ARTICLE

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 Bezginov[3], Kyle Chessman[4,5], Swati Pal[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[6], Mihail Sarov[7], Jack Greenblatt[1,4], Mohan Babu[6], W. Brent Derry[4,5], Elisabeth R. Tillier[3], John B. Wallingford[2,8], John Parkinson[4,5], Edward M. Marcotte[2,8] & Andrew Emili[1,4]

Macromolecular complexes are essential to conserved biological processes, but their prevalence across animals is unclear. By combining extensive biochemical fractionation with quantitative mass spectrometry, here we directly examined the composition of soluble multiprotein complexes among diverse metazoan models. Using an integrative approach, we generated a draft conservation map consisting of more than one million putative high-confidence co-complex interactions for species with fully sequenced genomes that encompasses functional modules present broadly across all extant animals. Clustering reveals a spectrum of conservation, ranging from ancient eukaryotic assemblies that have probably served cellular housekeeping roles for at least one 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, affinity purification and functional analyses. The comprehensiveness, centrality and modularity of these reconstructed interactomes reflect their fundamental mechanistic importance and adaptive value to animal cell systems.

Introduction

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 (PPIs) 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]. Here we define: (i) the extent to which physical interactions are preserved between phyla; (ii) the identity of protein complexes that are evolutionarily stable across animals; and (iii) the unique attributes of macromolecule composition, phylogenetic distribution and phenotypic significance.

Generating a high-quality conserved interaction dataset

As 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 Supplementary 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 on the basis of 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 were 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) were 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 Supplementary Methods) 13,386 protein orthologues 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 Supplementary Methods) to derive conserved interactions for human proteins and their orthologues 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 fivefold cross-validation) on correlation scores obtained for conserved reference annotated protein complexes (see Supplementary 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). Measurements of overall performance showed high precision with reasonable recall by the co-fractionation data alone (Fig. 2b), with external data sets serving only to increase

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [2]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA. [3]Department of Medical Biophysics, Toronto, Ontario M5G 1L7, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [5]Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. [6]Department of Biochemistry, University of Regina, Regina, Saskatchewan S4S 0A2, Canada. [7]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany. [8]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas 78712, USA.
*These authors contributed equally to this work.

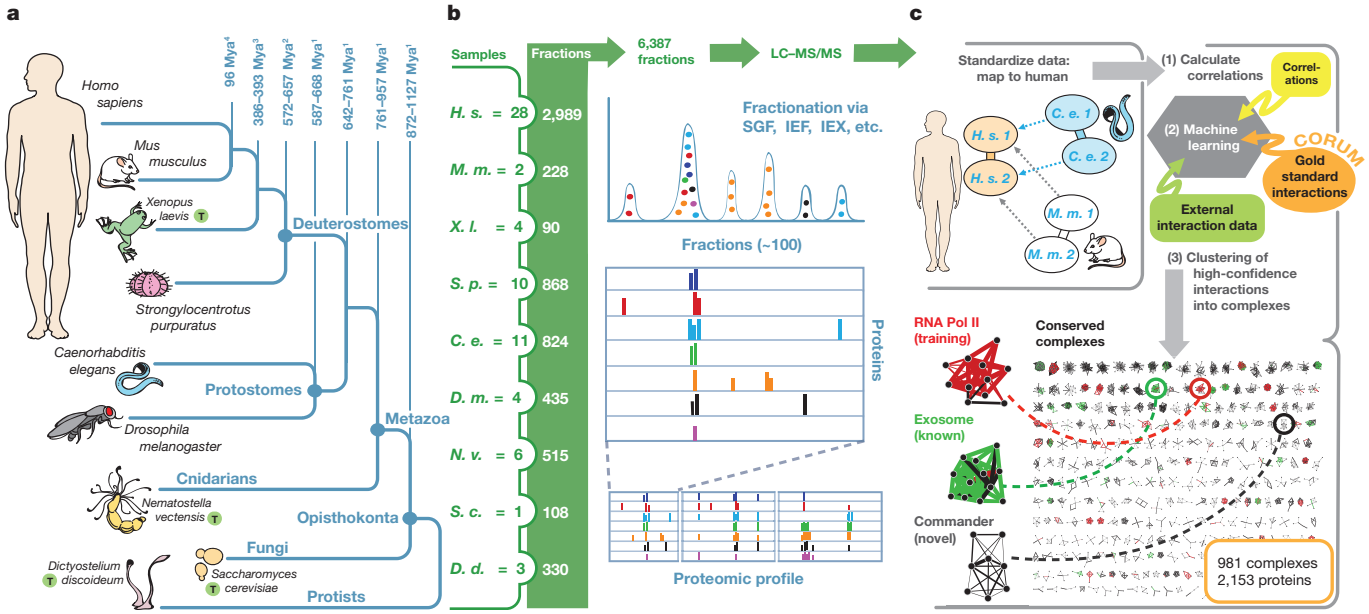


Figure 1 | Workflow. **a**, Phylogenetic relationships of organisms analysed in this study. We fractionated soluble protein complexes from worm (*C. elegans*) larvae, fly (*D. melanogaster*) S2 cells, mouse (*M. musculus*) embryonic stem cells, sea urchin (*S. purpuratus*) eggs and human (HEK293/HeLa) cell lines. Holdout species ('T', for test) likewise analysed were frog (*X. laevis*), an amphibian; sea anemone (*N. vectensis*), a cnidarian with primitive eumetazoan tissue organization; slime mould (*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 orthologue 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.

precision and recall as we required all derived interactions to have extensive biochemical support (see Supplementary Methods). Co-fractionation data of each input species affected 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 three or more (Extended Data Fig. 1b), enabling cross-species modelling and functional inference.

Benchmarking protein complexes

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 sixfold enriched (hypergeometric $P < 1 \times 10^{-24}$) for shared subcellular localization annotations in the Human Protein Atlas Database[21], 21-fold enriched ($P < 1 \times 10^{-56}$) 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 orthologues (interologues) in the four holdout species, as obtained by protein quantification across 1,127 biochemical fractions (see Supplementary Methods). 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 those of the input species used for learning (Fig. 2c). The biochemical data obtained for frog and sea anemone showed slightly better agreement than that for *Dictyostelium* and yeast that was proportional to evolutionary distance[24].

Besides indicating stably associated proteins, our multispecies biochemical profiles faithfully recapitulated the architecture of multiprotein complexes of known three-dimensional 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 sub-complex (Fig. 2d), reflecting known hierarchies of complex formation and disassembly.

Landscape of interaction conservation across species

Because most of the interacting components were phylogenetically conserved across vast evolutionary timescales, we were able to predict over one 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 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 probably reflects inherent clade diversity, especially in parasite genomes (for example, 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 (less than one-third 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 Supplementary Methods) that allowed proteins to participate in multiple complexes (that is, moonlighting) as merited (Extended Data Fig. 3a). The 981 putative multiprotein groupings (Fig. 3a; see Supplementary Table 4) include both

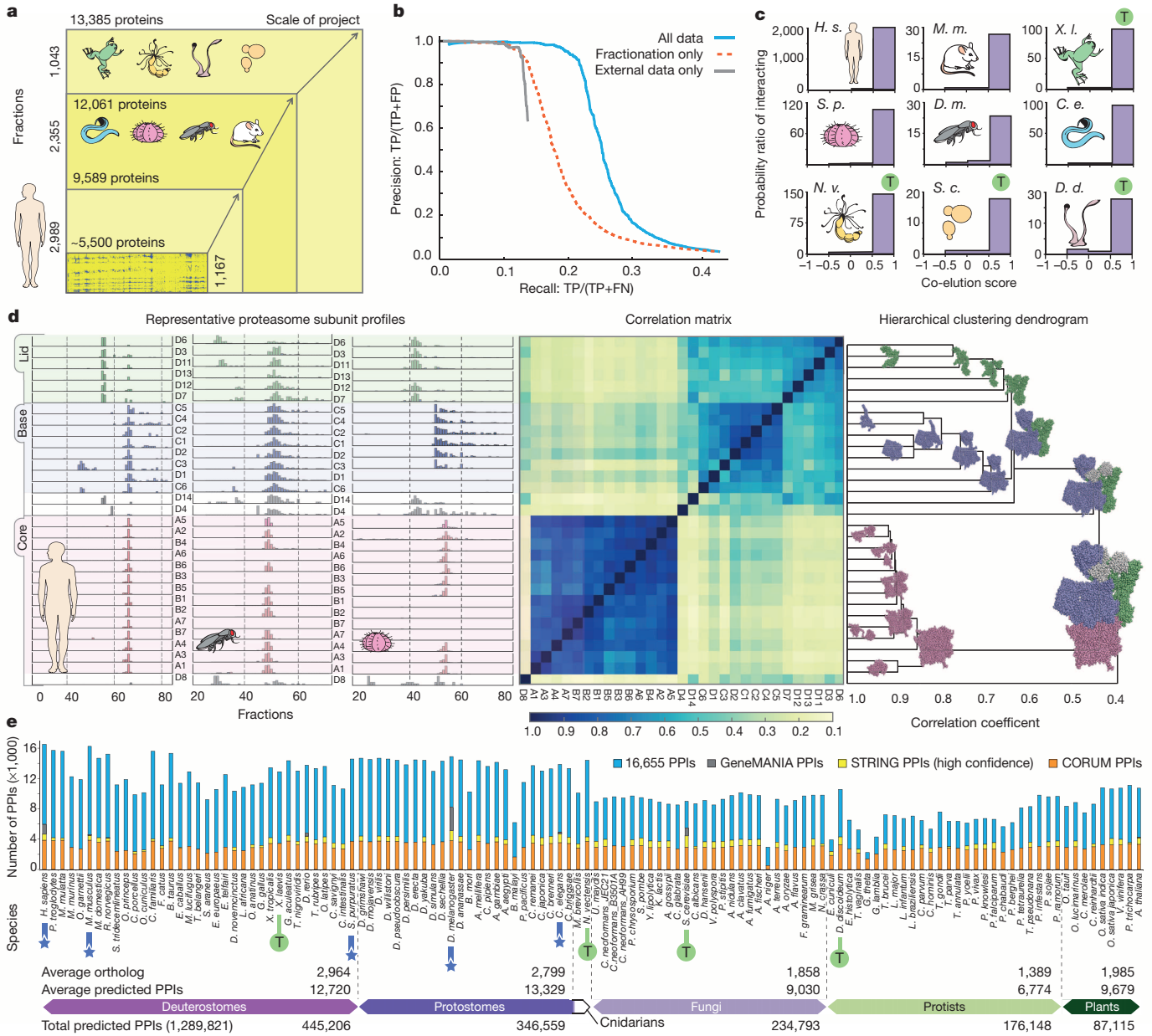


Figure 2 | Derivation and projection of protein co-complex associations across taxa. **a**, Expanded coverage via experimental scale-up relative to our previous human study[6]. Chart shows number of proteins detected, most (63%) in two or more species. **b**, Performance benchmarks, measuring precision 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; fivefold 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. TP, true positive; FP, false positive; FN, false negative. **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 (colours 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 sub-complexes. Hierarchical clustering (right) of pan-species pairwise Pearson correlation scores (centre) is consistent with accepted structural models (Protein Data Bank 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 occurrences across clades indicated at bottom.

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 (metazoan-specific, or 'new') to over one billion years (ancient, or 'old') of evolutionary divergence. Details of species, orthologues, taxonomic groups, protein ages and evolutionary distances are provided in Supplementary Tables 3 and 5 and Supplementary Methods.

Although proteins arising in metazoa (by gene duplication or other means) account for about three quarters of all human gene products, they form only about a third (39%; 147) of the clusters (Fig. 3a). These 'new' complexes tend to be smaller (≤3 components; Fig. 3b) and specific (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

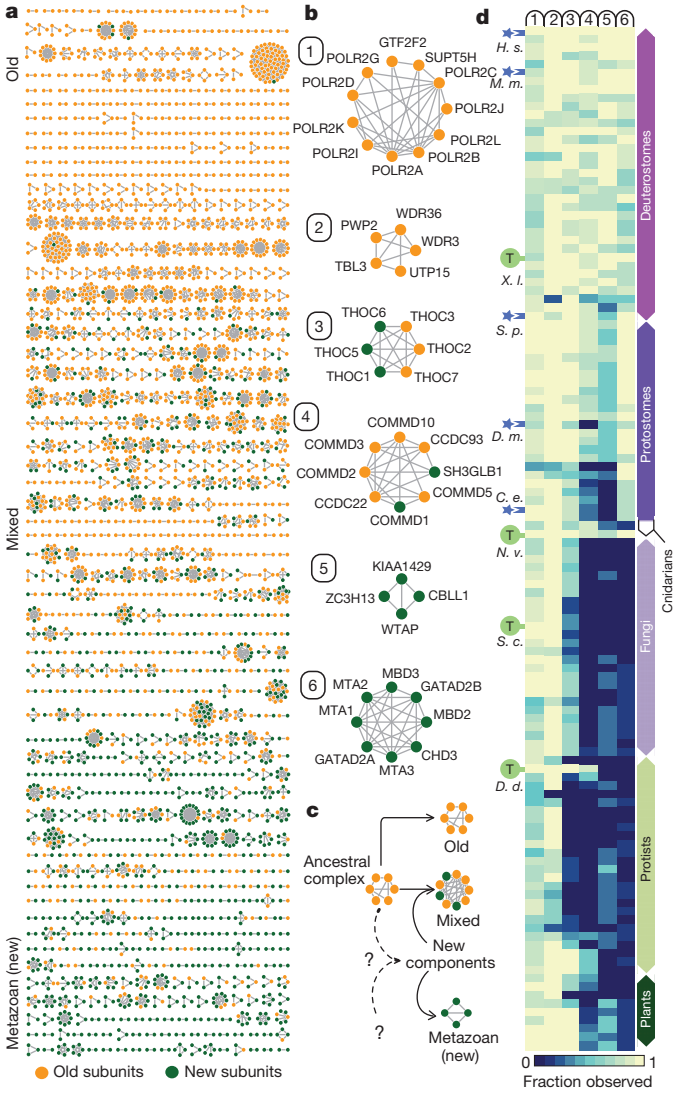


Figure 3 | Prevalence of conservation of protein complexes across Metazoa and beyond. **a**, Conserved multiprotein complexes, identified by clustering, arranged according to average estimated component age (see Supplementary Methods and ref. 25). Proteins (nodes) classified as metazoan (green) or ancient (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**, Heat map showing prevalence of selected complexes across phyla. Colour reflects fraction of components with detectable orthologues (absence, dark blue). Sea anemone (*N. vectensis*) is the most distant metazoan (cnidarian) analysed biochemically.

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 additional components to pre-existing macromolecules (Fig. 3c)[28].

These primordial complexes are present throughout the Opisthokonta supergroup (animals and fungi), estimated to be more than one 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 Supplementary Methods), suggesting multi-domain architectures underlie more transient or tissue-specific interactions. 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 $P \leq 1 \times 10^{-5}$) compared to strictly old (15%; 53/344; $P \leq 1 \times 10^{-3}$) 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).

Independent biological assessment

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 and 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 (E. L. Huttlin *et al.*, BioGRID preprint 166968), and observed a partial, but highly statistically 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 stoichiometries) and migration by size-exclusion chromatography (Fig. 4c and 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 un-annotated components (Fig. 4d and 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 website, http://metazoa.med.utoronto.ca/php/view_elution_image.php?id=71&cond=ms2).

We found an unexpected role in embryonic development for Commander, whose subunits are strongly co-expressed in developing frog (Extended Data Fig. 9d, e). *COMMD2/3*-knockdown (morpholino) tadpoles showed impaired head and eye development (Fig. 5a and Extended Data Fig. 9f, h), and defective neural patterning and expression changes in brain markers *PAX6*, *EN2* and *KROX20/EGR1* (Fig. 5b and Extended Data Fig. 9g, h). Given the recently discovered link[33,34] between *CCDC22* and 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 aetiology 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 orthologues 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 the *ZNF207* orthologue in *C. elegans* (*B0035.1*) enhanced lethality owing 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

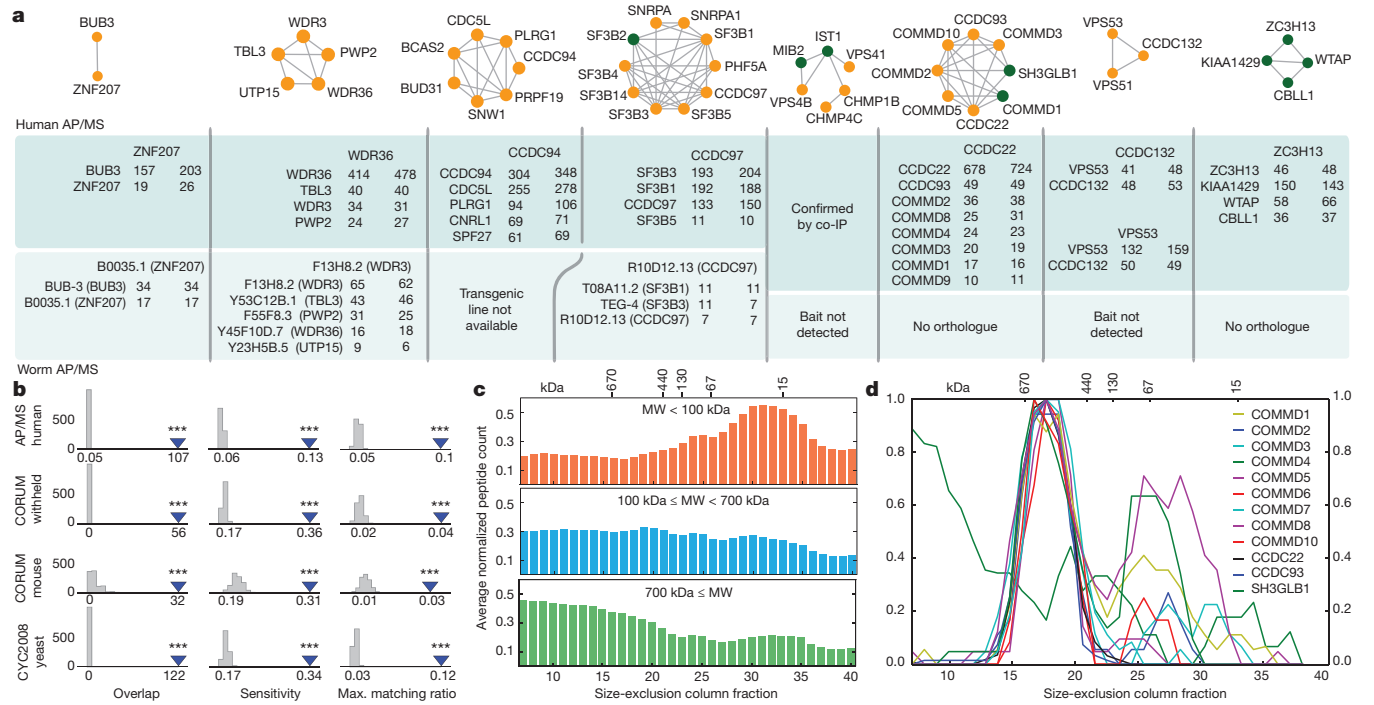


Figure 4 | Physical validation of complexes. **a**, Verification of complexes from tagged human cell lines and transgenic worms (see Supplementary Methods; complexes drawn as in Fig. 3). Inset reports spectral counts obtained in replicate AP/MS analyses of indicated bait protein (header). MIB2–VPS4 complex confirmed by co-immunoprecipitation (co-IP; Extended Data Fig. 6a). **b**, Conserved complexes significantly overlap large-scale AP/MS data reported for human cell lines (E. L. Huttlin *et al.*, BioGRID preprint 166968) to a comparable extent as literature reference sets[39,42], using three measures of complex-level agreement (see Supplementary Methods, Extended Data Fig. 6b); $***P < 0.001$, determined by shuffling (grey distributions). **c**, Agreement of inferred molecular weights (MW) of human protein complexes with size-exclusion chromatography profiles (data in **c**, **d**, from ref. 43). **d**, Co-elution of human Commander complex subunits by size-exclusion chromatography consistent with an approximately 500-kDa particle.

pre-mRNA splicing, CRISPR-based *CCDC97*-knockout human cells were slower growing than were control lines (Extended Data Fig. 10b, c) and hypersensitive to pladienolide B (Fig. 5d), a macrolide inhibitor of SF3b[36].

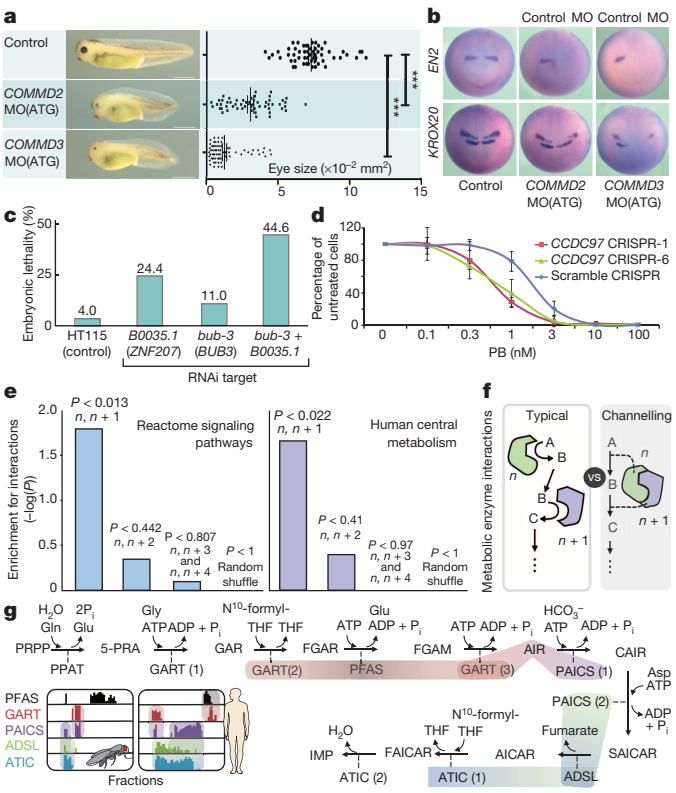


Figure 5 | Functional validation of complexes. **a**, Morpholino (MO(ATG)) targeting start codon to block translation) knockdown of *COMMD2* ($n = 55$ animals, 2 clutches, 1 eye each) or *COMMD3* ($n = 64$) in *X. laevis* embryos causes defective head and eye development (control $n = 57$; Extended Data Fig. 9f, h). $***P < 0.0001$, 2-sided Mann–Whitney test. **b**, *COMMD2/3* knockdown animals (five 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 (epistasis) following RNAi knockdown in *C. elegans* of *B0035.1* (*ZNF207*) and *bub-3* together (eggs laid: HT115, 1,308; *B0035.1*, 1,096; *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, and so on, as described in Supplementary Information. **f**, Metabolic channelling as opposed to traditional (typical) two-step cascade model. **g**, Conserved interactions among consecutively acting enzymes involved in purine biosynthesis (two representative co-fractionation profiles of the 69 total generated are shown).

Network perspective into conserved biological systems

Knowledge of conserved macromolecular associations provides a road map for additional functional inferences. For instance, fractionation profiles can be compared for any pair of proteins in our data set to search for evidence of interactions. 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), the significance of which decayed with more intervening steps (Fig. 5e). For example, strong consecutive

interactions were apparent within the widely conserved purine biosynthetic pathway, with enzymes (for example, PAICS, GART) eluting in two peaks (Fig. 5g), one coincident with the prior enzyme and the second with the downstream enzyme, suggestive of substrate channelling[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 extremely large set of supporting biochemical fractionation data (via ProteomeXchange with identifiers PXD002319–PXD002328), PPIs (via BioGRID; http://thebiogrid.org/185267/publication/) 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 analysed at the individual animal species or complex levels to assess the variety and functional adaptations of particular protein assemblies across phyla.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 December 2014; accepted 30 June 2015.
Published online 7 September 2015.

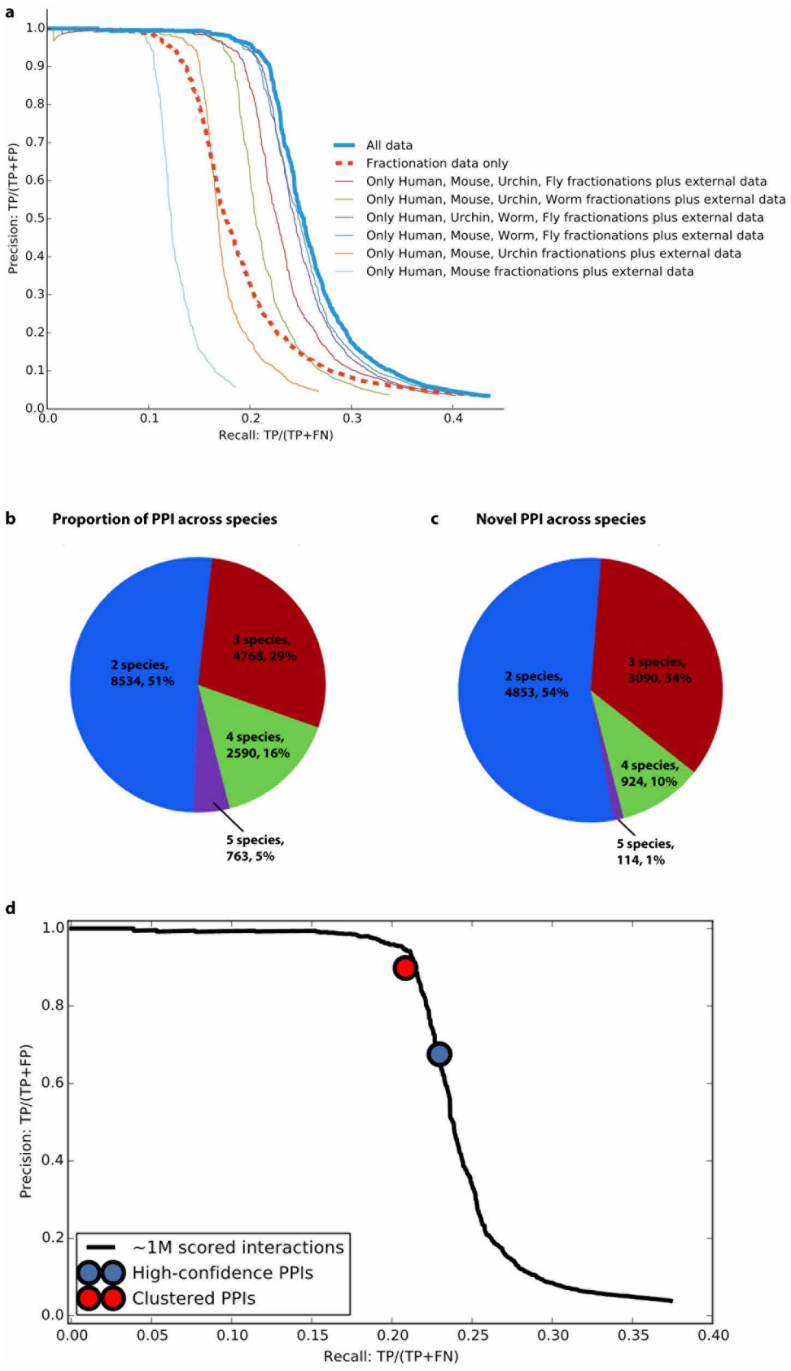
1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
2. Alberts, B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell* **92,** 291–294 (1998).
3. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433,** 531–537 (2005).
4. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643 (2006).
5. Guruharsha, K. G. *et al.* A protein complex network of *Drosophila melanogaster*. *Cell* **147,** 690–703 (2011).
6. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
7. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122,** 957–968 (2005).
8. Li, S. *et al.* A map of the interactome network of the metazoan C-elegans. *Science* **303,** 540–543 (2004).
9. Hu, P. *et al.* Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* **7,** e1000096 (2009).
10. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).
11. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA* **102,** 1974–1979 (2005).
12. Gandhi, T. K. B. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genet.* **38,** 285–293 (2006).
13. Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. Transcriptional regulation of protein complexes within and across species. *Proc. Natl Acad. Sci. USA* **104,** 1283–1288 (2007).
14. Singh, R., Xu, J. B. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA* **105,** 12763–12768 (2008).
15. Yu, H. *et al.* Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.* **14,** 1107–1118 (2004).
16. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8,** 565 (2012).
17. Kiemer, L. & Cesareni, G. Comparative interactomics: comparing apples and pears? *Trends Biotechnol.* **25,** 448–454 (2007).
18. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
19. Malovannaya, A. *et al.* Analysis of the human endogenous coregulator complexome. *Cell* **145,** 787–799 (2011).
20. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21,** 1109–1121 (2011).
21. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnol.* **28,** 1248–1250 (2010).
22. McKusick, V. A. *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*. (Johns Hopkins Univ. Press, 1998).
23. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
24. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
25. Bezginov, A., Clark, G. W., Charlebois, R. L., Dar, V. U. N. & Tillier, E. R. M. Coevolution reveals a network of human proteins originating with multicellularity. *Mol. Biol. Evol.* **30,** 332–346 (2013).
26. Stumpf, M. P. H. *et al.* Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA* **105,** 6959–6964 (2008).
27. Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7,** 120 (2006).
28. Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91,** 138701 (2003).
29. Knoll, A. H. The early evolution of eukaryotes: a geological perspective. *Science* **256,** 622–627 (1992).
30. Burstein, E. *et al.* COMMD proteins, a novel family of structural and functional homologs of MURR1. *J. Biol. Chem.* **280,** 22222–22232 (2005).
31. van de Sluis, B., Rothuizen, J., Pearson, P. L., van Oost, B. A. & Wijmenga, C. Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Hum. Mol. Genet.* **11,** 165–173 (2002).
32. McDonald, F. J. 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.* **305,** C129–C130 (2013).
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.* **109,** 1–6 (2014).
34. Voineagu, I. *et al.* CCDC22: a novel candidate gene for syndromic X-linked intellectual disability. *Mol. Psychiatry* **17,** 4–7 (2012).
35. Toledo, C. M. *et al.* BuGZ is required for Bub3 stability, Bub1 kinetochore function, and chromosome alignment. *Dev. Cell* **28,** 282–294 (2014).
36. Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature Chem. Biol.* **3,** 570–575 (2007).
37. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–D477 (2014).
38. Ovádi, J. *Cell Architecture and Metabolite Channeling*. (RG Landes Company, 1995).
39. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res.* **38,** D497–D501 (2010).
40. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38,** W214–W220 (2010).
41. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–D815 (2013).
42. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37,** 825–831 (2009).
43. Kirkwood, K. J., Ahmad, Y., Larance, M. & Lamond, A. I. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Mol. Cell. Proteomics* **12,** 3851–3873 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Bader, P. Kim, G. Moreno-Hagelsieb, S. Pu and S. Wodak for critical suggestions, illustrator A. Syrett for expert help drafting figures, T. Kwon (University of Texas) for *X. laevis* gene models, and K. Foltz (University of California, Santa Barbara), A. Brehm (Philipps-University Marburg), P. Paddison (Fred Hutchinson Cancer Research Center), J. Smith (Woods Hole Marine Biological Laboratory), P. Zandstra and J. Moffat (University of Toronto) for providing biological specimens and reagents. We thank members of the Emili and Marcotte laboratories for 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 CIHR, NSERC, ORF and the CFI to A.E., from the CIHR and Heart and Stroke to J. P., from the NIH (F32GM112495) to K.D., and from the NIH, NSF, CPRIT, and Welch Foundation (F-1515) to E.M.M.

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

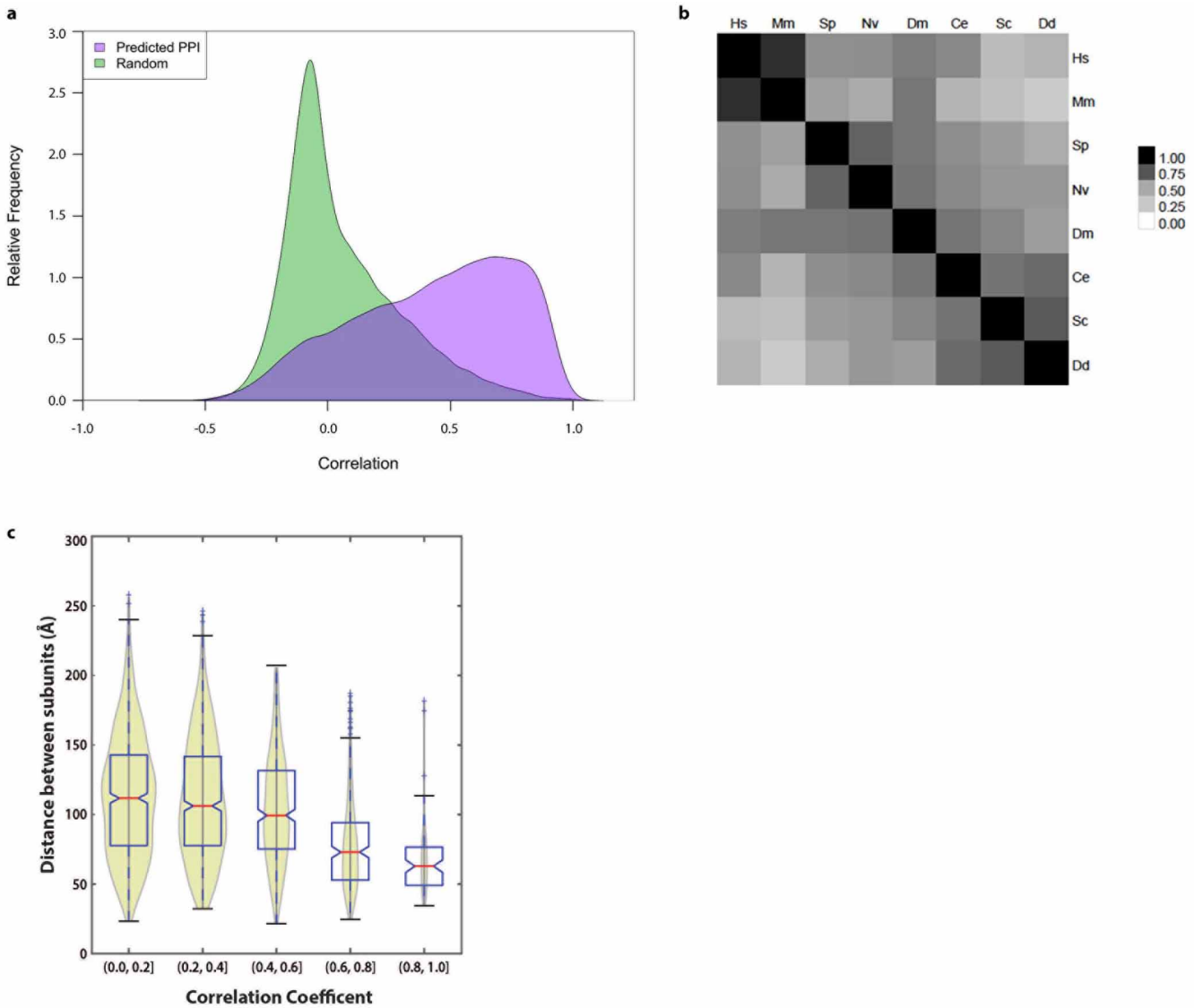
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.M.M. (marcotte@icmb.utexas.edu) or A.E. (andrew.emili@utoronto.ca).



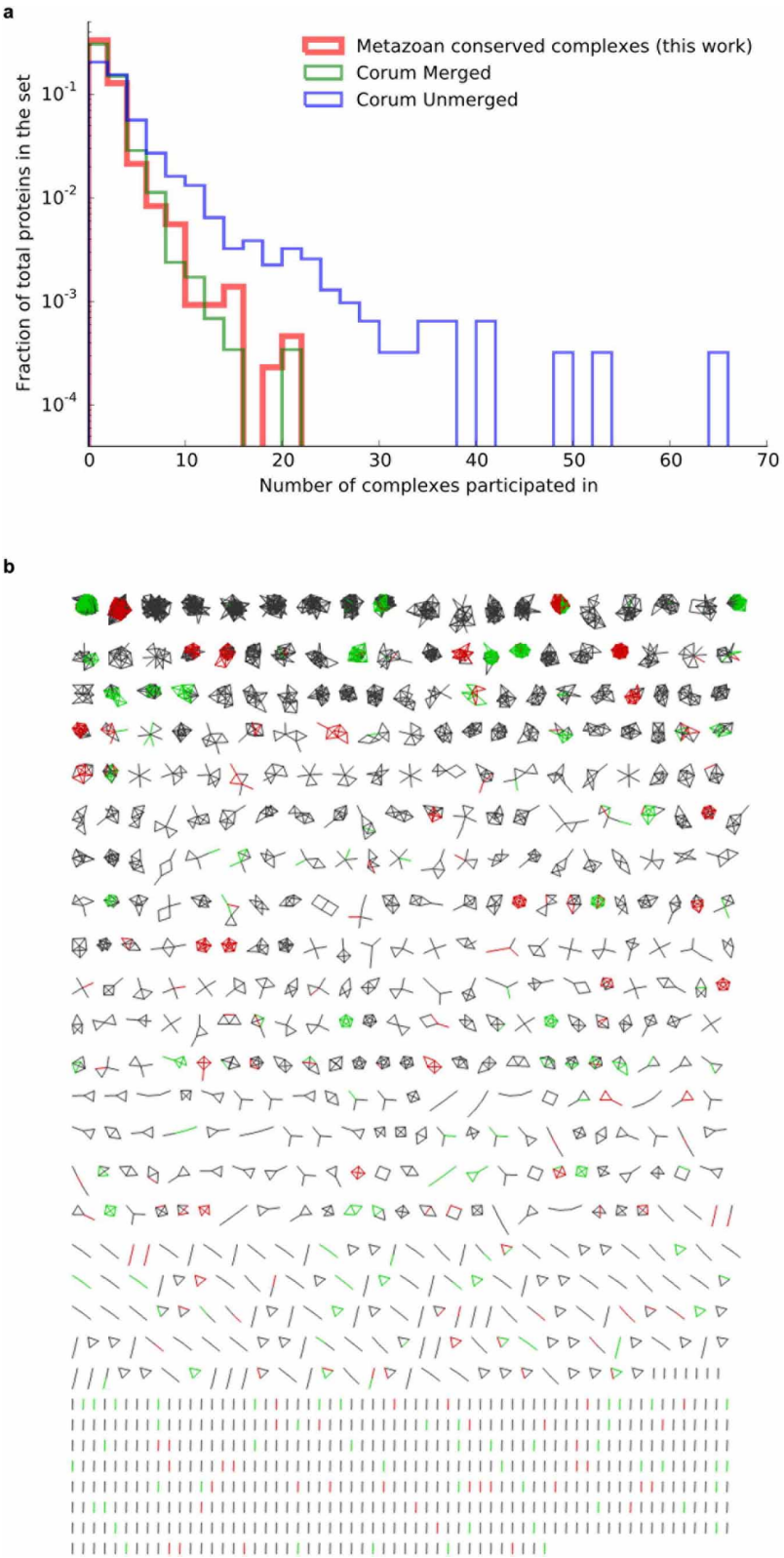
Extended Data Figure 1 | Performance measures. **a**, Performance benchmarks, measuring the precision and recall of our method and data in identifying known co-complex interactions from a withheld reference set of annotated human complexes (from CORUM[39]; as in Fig. 2b). Fivefold cross-validation against 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 five input species: human, mouse, urchin, fly and worm; the latter curve omits all external data. In all cases, at least two species were required to show supporting biochemical evidence. Recall refers to the 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 iRefWeb[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 display evidence of co-fractionation in three or more species. **d**, Final precision/recall performance on withheld interaction test set. A support vector machine classifier was trained using interactions derived from our training set of CORUM complexes, then ~1 million protein pairs found to co-elute in at least two of the five input species were scored by the classifier. Black curve shows precision and recall for ranked list of co-eluting pairs, with recall representing the 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 Supplementary Methods, yielding a final set of 7,669 interactions, which form the 981 identified complexes (red circle; precision = 90.0%, recall = 20.8%).

44. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* **2010,** baq023 (2010).

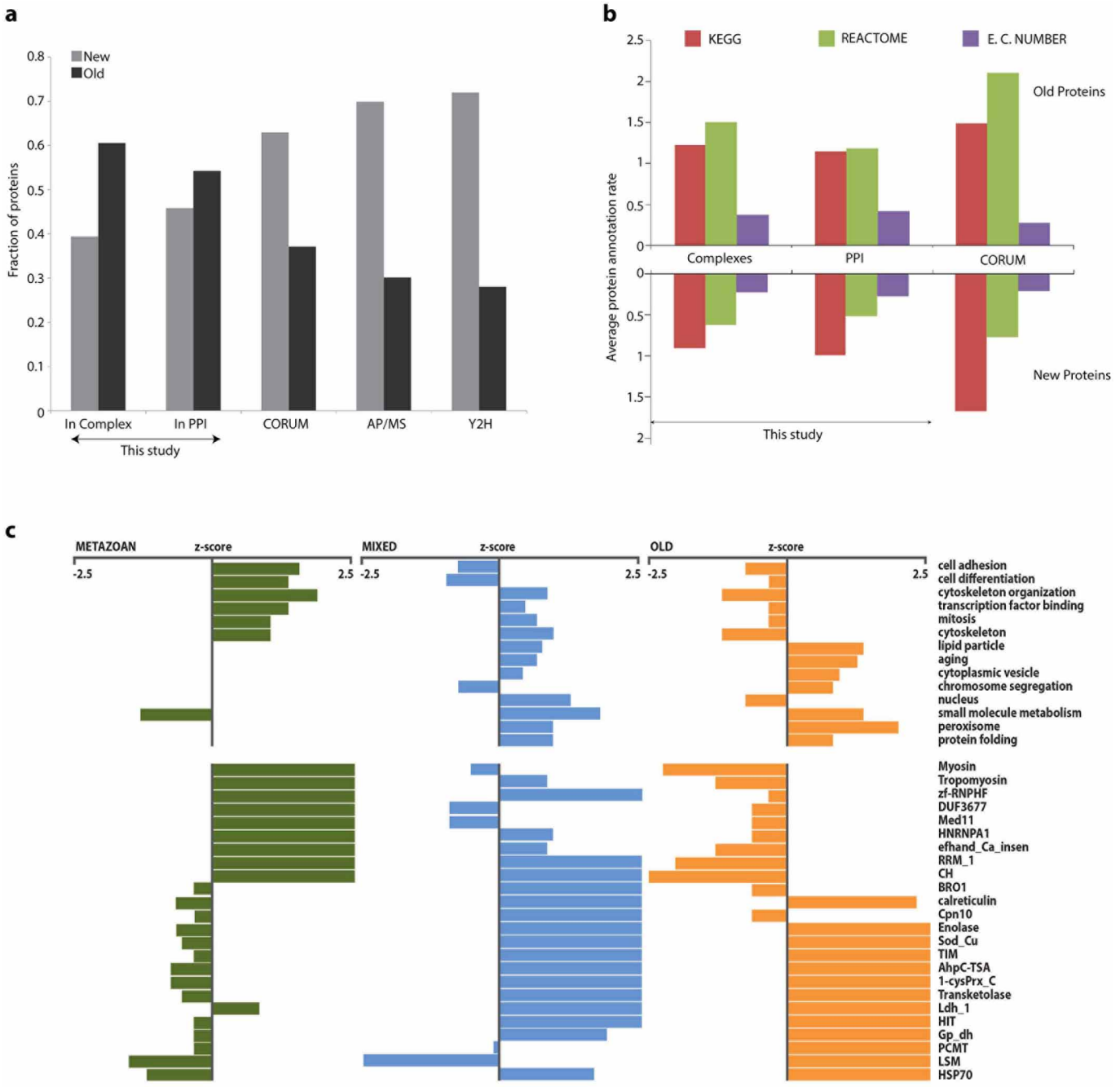
45. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–D539 (2006).



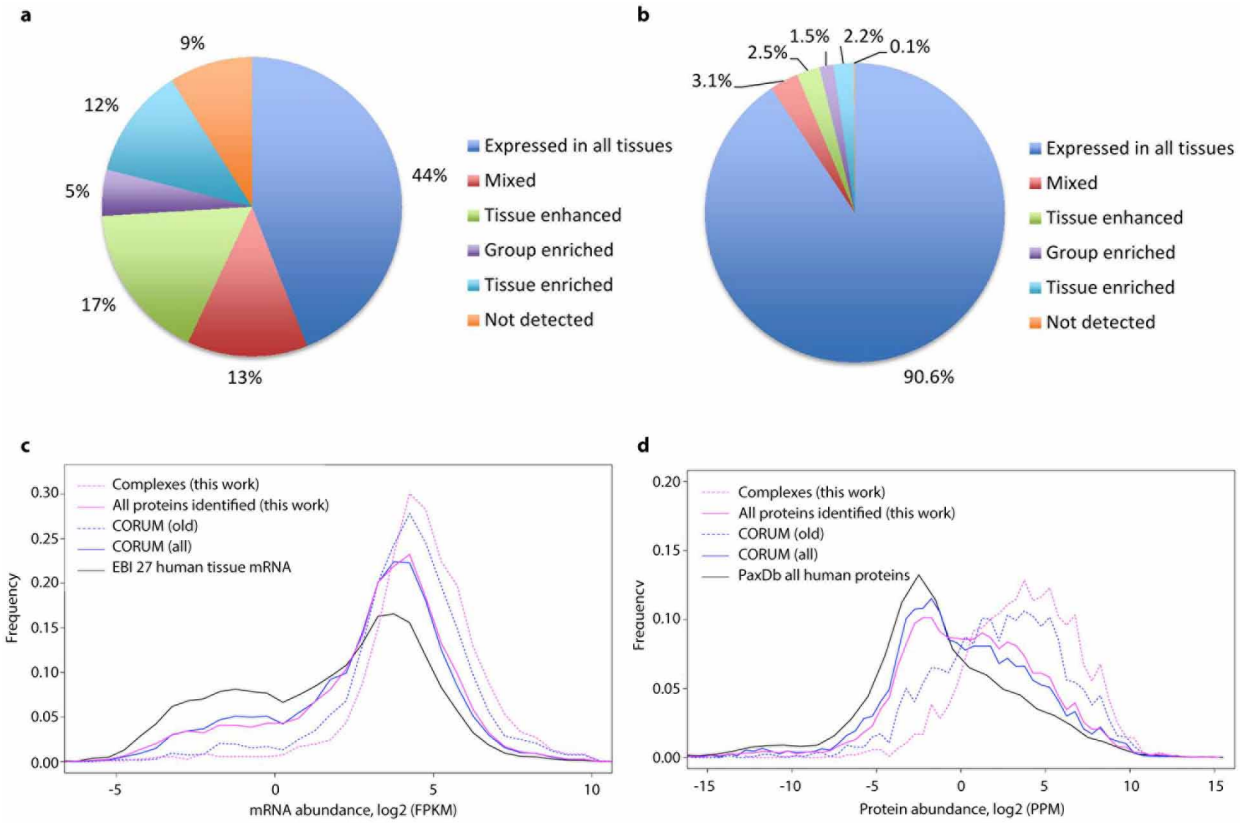
Extended Data Figure 2 | Properties of protein elution profiles. **a**, Distribution of global protein tissue expression pattern similarity, measured as the Pearson correlation coefficient of protein abundance across 30 human tissues[23], showing markedly higher correlations for 16,468 protein–protein 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**, Heat map 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 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 three-dimensional structures (see Supplementary 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 Supplementary 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. Grey interactions were not previously annotated in CORUM.

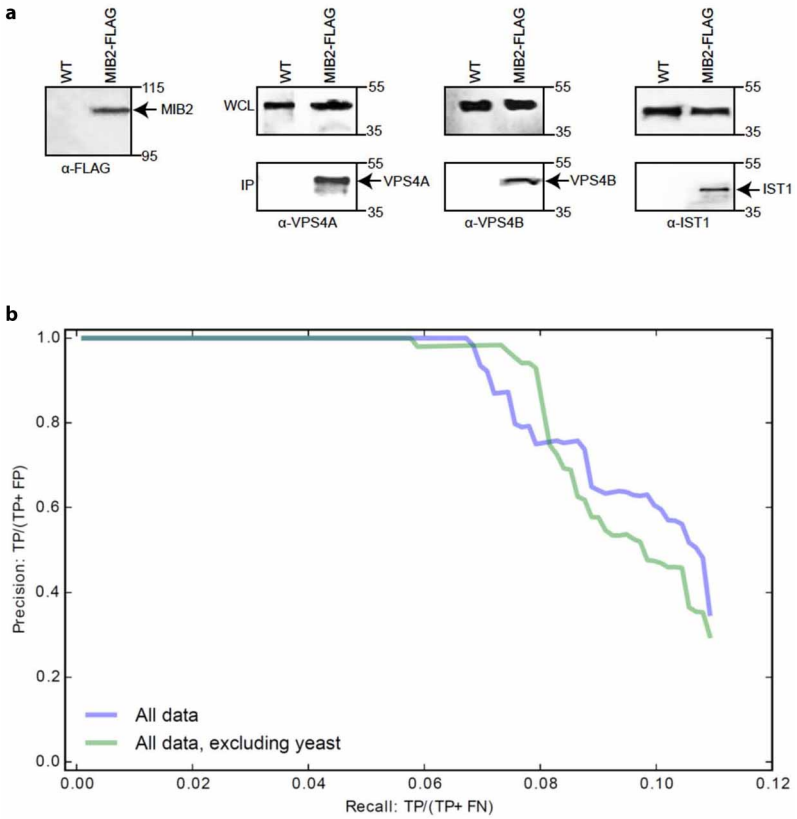


Extended Data Figure 4 | Properties of new and old proteins and complexes. **a**, The 2,153 protein components in the conserved animal complexes tend to be more ancient than the 2,301 proteins reported in the CORUM reference complexes or in two recent large-scale protein interaction assays, based on either the 7,062 proteins found by affinity purification/mass spectrometry (AP/MS; E. L. Huttlin *et al.*, BioGRID preprint 166968, http://thebiogrid.org/166968/publication/) or the 3,667 proteins analysed by yeast two-hybrid assays (Y2H)[10]. Ages are derived from OMA (Orthologous Matrix database) 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 PPIs, 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).

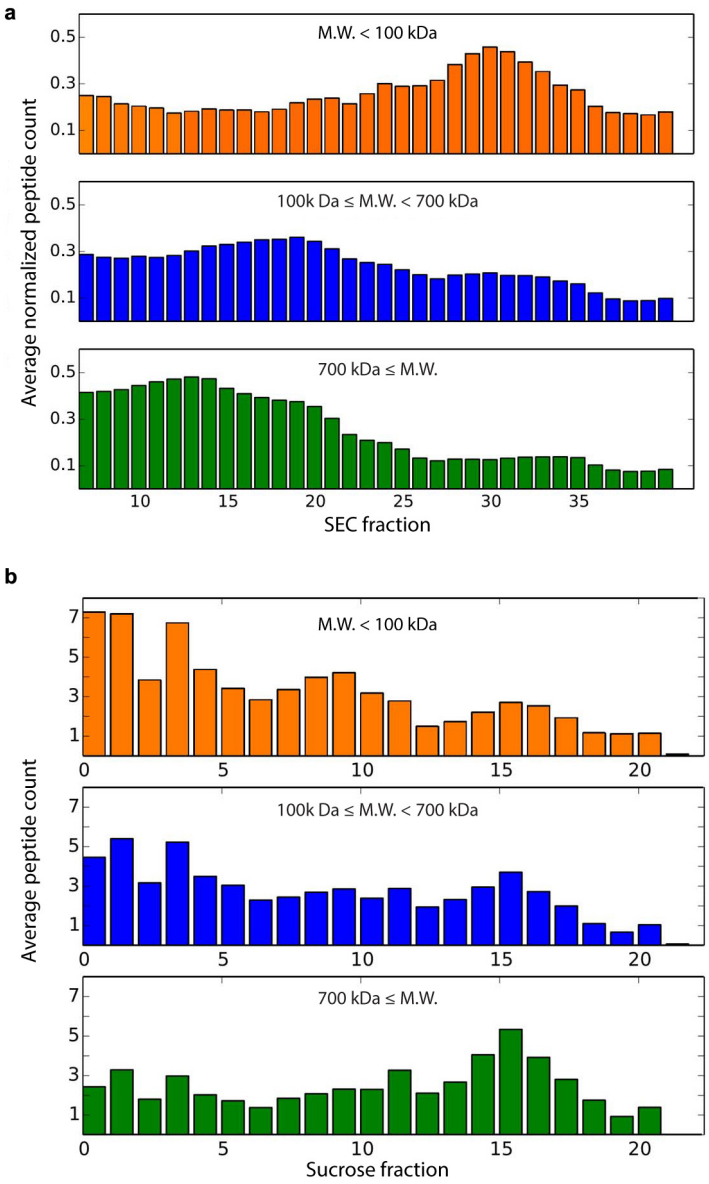


Extended Data Figure 5 | Abundance and expression trends for proteins in complexes. Proteins within the identified complexes tend to be ubiquitously expressed across human tissues. **a**, **b**, Pie charts show the proportions of proteins with varying tissue expression patterns, from a recently published human tissue proteome map[46], comparing the full set of 20,258 human proteins (**a**) with the 2,131 proteins within the identified complexes (**b**). 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 $P < 0.001$). **c**, **d**, The distributions of average mRNA (**c**, data from EBI accession E-MTAB-1733) and protein (**d**, data from PaxDb integrated data set, 9606-H.sapiens_whole_organism-integrated_data set) abundances for all proteins identified and those within complexes. Evolutionarily old proteins (defined by OMA as described in ref. 25 and mentioned earlier) tend towards higher abundances, even for proteins in reference complexes.

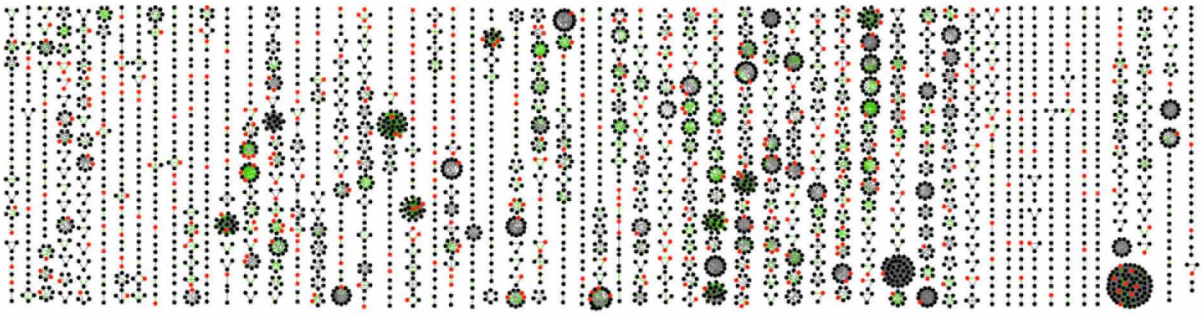
46. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 6220 (2015).



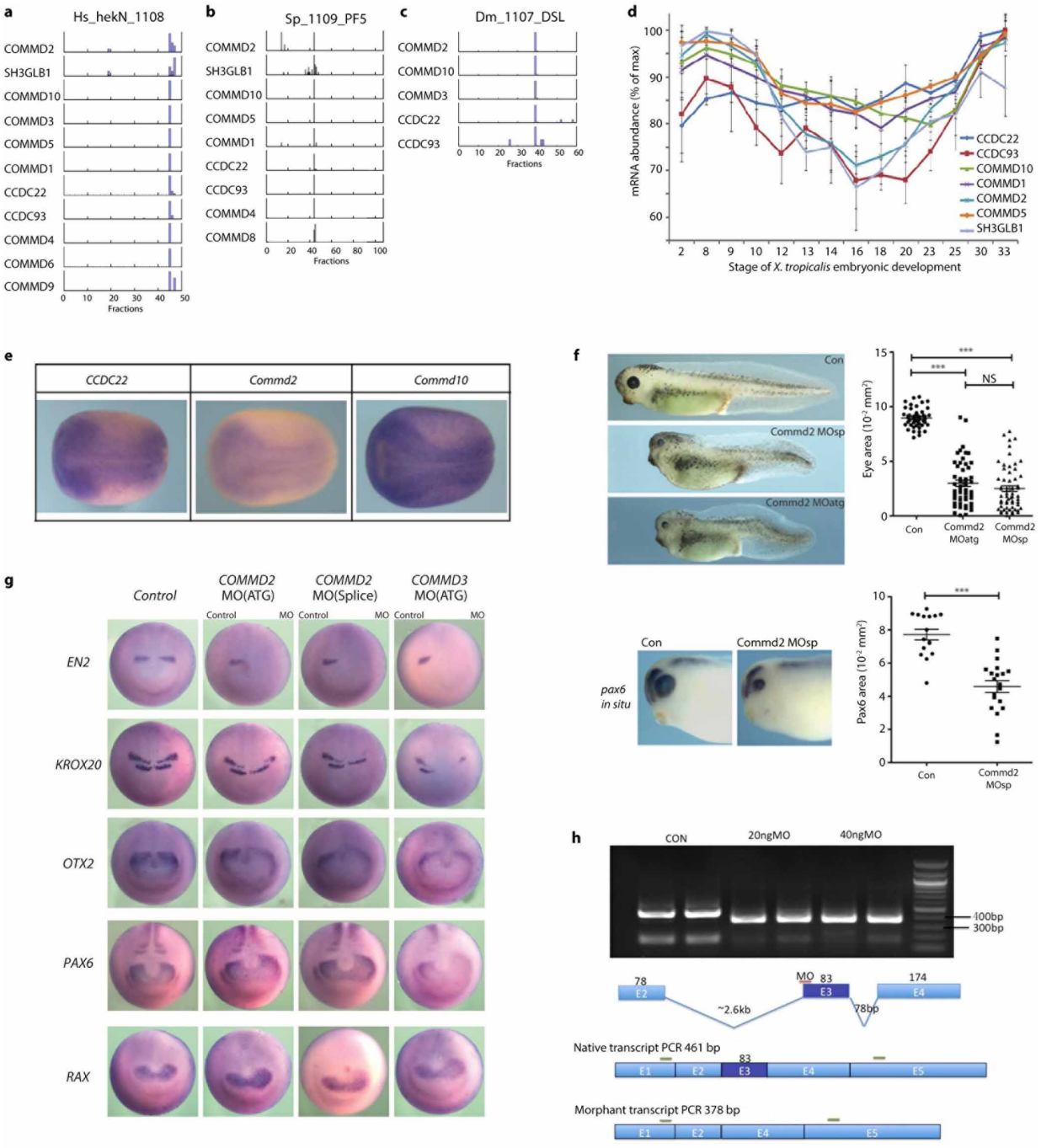
Extended Data Figure 6 | Additional validation data. **a**, Confirmation of MIB2 interactions by co-immunoprecipitation. Extract (~10 mg protein) from cultured human HCT116 cells expressing Flag-tagged MIB2 or control (WT) cells was incubated with 100 µl anti-Flag M2 resin for 4 h while gently rotating 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 Supplementary Information). **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 weight of each complex was calculated as the sum of putative component molecular weights, assuming 1:1 stoichiometry. Data from ref. 43 were analysed as in Fig. 4c and show a similar trend as for the derived complexes. **b**, Derived complexes' inferred molecular weights are broadly consistent with their components' average cumulative ultracentrifugation profiles on a sucrose density gradient. Average profiles are plotted for *X. laevis* orthologues, based on a preparation of haemoglobin-depleted heart and liver proteins separated on a 7–47% sucrose density gradient, as described in the Supplementary Methods.

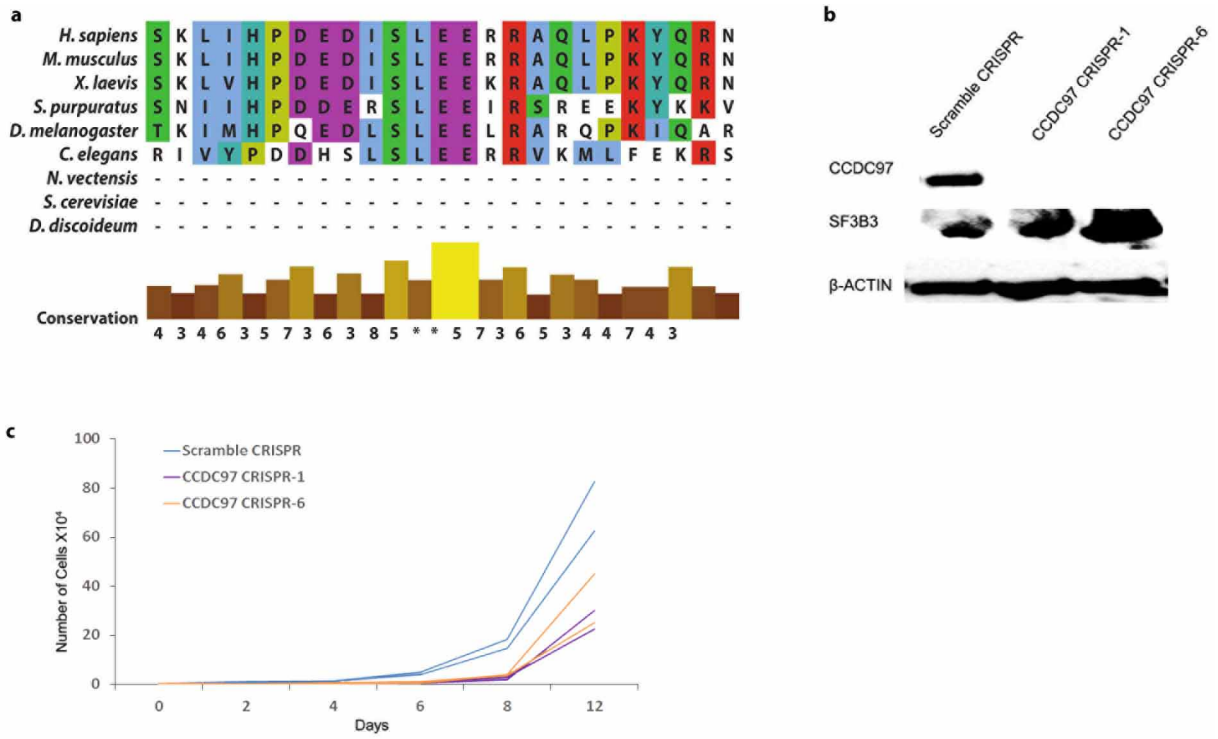


Extended Data Figure 8 | Distribution of uncharacterized proteins and novel interactions across the 981 derived complexes. Complexes were sorted by median age (defined by OMA). Among 2,153 unique proteins, 293 (red) lack Gene Ontology (GO) functional annotations, while 1,756 of 7,665 co-complex interactions are novel (light green) (not listed in iRefWeb 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: HEK293 cell nuclear extract (**a**); sea urchin embryonic (5 days post-fertilization) extract (**b**); and fly SL2 cell nuclear extract (**c**); each fractionated by heparin affinity chromatography. **d**, Co-expression of Commander complex subunits during embryonic development of *X. tropicalis* (plotting mean ± s.d. of three clutches; data from ref. 49). **e**, Messenger RNA 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 (three 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 three clutches; control $n = 47$ animals, one eye each; $^{***}P < 0.0001$, two-sided Mann–Whitney test); 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). Bottom, *Commd2*-knockdown induced altered *Pax6* patterning in the embryonic eye (control $n = 8$ animals, two 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; five 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. Notably, 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.

47. de Bie, P. *et al.* Characterization of COMMD protein–protein interactions in NF-κB signalling. *Biochem. J.* **398,** 63–71 (2006).
48. Phillips-Krawczak, C. A. *et al.* COMMD1 is linked to the WASH complex and regulates endosomal trafficking of the copper transporter ATP7A. *Mol. Biol. Cell* **26,** 91–103 (2015).
49. Yanai, I., Peshkin, L., Jorgensen, P. & Kirschner, M. W. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. *Dev. Cell* **20,** 483–496 (2011).



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 Supplementary Information). **c**, Loss of *CCDC97* impairs cell growth. Lines show growth curves of control versus knockout cell lines in two biological replicate assays.