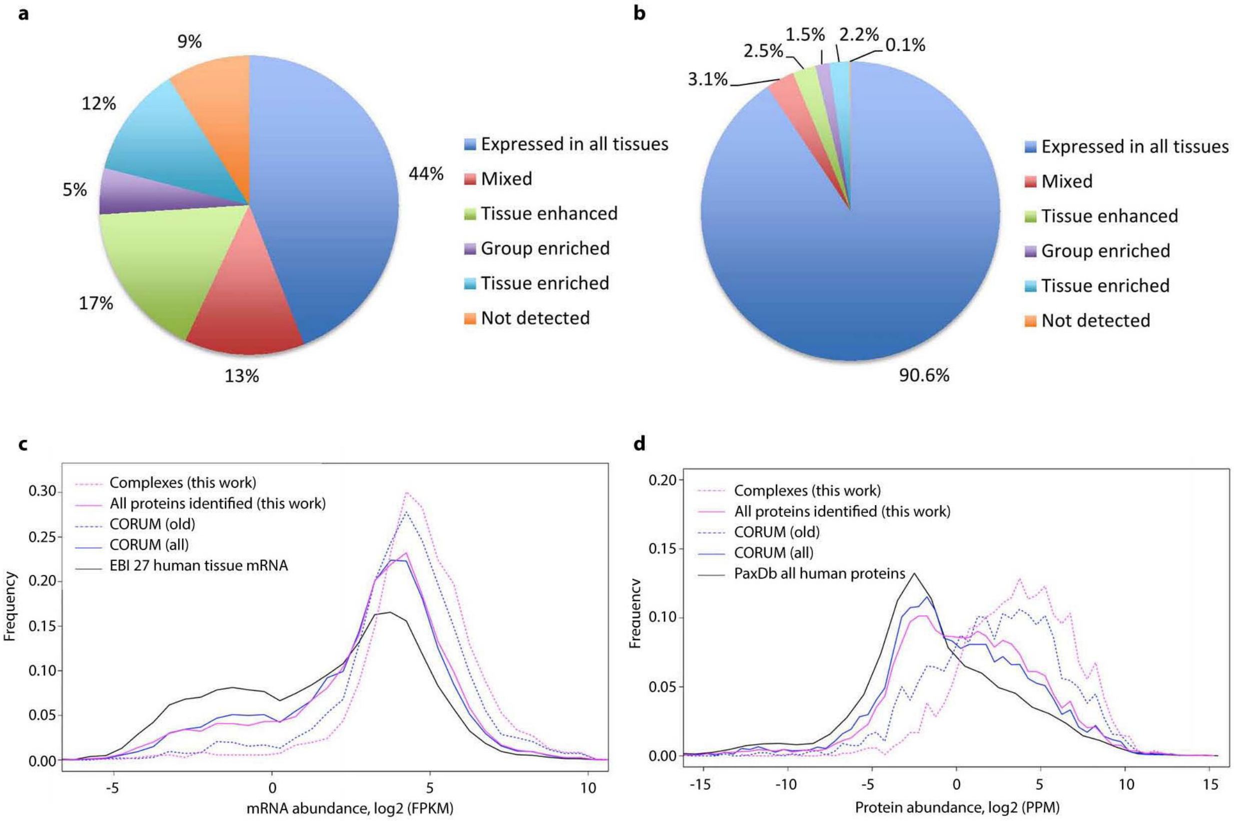
Extended Data Figure 3. Derivation of complexes

a, The 2,153 proteins present in the 981 derived metazoan complexes participate in multiple assemblies (‘moonlighting’) to an extent comparable to the sharing of subunits reported for literature-derived complexes (CORUM). For comparison, we examined the 1,550 unique proteins from the full CORUM set of 1,216 human complexes passing our selection criteria for supporting evidence (‘Unmerged’) and the 1,461 unique proteins from the non-redundant set of 501 merged complexes used as the reference for splitting our training and testing sets, with some of the largest complexes removed to avoid bias in training (‘Merged’; see ‘Optimizing the two-stage clustering’ in Extended Methods for details). **b**, Schematic of 981 identified complexes containing 2,153 unique proteins. In this graphical representation, 7,669 co-complex interactions are shown as lines, and proteins as nodes. Red and green interactions were previously annotated in CORUM. Red interactions were used in training the classifier and/or clustering procedure, while green interactions were held out for validation purposes. Gray interactions were not previously annotated in CORUM.



Extended Data Figure 4. Properties of new and old proteins and complexes

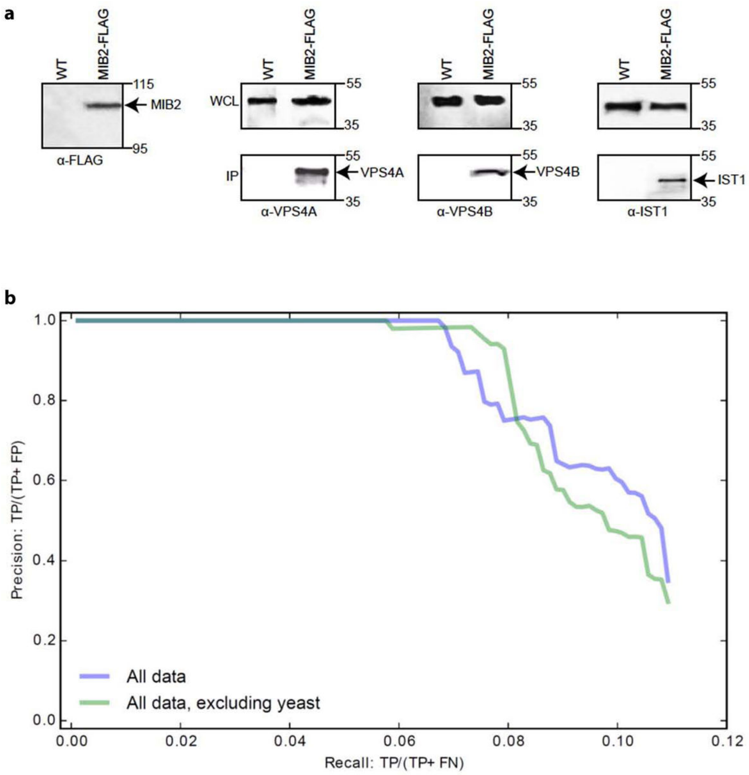
a, The 2,153 protein components in the conserved animal complexes tend to be more ancient than the 2,301 proteins reported in the CORUM reference complexes or in two recent large-scale protein interaction assays, based on either the 7,062 proteins found by affinity purification/mass spectrometry (AP/MS; BioGrid 166968, Huttlin EL (2014/pre-pub), downloaded Feb 10th 2015) or the 3,667 proteins analyzed by yeast two-hybrid assays (Y2H)[10]. Ages are derived from OMA as in ref. [25]. **b**, Annotation rates (mean count of annotation terms per protein) of old and new proteins in the derived complexes and pairwise PPI, compared with proteins in the CORUM reference complex set. Old proteins (defined by OMA) from the complexes generally exhibited higher annotation rates than new proteins. **c**, Differential enrichment of old, mixed and metazoan-specific protein complexes for functional annotations (select GO-slim biological process terms shown, top) and protein domains (Pfam, bottom).



Extended Data Figure 5. Abundance and expression trends for proteins in complexes

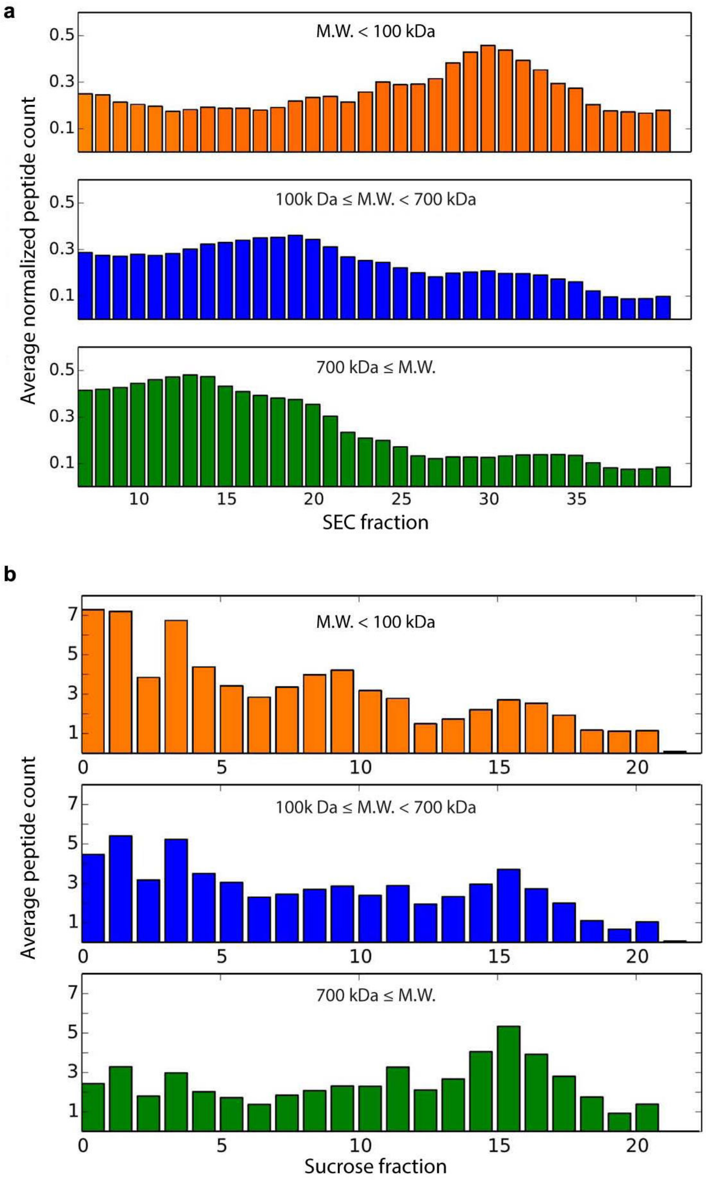
Proteins within the identified complexes tend to be ubiquitously expressed across human tissues. Pie charts show the proportions of proteins with varying tissue expression patterns, from a recently published human tissue proteome map[46], comparing: **a**, the full set of 20,258 human proteins, with **b**, the 2,131 proteins within the identified complexes. Consistent with these observations, 91% of the protein components in the complexes were expressed in >15 tissues in data from a reference human proteome[23], compared to less than half (46%) of the 17,294 proteins in the overall reference set (Z-test $p < 0.001$). The distributions of average mRNA and protein abundances for all proteins identified and those within complexes are shown in panel **c**, mRNA abundances (data from EBI accession E-MTAB-1733) and **d**, protein abundances (data from PaxDb integrated dataset, 9606-H.sapiens_whole_organism-integrated_dataset). Evolutionarily ‘old’ proteins (defined by OMA as described in ref. [25]

and mentioned earlier) tend towards higher abundances, even for proteins in reference complexes.



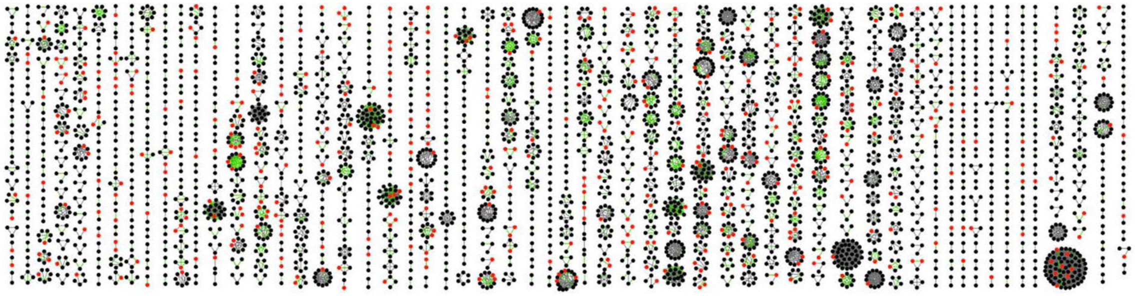
Extended Data Figure 6. Additional validation data

a, Confirmation of MIB2 interactions by co-immunoprecipitation. Extract (~10 mg protein) from cultured human HCT116 cells expressing FLAG-tagged MIB2 or control (WT) cells was incubated with 100 μl anti-FLAG M2 resin for 4 h by gently rotating at 4°C. After extensive washing with RIPA buffer, co-purifying proteins bound to the beads were eluted by the addition of 25 μl Laemmli loading buffer at 95 °C. Polypeptides were separated by SDS-PAGE and immunoblotted using FLAG, VPS4A, VPS4B or IST1 antibodies as indicated (expanded gel images provided in SI). **b**, Protein co-complex interactions reported in the CYC2008 yeast protein complex database[42] are reconstructed accurately from the co-fractionation data, regardless of whether the full set of co-fractionation plus external data are used to derive protein interactions (‘All data’, see also Fig. 4b) or if the external yeast data was specifically excluded from the analyses (‘All data, excluding yeast’).



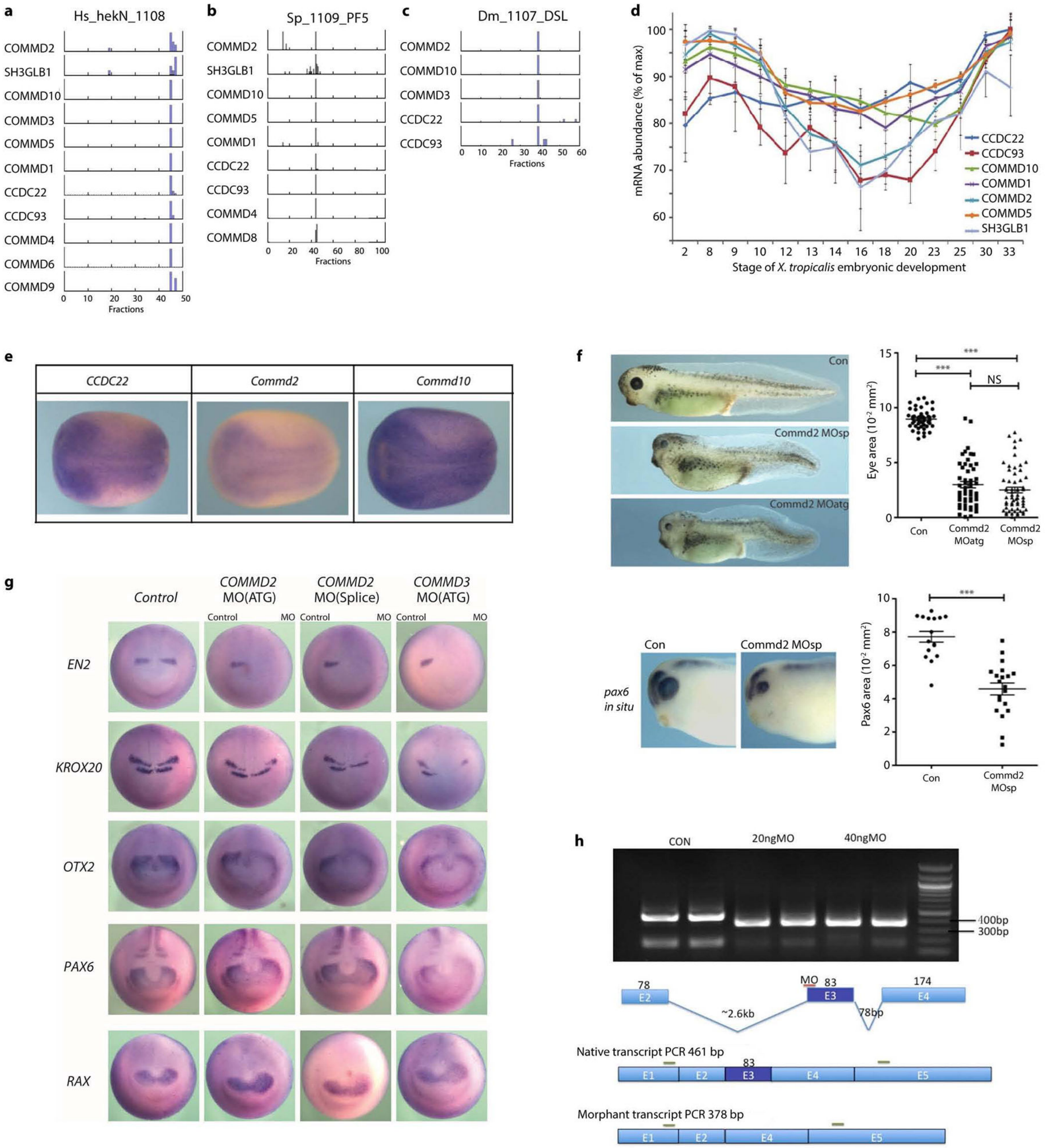
Extended Data Figure 7. Agreement of derived complexes' molecular weights with measurement by HPLC and density centrifugation

a, CORUM reference complexes' inferred molecular weights (MW) are consistent with their components' average cumulative size exclusion chromatograms. The molecular weights of each complex was calculated as the sum of putative component molecular weights, assuming 1:1 stoichiometry. Data from ref. [43] were analyzed as in Fig. 4c and show a similar trend as for the derived complexes. **b**, Derived complexes' inferred molecular weights (MW) are broadly consistent with their components' average cumulative ultracentrifugation profiles on a sucrose density gradient. Average profiles are plotted for *X. laevis* orthologs, based on a preparation of hemoglobin-depleted heart and liver proteins separated on a 7 – 47% sucrose density gradient, as described in the Extended Methods.



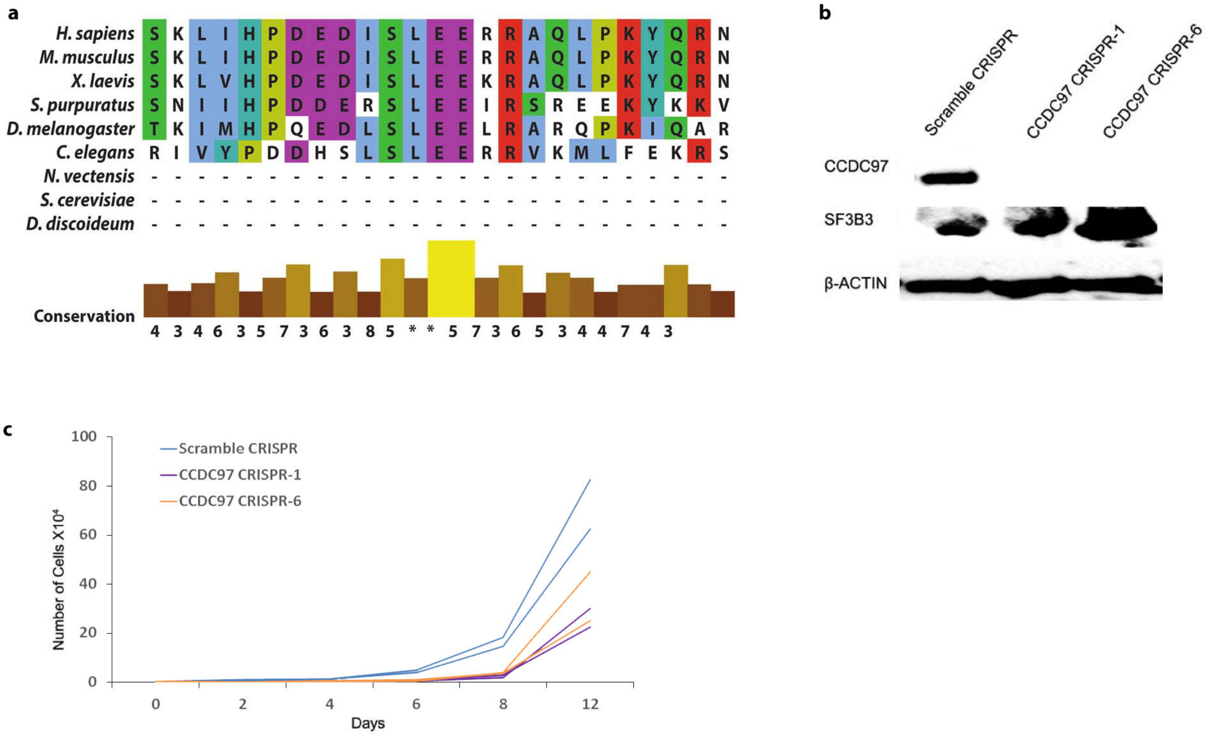
Extended Data Figure 8. Distribution of uncharacterized proteins and novel interactions across the 981 derived complexes

Complexes were sorted by median age (defined by OMA). Among 2,153 unique proteins, 293 (red) lack Gene Ontology (GO) functional annotations, while 1,756 of 7,665 co-complex interactions are novel (light green) (*i.e.,* not listed in iRef curation database).



Extended Data Figure 9. Properties of the Commander complex

The automatically-derived 8 subunit Commander complex (Fig. 3b) was subsequently extended to 13 subunits (COMMD1 to 10, CCDC22, CCDC93, and SH3GLB1) based on combined analysis of AP-MS (Fig. 4a), size exclusion chromatograms[43] (Fig. 4d), published pairwise interactions[30,47,48], and analysis of elution profiles of the remaining COMM domain containing proteins, as shown here. Example protein elution profiles are plotted for Commander complex subunits observed from: **a**, HEK293 cell nuclear extract; **b**, sea urchin embryonic (5 days post-fertilization) extract; and **c**, fly SL2 cell nuclear extract; each fractionated by heparin affinity chromatography. **d**, Co-expression of Commander complex subunits during embryonic development of *X. tropicalis* (plotting mean +/− s.d. of 3 clutches; data from ref. [49]). **e**, mRNA expression patterns of Commander complex subunits in stage 15 *X. laevis* embryos. Images show coordinated spatial expression in early vertebrate embryogenesis, as measured by *in situ* hybridization (3 embryos examined). **f**, Knockdown of Commd2 induced marked head and eye defects in developing *X. laevis*. (*top*) Commd2 antisense knockdown significantly decreased eye size, shown for stage 38 tadpoles (from 3 clutches; control $n = 47$ animals, 1 eye each); phenotypes were consistent between translation blocking (MOatg; $n = 60$) morpholino reagents, splice site blocking (MOsp; $n = 50$) morpholinos, and knockdowns of interaction partner Commd3 (see Fig. 5a). ***, $p < 0.0001$, 2-sided Mann-Whitney test. (*bottom*) Commd2 knockdown induced altered Pax6 patterning in the embryonic eye (control $n = 8$ animals, 2 eyes each; MO $n = 11$). **g**, Commd2/3 knockdown animals show altered neural patterning. Changes in stage 15 *X. laevis* embryos, measured by *in situ* hybridization (assayed in duplicates; 5 embryos per treatment), seen upon knockdown but not on controls: the forebrain marker PAX6 was expanded, while the mid-brain marker EN2 was strongly reduced. Strikingly, while expression of KROX20/EGR1 in rhombomere R3 was shifted posteriorly, expression in R5 was strongly reduced or entirely absent. Panels in Fig. 5b are reproduced from this figure and are directly comparable. **h**, Confirmation of splice-blocking Commd2 morpholino activity. Images and schematic show the basis and results of RT-PCR and agarose gel electrophoresis obtained with the corresponding *X. laevis* knockdown tadpoles.



Extended Data Figure 10. Supporting data for BUB3 and CCDC97 experiments

a, Sequence alignment showing conservation of ZNF207 GLEBS domain. **b**, Targeted CRISPR/Cas9 induced knockout of CCDC97 in two independent lines of human HEK293 cells, as verified by Western blotting (expanded gel images provided in SI), also results in a slight decrease in annotated SF3B3 component levels. **c**, Loss of CCDC97 impairs cell growth. Lines show growth curves of control versus knockout cell lines in two biological replicate assays.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Gary Bader, Philip Kim, Gabriel Moreno-Hagelsieb, Shuye Pu, and Shoshana Wodak for critical suggestions, illustrator Angel Syrett for expert help drafting figures, Taejoon Kwon (University of Texas) for *X. laevis* gene models, and Kathy Foltz (University of California, Santa Barbara), Alexander Brehm (Philipps-University Marburg), Patrick Paddison (Fred Hutchinson Cancer Research Center), Joel Smith (Woods Hole Marine Biological Laboratory), Peter Zandstra and Jason Moffat (University of Toronto) for generously providing biological specimens and reagents. We thank members of the Emili and Marcotte laboratories for thoughtful assistance and guidance, and SciNet (University of Toronto) and the Texas Advanced Computing Center (University of Texas) for high-performance computing resources. This work was supported by grants from the C.I.H.R., N.S.E.R.C., O.R.F. and the C.F.I. to A.E., from the C.I.H.R. and Heart and Stroke to J. P., from the N.I.H. (F32GM112495) to K.D., and from the N.I.H., N.S.F., C.P.R.I.T., and Welch Foundation (F-1515) to E.M.M.

References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999; 402:C47–C52. DOI: 10.1038/35011540 [PubMed: 10591225]
2. Alberts B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell. 1998; 92:291–294. DOI: 10.1016/s0092-8674(00)80922-8 [PubMed: 9476889]
3. Butland G, et al. Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature. 2005; 433:531–537. [PubMed: 15690043]
4. Krogan NJ, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006; 440:637–643. [PubMed: 16554755]
5. Guruharsha KG, et al. A Protein Complex Network of Drosophila melanogaster. Cell. 2011; 147:690–703. [PubMed: 22036573]
6. Havugimana PC, et al. A Census of Human Soluble Protein Complexes. Cell. 2012; 150:1068–1081. [PubMed: 22939629]
7. Stelzl U, et al. A human protein-protein interaction network: A resource for annotating the proteome. Cell. 2005; 122:957–968. [PubMed: 16169070]
8. Li SM, et al. A map of the interactome network of the metazoan C-elegans. Science. 2004; 303:540–543. [PubMed: 14704431]
9. Hu PZ, et al. Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins. PLoS Biol. 2009; 7:929–947. DOI: 10.1371/journal.pbio.1000096
10. Rolland T, et al. A proteome-scale map of the human interactome network. Cell. 2014; 159:1212–1226. DOI: 10.1016/j.cell.2014.10.050 [PubMed: 25416956]
11. Sharan R, et al. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A. 2005; 102:1974–1979. [PubMed: 15687504]
12. Gandhi TKB, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature Genet. 2006; 38:285–293. [PubMed: 16501559]
13. Tan K, Shlomi T, Feizi H, Ideker T, Sharan R. Transcriptional regulation of protein complexes within and across species. Proc Natl Acad Sci U S A. 2007; 104:1283–1288. [PubMed: 17227853]
14. Singh R, Xu JB, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci U S A. 2008; 105:12763–12768. DOI: 10.1073/pnas.0806627105 [PubMed: 18725631]
15. Yu HY, et al. Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. Genome Res. 2004; 14:1107–1118. DOI: 10.1101/gr.1774904 [PubMed: 15173116]

16. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012; 8

17. Kiemer L, Cesareni G. Comparative interactomics: comparing apples and pears? Trends in biotechnology. 2007; 25:448–454. [PubMed: 17825444]

18. von Mering C, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002; 417:399–403. [PubMed: 12000970]

19. Malovannaya A, et al. Analysis of the Human Endogenous Coregulator Complexome. Cell. 2011; 145:787–799. DOI: 10.1016/j.cell.2011.05.006 [PubMed: 21620140]

20. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21:1109–1121. DOI: 10.1101/gr.118992.110 [PubMed: 21536720]

21. Uhlen M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010; 28:1248–1250. DOI: 10.1038/nbt1210-1248 [PubMed: 21139605]

22. McKusick, VA. MENDELIAN INHERITANCE IN MAN: A CATALOGS OF HUMAN GENES AND GENETIC DISORDERS. Johns Hopkins University Press; Baltimore, Maryland: 1998.

23. Kim MS, et al. A draft map of the human proteome. Nature. 2014; 509:575. [PubMed: 24870542]

24. Rubin GM, et al. Comparative genomics of the eukaryotes. Science. 2000; 287:2204–2215. DOI: 10.1126/science.287.5461.2204 [PubMed: 10731134]

25. Bezginov A, Clark GW, Charlebois RL, Dar VUN, Tillier ERM. Coevolution Reveals a Network of Human Proteins Originating with Multicellularity. Mol Biol Evol. 2013; 30:332–346. DOI: 10.1093/molbev/mss218 [PubMed: 22977115]

26. Stumpf MPH, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008; 105:6959–6964. DOI: 10.1073/pnas.0708078105 [PubMed: 18474861]

27. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? Genome Biol. 2006; 7:9.

28. Eisenberg E, Levanon EY. Preferential attachment in the protein network evolution. Phys Rev Lett. 2003; 91:4.

29. Knoll AH. The early evolution of eukaryotes: a geological perspective. Science. 1992; 256:622–627. DOI: 10.1126/science.1585174 [PubMed: 1585174]

30. Burstein E, et al. COMMD proteins, a novel family of structural and functional homologs of MURR1. J Biol Chem. 2005; 280:22222–22232. DOI: 10.1074/jbc.M501928200 [PubMed: 15799966]

31. van de Sluis B, Rothuizen J, Pearson PL, van Oost BA, Wijmenga C. Identification of a new copper metabolism gene by positional cloning in a purebred dog population. Hum Mol Genet. 2002; 11:165–173. DOI: 10.1093/hmg/11.2.165 [PubMed: 11809725]

32. McDonald FJ. COMMD1 and ion transport proteins: what is the COMMection? Focus on “COMMD1 interacts with the COOH terminus of NKCC1 in Calu-3 airway epithelial cells to modulate NKCC1 ubiquitination”. Am J Physiol-Cell Physiol. 2013; 305:C129–C130. DOI: 10.1152/ajpcell.00128.2013 [PubMed: 23677795]

33. Kolanczyk M, et al. Missense variant in CCDC22 causes X-linked recessive intellectual disability with features of Ritscher-Schinzel/3C syndrome. Eur J Hum Genet. 2014; 109:1–6. DOI: 10.1038/ejhg.2014.109

34. Voineagu I, et al. CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. Mol Psychiatr. 2012; 17:4–7. DOI: 10.1038/mp.2011.95

35. Toledo CM, et al. BuGZ Is Required for Bub3 Stability, Bub1 Kinetochore Function, and Chromosome Alignment. Dev Cell. 2014; 28:282–294. DOI: 10.1016/j.devcel.2013.12.014 [PubMed: 24462187]

36. Kotake Y, et al. Splicing factor SF3b as a target of the antitumor natural product pladienolide. Nat Chem Biol. 2007; 3:570–575. DOI: 10.1038/nchembio.2007.16 [PubMed: 17643112]

37. Croft D, et al. The Reactome pathway knowledgebase. Nucleic Acids Research. 2014; 42:D472–D477. DOI: 10.1093/nar/gkt1102 [PubMed: 24243840]

38. Ovádi, J. CELL ARCHITECTURE AND METABOLITE CHANNELING. RG Landes Company; Austin, Texas: 1995.

39. Ruepp A, et al. CORUM: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Research. 2010; 38:D497–D501. [PubMed: 19884131]

40. Warde-Farley D, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 2010; 38:W214–W220. DOI: 10.1093/nar/gkq537 [PubMed: 20576703]

41. Franceschini A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research. 2013; 41:D808–D815. DOI: 10.1093/nar/gks1094 [PubMed: 23203871]

42. Pu SY, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research. 2009; 37:825–831. DOI: 10.1093/nar/gkn1005 [PubMed: 19095691]

43. Kirkwood KJ, Ahmad Y, Larance M, Lamond AI. Characterization of Native Protein Complexes and Protein Isoform Variation Using Size-fractionation-based Quantitative Proteomics. Mol Cell Proteomics. 2013; 12:3851–3873. DOI: 10.1074/mcp.M113.032367 [PubMed: 24043423]

44. Turner B, et al. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database. 2010; 2010:baq023. [PubMed: 20940177]

45. Stark C, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006; 34:D535–D539. [PubMed: 16381927]

46. Uhlen M, et al. Tissue-based map of the human proteome. Science. 2015; 347:394.

47. de Bie P, et al. Characterization of COMMD protein-protein interactions in NF-kappa B signalling. Biochem J. 2006; 398:63–71. DOI: 10.1042/bj20051664 [PubMed: 16573520]

48. Phillips-Krawczak CA, et al. COMMD1 is linked to the WASH complex and regulates endosomal trafficking of the copper transporter ATP7A. Mol Biol Cell. 2015; 26:91–103. DOI: 10.1091/mbc.E14-06-1073 [PubMed: 25355947]

49. Yanai I, Peshkin L, Jorgensen P, Kirschner MW. Mapping Gene Expression in Two Xenopus Species: Evolutionary Constraints and Developmental Flexibility. Dev Cell. 2011; 20:483–496. DOI: 10.1016/j.devcel.2011.03.015 [PubMed: 21497761]

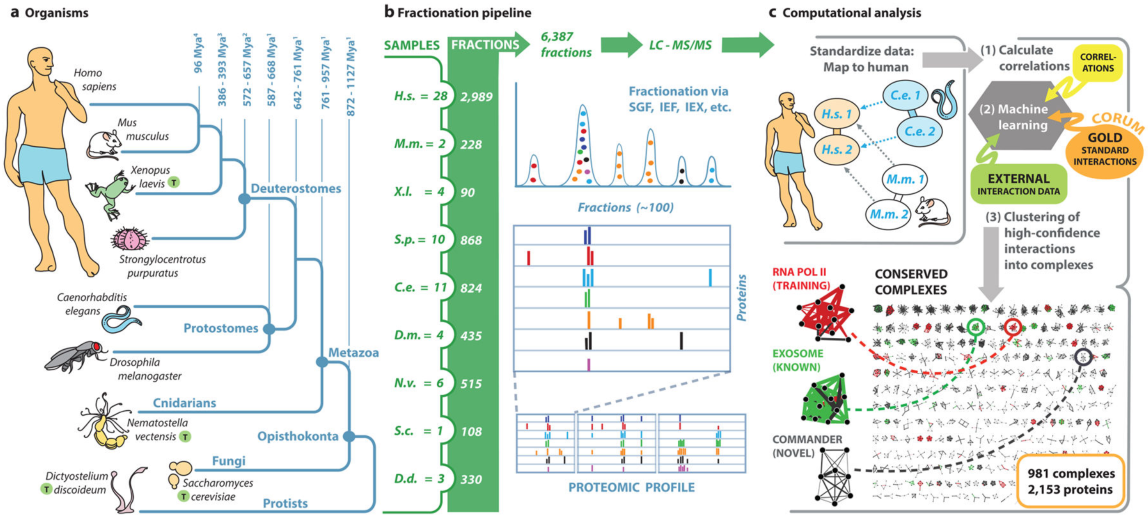


Figure 1. Workflow

a, Phylogenetic relationships of organisms analyzed in this study. We fractionated soluble protein complexes from worm (*C. elegans*) larvae, fly (*D. melanogaster*) S2 cells, mouse (*M. musculus*) embryonic stem cells, sea urchin (*S. purpuratus*) eggs, and human (HEK293/ HeLa) cell lines. Holdout species ('*T*', for test) likewise analyzed were frog (*X. laevis*), an amphibian; sea anemone (*N. vectensis*), a Cnidarian with primitive Eumetazoan tissue organization; slime mold (*D. discoideum*), an amoeba; and yeast (*S. cerevisiae*), a unicellular eukaryote. ***b***, Protein fractions were digested and analysed by high performance liquid chromatography-tandem mass spectrometry (LC-MS/MS), measuring peptide spectral counts and precursor ion intensities. ***c***. Integrative computational analysis: after ortholog mapping to human, correlation scores of co-eluting protein pairs detected in each 'input' species were subjected to machine learning together with additional external association evidence, using the CORUM complex database as a reference standard for training. High-confidence interactions were clustered to define co-complex membership.

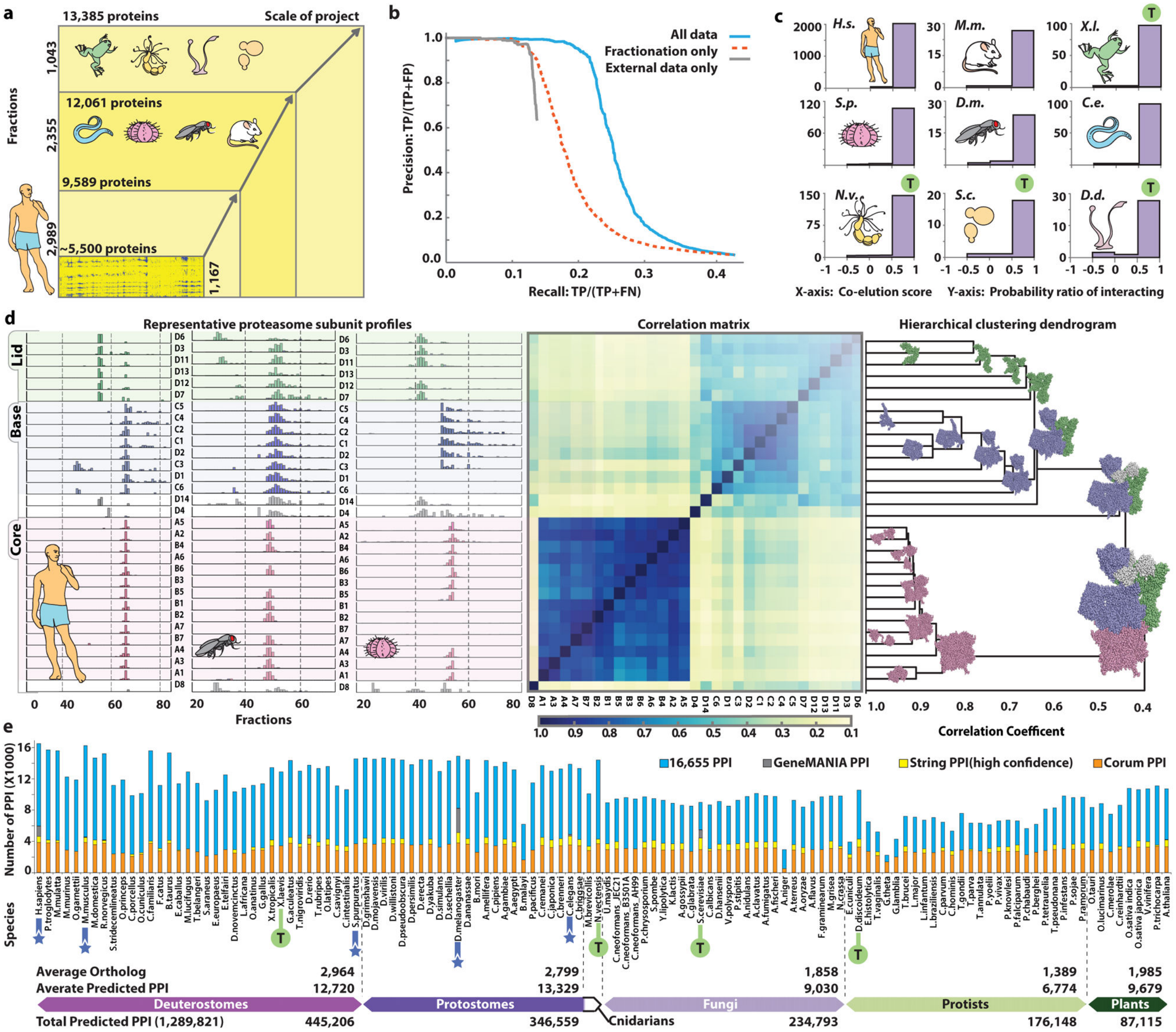


Figure 2. Derivation and projection of protein co-complex associations across taxa

a, Expanded coverage *via* experimental scale-up relative to our previous human study[6]. Chart shows number of proteins detected, most (63%) in two or more species. ***b***, Performance benchmarks, measuring precision and recall of our method and data in identifying known co-complex interactions (annotated human complexes from CORUM[39]). Complexes were split into training and withheld test sets; 5-fold cross-validation against 4,528 interactions derived from the withheld test set shows strong performance gains, beyond baselines achieved using only co-fractionation or external evidence alone. ***c***, Plots showing high enrichment (probability ratio of interacting) of predicted interacting orthologous protein pairs (relative to non-interacting pairs) among highly correlated fractionation profiles, in both the holdout validation (test, '*T*') and input species (colors reflect clade memberships). ***d***, (*left*) Representative co-fractionation data (normalized spectral counts shown for portions of 3 of 42 experimental profiles) from human, fly, and sea urchin showing characteristic profiles of proteasome core, base and lid subcomplexes.

Hierarchical clustering (*right*) of pan-species pairwise Pearson correlation scores (*centre*) is consistent with accepted structural models (PDB id: 4CR2; core, *red*; base, *blue*; lid, *green*; out-clusters, *white*). ***e***, Projection of conserved co-complex interactions across 122 eukaryotic species, indicating overlap with leading public PPI reference databases[39–41]. STRING bars indicate excess over CORUM; GeneMania bars indicate excess over both; component and interaction occurences across Clades indicated at bottom.

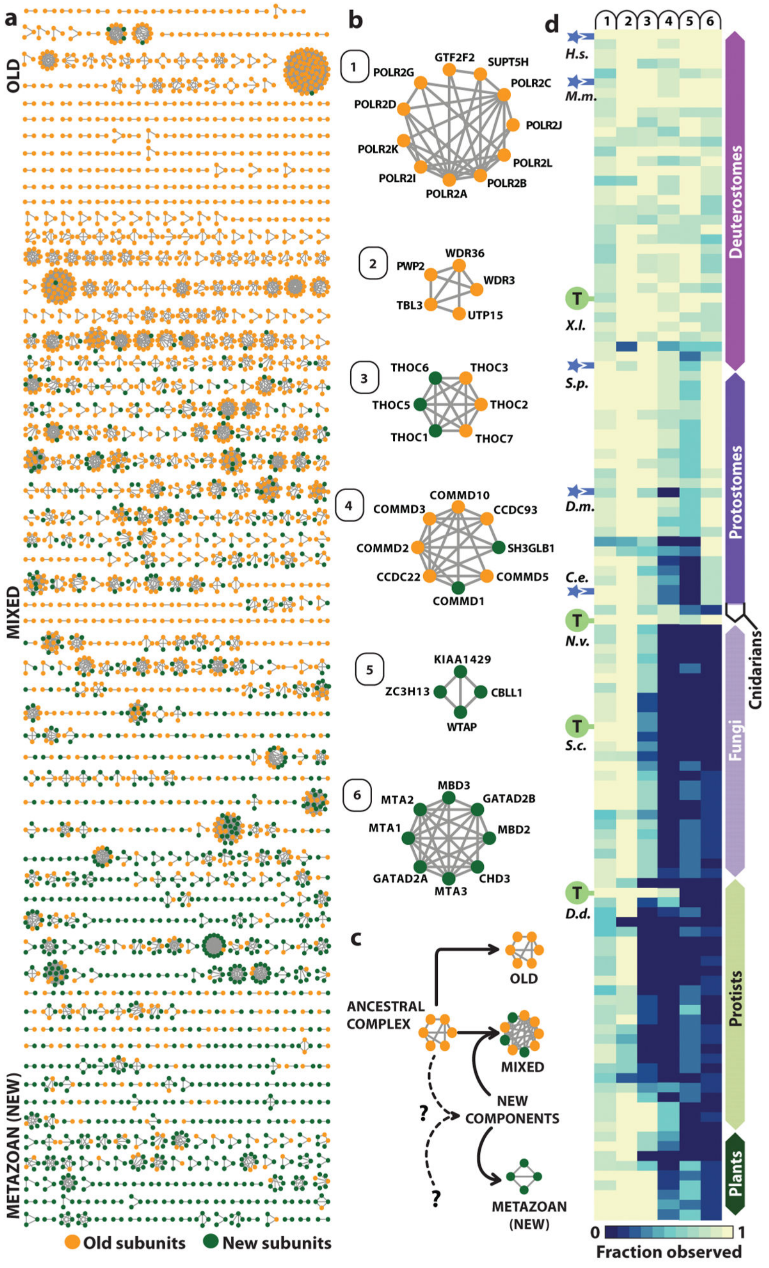
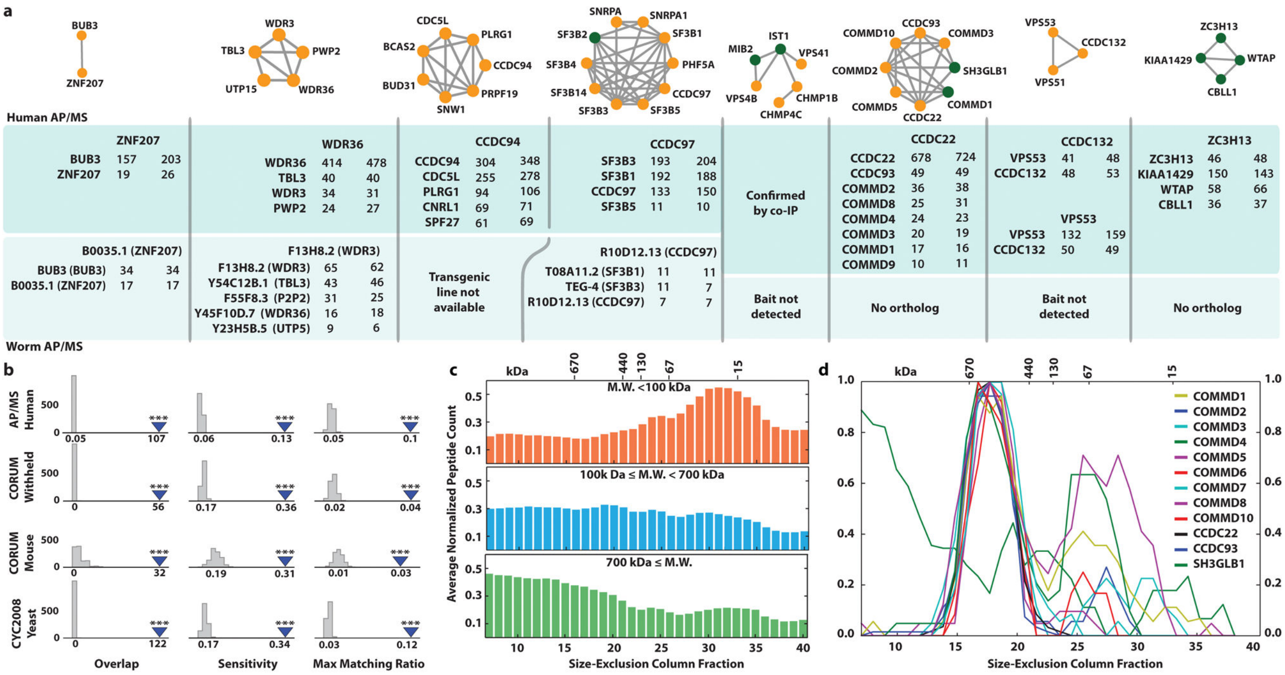


Figure 3. Prevalence of conservation of protein complexes across metazoa and beyond

a, Conserved multiprotein complexes, identified by clustering, arranged according to average estimated component age (see Extended Methods and ref. [25]). Proteins (nodes) classified as metazoan (*green*) or ancient (*orange*); assemblies showing divergent phylogenetic trajectories termed '*mixed*'. ***b***, Example complexes with different proportions of old and new subunits. ***c***, Presumed origins of metazoan (new), mixed, and old complexes; '?' indicates variable origins of new genes. ***d***, Heatmap showing prevalence of selected complexes across phyla. Color reflects fraction of components with detectable orthologs (absence, *dark blue*). Sea anemone (*N. vectensis*) most distant metazoan (Cnidarian) analyzed biochemically.



Human AP/MS

ZNF207		
BUB3	157	203
ZNF207	19	26

WDR36		
WDR36	414	478
TBL3	40	40
WDR3	34	31
PWP2	24	27

CCDC94		
CCDC94	304	348
CDC5L	255	278
PLRG1	94	106
CNRL1	69	71
SPF27	61	69

CCDC97		
SF3B3	193	204
SF3B1	192	188
CCDC97	133	150
SF3B5	11	10

Confirmed by co-IP

CCDC22		
CCDC22	678	724
CCDC93	49	49
COMMD2	36	38
COMMD8	25	31
COMMD4	24	23
COMMD3	20	19
COMMD1	17	16
COMMD9	10	11

CCDC132		
VPS53	41	48
CCDC132	48	53

VPS53		
VPS53	132	159
CCDC132	50	49

ZC3H13		
ZC3H13	46	48
KIAA1429	150	143
WTAP	58	66
CBLL1	36	37

Worm AP/MS

B0035.1 (ZNF207)		
BUB3 (BUB3)	34	34
B0035.1 (ZNF207)	17	17

F13H8.2 (WDR3)		
F13H8.2 (WDR3)	65	62
Y54C12B.1 (TBL3)	43	46
F55F8.3 (P2P2)	31	25
Y45F10D.7 (WDR36)	16	18
Y23H5B.5 (UTP5)	9	6

Transgenic line not available

R10D12.13 (CCDC97)		
T08A11.2 (SF3B1)	11	11
TEG-4 (SF3B3)	11	7
R10D12.13 (CCDC97)	7	7

Bait not detected

No ortholog

Bait not detected

No ortholog

Figure 4. Physical validation of complexes

a, Verification of complexes from tagged human cell lines and transgenic worms (see Extended Methods). Inset reports spectral counts obtained in replicate AP/MS analyses of indicated bait protein (header). MIB2-VPS4 complex confirmed by co-IP (Extended Data Fig. 6a). ***b***, Conserved complexes significantly overlap large-scale AP/MS data reported for human cell lines (BioGrid pre-pub 166968, Huttlin *et al.*, 2015) to a comparable extent as literature reference sets[39,42], using 3 measures of complex-level agreement (see Extended Methods, Extended Data Fig. 6b); ***, p-value < 0.001, determined by shuffling (gray distributions). ***c***, Agreement of inferred molecular weights (MW) of human protein complexes with size exclusion chromatography (SEC) profiles (data in ***c, d*** from ref. [43]). ***d***, Co-elution of human Commander complex subunits by SEC consistent with an approx. 500 kDa particle.

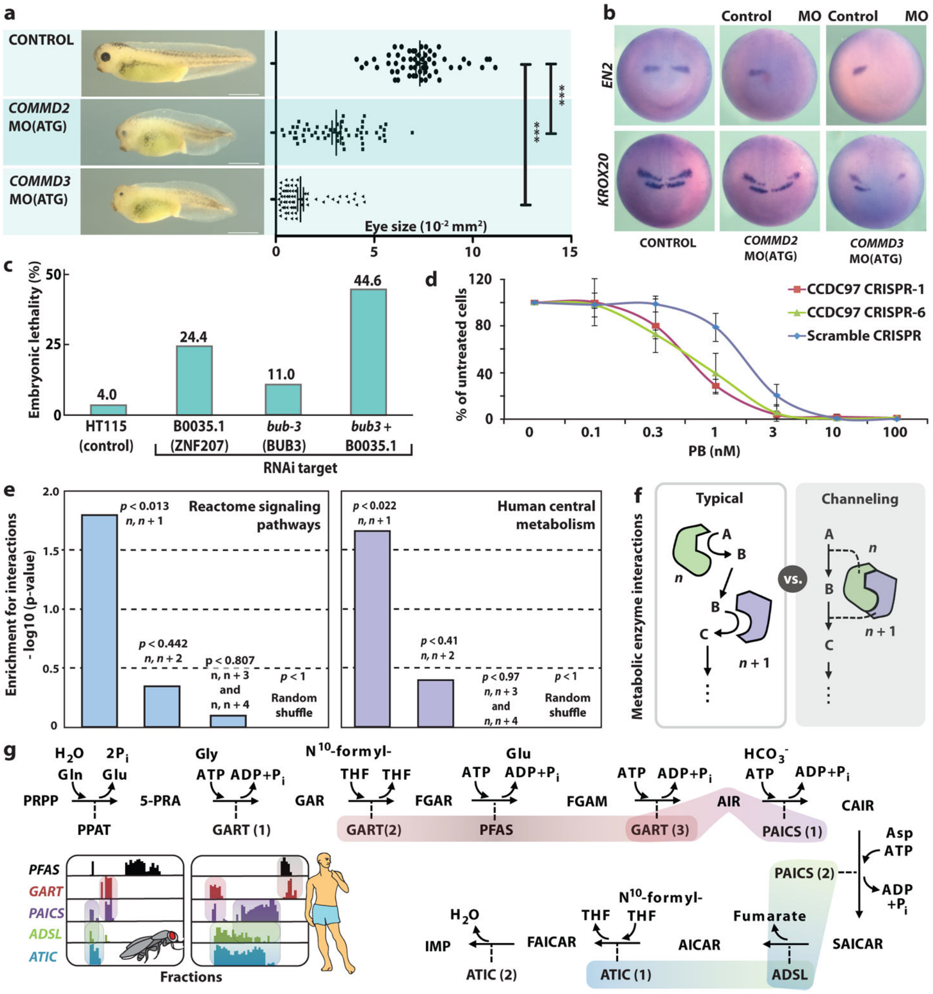


Figure 5. Functional validation of complexes

a, Morpholino knockdown of COMMD2 (*n* = 55 animals, 2 clutches, 1 eye each) or COMMD3 (*n* = 64) in *X. laevis* embryos causes defective head and eye development (control *n* = 57; Extended Data Fig. 9f, h). ***, *p* < 0.0001, 2-sided Mann-Whitney test. ***b***, COMMD2/3 knockdown animals (5 embryos per treatment examined) show altered neural patterning, including posterior shift or loss of expression of mid-brain marker EN2 and KROX20(EGR1), the latter in rhombomeres R3/R5 (compare to Extended Data Fig. 9g, h). ***c***, Enhanced embryonic lethality (*i.e.,* epistasis) following RNAi knockdown in *C. elegans* of B0035.1 (ZNF207) and *bub-3* together (eggs laid: HT115, 1308; B0035.1, 1096; *bub-3,* 445; *bub-3* + B0035.1, 341). ***d,*** Enhanced sensitivity (mean +/− s.d. across four cell culture experiments) of two independent CCDC97-knockout lines to the SF3b inhibitor pladienolide B (PB) relative to control HEK293 cells. ***e***, Enrichment (permutation test p-value) for interactions among sequential pathway components and metabolic enzymes relative to shuffled controls (n refers to enzyme index, where n,n+1 denotes sequential enzymes, n,n+2 sequential-but-one, etc, as described in SI ("Analysis of consecutively acting signal transduction and metabolic enzyme interactions"). ***f***, Metabolic channeling as opposed to traditional (typical) two-step cascade model. ***g***, Conserved interactions among consecutively acting enzymes involved in purine biosynthesis (2 representative co-fractionation profiles of the 69 total generated are shown).