Ancient eukaryotic protein interactions illuminate modern genetic traits and disorders

Rachael M. Cox[1], Ophelia Papoulas[1], Shirlee Shril[2], Chanjae Lee[1], Tynan Gardner[1], Anna M. Battenhouse[1], Muyoung Lee[1], Kevin Drew[3], Claire D. McWhite[4], David Yang[1], Janelle C. Leggere[1], Dannie Durand[5], Friedhelm Hildebrandt[2], John B. Wallingford[1], Edward M. Marcotte[1]

[1]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712, USA

[2]Division of Nephrology, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA 02215, USA

[3]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA

[4]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

[5]Department of Biological Sciences, Carnegie Mellon University, 4400 5th Avenue Pittsburgh, PA 15213, USA

ORCIDs: R.M.C., 0000-0001-5407-1101; O.P., 0000-0002-6370-0616; C.L., 0000-0001-8748-1369; A.M.B., 0000-0002-7455-9064; M.L., 0000-0002-4913-6289; K.D., 0000-0002-1260-4413; C.D.M., 0000-0001-7346-3047; J.C.L., 0000-0002-5225-6046; D.D., 0000-0002-3505-6640; F.H., 0000-0002-7130-0030; J.B.W., 0000-0001-8701-4293; E.M.M., 0000-0001-8808-180X

*Correspondence to: E.M.M., marcotte@utexas.edu

SUMMARY

All eukaryotes share a common ancestor from roughly 1.5 - 1.8 billion years ago, a single-celled, swimming microbe known as LECA, the Last Eukaryotic Common Ancestor. Nearly half of the genes in modern eukaryotes were present in LECA, and many current genetic diseases and traits stem from these ancient molecular systems. To better understand these systems, we compared genes across modern organisms and identified a core set of 10,092 shared protein-coding gene families likely present in LECA, a quarter of which are uncharacterized. We then integrated >26,000 mass spectrometry proteomics analyses from 31 species to infer how these proteins interact in higher-order complexes. The resulting interactome describes the biochemical organization of LECA, revealing both known and new assemblies. We analyzed these ancient protein interactions to find new human gene-disease relationships for bone density and congenital birth defects, demonstrating the value of ancestral protein interactions for guiding functional genetics today.

INTRODUCTION

The last eukaryotic common ancestor (LECA), existing approximately 1.5 to 1.8 billion years ago, was the unicellular ancestor of all extant eukaryotes [1,2]. The molecular makeup of this basal eukaryote may offer insights into the ancient genetic innovations that gave rise to the vast eukaryotic cellular complexity we observe today. As such, understanding the core genetic toolkit of LECA has been a long-standing goal in genetics and evolutionary biology.

Previous phylogenomic analyses are in strong agreement that LECA was highly complex and almost certainly contained most hallmarks of the eukaryotic cell. Synthesis of these studies suggests LECA had at least one nucleus [3,4] with linear chromosomes and centromeres [5]; an interconnected endomembrane system comprised of an endoplasmic reticulum [6], Golgi apparatus [7], vesicle trafficking system [8], and nuclear envelope [9]; a dynamic actin- and tubulin-based cytoskeleton [10] including pseudopodia [11], centrioles [12], and at least one cilium [13]; distinct degradative vesicles such as lysosomes [14] and peroxisomes [15]; and mitochondria capable of both aerobic and anaerobic respiration [2,16].

While there are many partial descriptions of the genetic content of LECA, systematic reconstructions of LECA genes are sparse [17–20]. Moreover, there is as yet no integrated picture of LECA's proteome or how these proteins interact in higher-order assemblies. Such interactome data would be useful because nearly every significant

cellular process appears to rely on assemblies of proteins working together [21], often organized into extended networks. Thus, an interactome for LECA would provide a richer portrait of its genetics and biochemistry than is possible from genomic reconstructions alone. Moreover, since proteins are the primary drivers of molecular phenotype [22], protein interaction networks are valuable tools for inferring protein functions and uncovering genotype-to-phenotype relationships of medical and agricultural interest.

Previous efforts to map eukaryotic protein interactomes systematically have primarily concentrated on opisthokonts (*i.e.*, animals and fungi) [23–29] with the exception of a handful of studies in plants (e.g. [30–32]) and protists [33,34]. However, comparing protein networks across more divergent species could provide many new insights not only into evolutionary conservation, but also the divergence of specific biological processes. For example, conserved cellular machinery can be used in different organism-specific contexts to produce distinct organism-specific phenotypes (e.g. the “phenolog hypothesis” [35]).

Here, we sought to reconstruct the LECA protein interaction network and to use this network to illuminate the etiology of human genetic disease. We first derived a conservative estimate of the LECA protein-coding gene set. We then integrated ~26,000 mass spectrometry experiments across 31 eukaryotes spanning ~1.8 billion years of evolution to generate a draft map of protein interactions present in LECA and widely conserved across the eukaryotic tree of life. We then used directed analysis of frogs, mice, and humans to demonstrate that these ancient interactions are enriched for human disease-linked proteins. Thus, the LECA interactome serves as both a guide for better understanding the deep origins of eukaryotic systems and as a framework for identifying new disease associations for human proteins.

RESULTS AND DISCUSSION

Inferring the gene content of LECA, the last eukaryotic common ancestor

We first determined a set of 10,092 groups of orthologous genes tracing back to the last eukaryotic common ancestor using Dollo parsimony [36] on reference proteomes from 156 species (see **Methods** and supporting Zenodo data repository). These were defined using the eggNOG algorithm [37] and are referred to hereafter as LECA OrthoGroups (OGs) (**Table S1**). Within these, we recover genes consistent with the suite of eukaryotic features inferred from ancestral trait reconstructions. We mapped these LECA OGs to known functional annotations [37]. **Figure 1** highlights LECA OG functions associated with the nucleus, endoplasmic reticulum, Golgi apparatus,

endosomes, digestive vesicles, transport vesicles, secretory vesicles, mitochondria, cilium and an extensive cytoskeleton likely capable of cell projection.

A substantial proportion of functionally annotated LECA OGs are dedicated to DNA replication/repair, transcription, translation, and RNA processing (~25%), underscoring eukaryote-specific innovations related to the segregation of transcription and translation within the nucleus and cytoplasm, respectively [38]. Examples of these innovations are the spliceosome, nuclear envelope, nuclear pore complex, and transport factors such as karyopherins regulated by the Ran-GTP system [9,39]. Another large group of genes (~30%) reflects the expansion and specialization of proteins and pathways related to the compartmentalization of energy production, energy conversion, and metabolism [40]. For example, we recover all V-ATPase subunits, a conserved protein complex responsible for the acidification of lysosomes and peroxisomes. Additionally, we traced mitochondrial-specific genes such as TIM and TOM translocases to LECA, in addition to a large suite of mitochondrial carrier family (MLF) proteins and generalist solute carrier (SLC) family proteins. Continuous MLF- and SLC-mediated transport of myriad metabolites across the mitochondrial membrane enables multiple modes of compartmentalized energy conversion [41], suggesting that LECA had already evolved sophisticated discriminatory pathways necessary for precise partitioning of varied substrates.

LECA also possessed a rich system of endomembranes, with ~15% of LECA OGs associated with membrane biogenesis, intracellular trafficking, and signal transduction. These include endosomal coat proteins such as clathrin, adaptin, and the COPI and COPII vesicle coat complexes which facilitate vesicle budding from the Golgi apparatus and endoplasmic reticulum, respectively [8]. Proteins related to signal transduction underwent a massive expansion at the root of eukaryotes, particularly within GTPases, such as the Ras, Ran, Rho, Rab, Arf, and dynamin superfamilies [42], highlighting the complexity and diversity of early signaling pathways and their relationships with endomembranes. Surprisingly, using eggNOG functional annotations we only found ~300 LECA OGs (~3%) associated with the cytoskeleton and cell motility (*i.e.*, a cilium) while previous studies reported ~500 OGs [43,44].

Finally, the eggNOG algorithm identified 2,387 LECA OGs (~25%) of unknown function (**Figure 1, bottom right**), the majority of which are absent from the human genome. The high proportion of uncharacterized genes and low count of ciliary genes prompted us to attempt to assign function, or at a minimum, subcellular location, to LECA OGs by using annotations for specific extant proteins in these families present in the UniProt database [45] (see **Methods** and Zenodo repository). This allowed us to recover ~200 LECA OGs associated with cytoskeletal and cell motility pathways that

were previously classified as “function unknown” by the eggNOG algorithm, bringing the total up to the ~500 expected OGs based on previous literature [43,44]. In sum, we assign tentative UniProt annotations and subcellular compartments to 1,066 of 2,387 LECA OGs originally categorized as “function unknown” by eggNOG.

Overall, slightly less than half of human genes (n = 9,908) map to 4,777 unique LECA OGs, and these differences can be instructive. For example, the human genome appears to have lost a large number of genes associated with carbohydrate metabolism, amino acid metabolism, and membrane biogenesis, including those for isocitrate lyase (ICL) and malate synthase, enzymes of the glyoxylate cycle. Glyoxylate is a highly reactive aldehyde and glyoxylate cycles have been demonstrated in nearly every other branch of life outside of mammals [46,47]. Because of malate synthase loss, human cells must rely on other enzymes to neutralize glyoxylate: (alanine-glyoxylate aminotransferase (AGT) in the peroxisome or glyoxylate reductase (GRHPR) in the cytosol). Defects in either human enzyme produce primary hyperoxaluria, *i.e.*, renal disease caused by the failure to detoxify glyoxylate [48]. This example highlights how comparative evolution informs mechanisms of human disease.

Mapping the LECA protein interactome

We next sought to determine the higher-order organization of the LECA proteome using co-fractionation mass spectrometry (CFMS). CFMS is a high-throughput method for measuring protein-protein interactions (PPIs) *en masse* from virtually any species or cell line without recombinant tags or affinity reagents (**Figure 2**) [26,49]. CFMS reduces the complexity of a native protein lysate by gently separating protein complexes based on properties such as size, charge, or hydrophobicity, and utilizes the premise that stably interacting proteins will generally co-elute together irrespective of the native separation method used. While results from a single CFMS experiment are insufficient to draw reliable conclusions about specific protein-protein interactions, integrating observations from orthogonal separations from multiple species, cell types, and tissues confers strong statistical power for inferring conserved interactions [27,29,31].

To detect conserved protein interactions, we integrated raw CFMS data from more than 10,000 individual biochemical fractions [26,27,31,33,34] across 31 diverse eukaryotic species spanning ~1.8 billion years of evolution (**Figure 2B**). In addition to external data sets, we generated new CFMS data for four minimally characterized and phylogenetically diverse unicellular eukaryotes: *Brachionus rotundiformis* (Amorphea, rotifer), *Euglena gracilis* (Excavata, algae), *Phaeodactylum tricornutum* (TSAR, diatom), and *Tetrahymena thermophila* (TSAR, ciliate). Because LECA was ciliated, we expanded our coverage of ciliary proteomes by collecting CFMS data from *Sus scrofa*

(Amorphea, pig) tracheal tissue and *Xenopus laevis* (Amorphea, frog) sperm. Experimental details concerning cell types, tissues, developmental stages, and fractionation procedures for each separation can be found either in the **Methods** section or in the PRIDE database (accessions in **Table S2)**.

In all, we measured 379,758,411 peptides that were uniquely assigned to 259,732 orthologous groups (or unique proteins not mapping to orthogroups) across the 31 species, 149 separations, and 10,491 fractions (**Figure 3A**). We further augmented our CFMS data with ~15,000 mass spectrometry proteomics experiments [28] that included affinity purification mass spectrometry (APMS) [50–52], proximity labeling [53,54], and RNA-pulldown data [55]. In total, we incorporated data from 26,297 mass spectrometry experiments. We then filtered this data such that we only retained LECA OGs that were strongly observed, in that the sum total of peptide spectral matches (PSMs) across all 149 fractionations was greater than or equal to 150 PSMs. This resulted in elution profiles for 5,989 well-measured LECA OGs, encompassing approximately 60% of the estimated LECA gene set.

While many complex subunits have co-elution profiles with visually detectable correlation (as for the members of the COPI vesicle coat complex, 20S proteasome, and eukaryotic initiation factor 3 in **Figure 3A, right**), a computational framework is required for systematic identification and to properly control for false positives (**Figure 2C**). To this end, we employed a supervised machine learning pipeline trained on the data we observed for known protein complexes. We assembled a set of 1,499 known complexes from two databases that record PPIs for a variety of eukaryotic species [56–58] (see **Methods**). We then assessed a number of interaction prediction models generated by three different classification algorithms: extremely randomized trees (“ExtraTreesClassifier”), linear support vector (“LinearSVC”), and stochastic gradient descent (“SGDClassifier”).

We constructed models for each with a variable number of features derived from the mass spectrometry datasets, ranking features on a classifier-by-classifier basis, and evaluated performance by measuring the precision and recall of known complexes withheld from the training data (**Figure 3B**).

The precision-recall of our initial models were 5-38% recall at 90% precision, similar to previous large-scale protein interaction maps papers [28,29,31]. However, after implementation of a custom data stratification approach (see **Methods**), we observed improvements to performance, with models ranging from 40-90% recall at 90% precision. We suspect that this large jump in performance stemmed from a combination of the large volume of high quality mass spectrometry data being

integrated and the data stratification approach, which we found significantly reduced overfitting (see **Methods**).

Using each classifier and its best performing feature set, we identified PPIs with a 10% false discovery rate (FDR) threshold (see Zenodo repository) and clustered the proteins into complexes using an unsupervised community detection "walktrap" algorithm [59], weighting the interactions by their confidence scores. The walktrap procedure determined the "optimal" number of subcommunities to range between 100-400, differing significantly by classifier. These large clusters capture the general organization of eukaryotic cells (e.g., partitions include a large spliceosomal cluster, a large chromosomal maintenance cluster, clusters broadly associated with cilia, etc). In order to obtain more granular protein complexes, we further divided these clusters into increasingly smaller communities to define a hierarchy of protein interactions (**Table S3**).

As a positive control, we noted that this approach successfully delineated known complexes. For example, a large spliceosome cluster was demarcated into LSM, PrP19 complex, and U4/U6 x U5 tri-snRNP complexes with increased granularity (**Table S3**). For each classifier, we quantified the performance of the walktrap procedure by computing precision and recall for each cluster at each level of the hierarchy (**Figure 3C**) and observed a clear tradeoff for increasingly fine-grained subcommunities to show increased precision but decreased recall. Overall, the support vector classifier (**Figure 3C**, red) netted the highest quality protein complexes and was chosen as our final model. The resulting final LECA complexome consists of the highest confidence 109,466 pairwise interactions between 3,193 unique OGs, hierarchically assembled into 199 (less granular) to 2,014 (more granular) protein complexes, which are portrayed schematically in **Figure 4**.

We sought to assess how well our final protein interaction model agreed with independent studies that defined interactions using orthogonal approaches. We observe that protein pairs within our highest scoring threshold (≤10% FDR) are significantly more likely than random chance to agree with yeast 2-hybrid (Y2H) [60], mRNA co-expression [61], and cross-linking mass spectrometry (XLMS) [62] interactions (**Figure 3D**), and performed comparably with a previous interaction map in plants [31].

Finally, current phylogenomic studies hypothesize myriad protein assemblies at the root of eukaryotes [5,8,9,18,63–65]. With an experimentally determined set of conserved LECA PPIs in hand, in the next sections, we examined the extent to which our interactome both recapitulates previously hypothesized and discovers new LECA complexes. We focused on ancient protein assemblies related to intracellular trafficking

and cell projection because these are highly relevant to modern human disease [66–68] and highlight below multiple examples where we uncovered previously undescribed interactions.

Deep conservation and loss of vesicle tethering complexes

One hallmark of eukaryotic cells is their system of intracellular trafficking by cargo-laden vesicles that bud from one compartment and fuse to another, supported by coat proteins (e.g. clathrin, COPI, and COPII), membrane-anchored SNARE proteins to facilitate membrane fusion [69], and tethering factors ensure target specificity [70]. Many compartment-specific tethering modules are thought to have been present in LECA (as reviewed in [63]), including the ER-associated TRAPP-I, TRAPP-II, and TRAPP-III complexes, the Golgi-associated retrograde protein (GARP), the conserved oligomeric Golgi (COG) complexes, the endosome-associated recycling protein (EARP), and the endolysosomal homotypic fusion and protein sorting (HOPS) complexes. However, the extent to which particular protein interactions are conserved remains an open area of research. In our LECA interactome, we recover all of these core tethering assemblies, along with some unexpected members (**Figure 5A**).

The GARP and EARP protein complexes are closely related and share three subunits (VPS51, VPS52, VPS53) [71]. Localization is conferred by additional subunits: VPS50 for endosomes or VPS54 for the Golgi apparatus [71]. We observe strongly conserved interactions between all of these subunits, in addition to EIPR1 (EARP and GARP complex-interacting protein 1) (**Figure 5A, top right**). The interaction of EIPR1 with the GARP/EARP complexes was only recently discovered, first in high-throughput screens of human proteins [50,52] and then confirmed in targeted study in human neuroglioma cells [72]. While EIPR1 is speculated to be widely conserved, we find that its interaction with GARP/EARP is indeed ancient and likely traces back to LECA.

In modern eukaryotes the HOPS complex shares four of its six subunits (VPS11, VPS16, VPS18, VPS33) with the related CORVET complex, while the remaining two subunits (VPS39 and VPS41) are unique to HOPS. In yeast, the CORVET subunits direct the fusion of early and recycling endosomes while HOPS directs the fusion of late endosomes, lysosomes, and autophagosomes [73]. Interestingly, we observe conserved interactions between VPS8 (previously thought to be CORVET-specific) and the VPS16, VPS18, VPS39 and VPS41 subunits of the HOPS complex (**Figure 5A, top left**), raising the possibility that a single HOPS-like complex in LECA may have governed the endolysosomal vesicle fusion pathway, with subsequent lineage-specific duplication and specialization of subunits for different compartments.

Analogously, the ER/Golgi-associated TRAPP complex is thought to be composed of five core subunits (TRAPPC1-5) with additional subunits in distinct TRAPP-I (TRAPPC6), TRAPP-II (TRAPPC9, TRAPPC10, TRAPPC13), and TRAPP-III (TRAPPC8) complexes. However, the number and identity of proteins in TRAPP-I/II/III vary significantly by species [74]. In our ancient interactome, we observe strong interactions between each member of the core C1-C5 complex, conserved across all sampled eukaryotic supergroups. Unexpectedly, we find pan-eukaryotic evidence for TRAPPC12 in this core complex, previously thought to be metazoan-specific [75,76]. The remaining interactions are differentially lost in specific eukaryotic lineages, with TRAPPC10 absent in all five sampled TSAR species. Our data, combined with conflicting literature on the exact composition of TRAPP-I/II/III, thus suggests an ancient and flexible core complex where subunits differentially specialize along different eukaryotic branches.

The eight-subunit COG assembly governs retrograde intra-Golgi trafficking and comprises two heterotrimeric subcomplexes (COG2-4 and COG5-7) linked by a COG1-COG8 heterodimer [77]. Our LECA interactome recapitulates this assembly and includes interactions with TMF1 and the ubiquitin ligase complex RNF20-RNF40 (**Figure 5A, bottom right**). TMF1-COG interactions have only been previously observed for metazoan COG2 and COG6 [78], but our data show confident TMF1-COG interactions spanning Amorphea and Archaeplastida, with TMF1 lost in Excavata and TSAR. RNF20-RNF40 is generally described in nuclear roles like histone ubiquitination, transcription regulation, and DNA damage repair [79,80], but has been also linked to the Golgi-associated adapter protein WAC [81], involved in Golgi membrane fusion [82,83]. We see strong conservation of RNF20-RNF40 interactions with COG across Amorphea, TSAR, and Archaeplastida, suggesting that the nuclear-repurposing of this ubiquitin ligase complex could be a recent mammalian innovation.

Thus, the LECA interactome reveals the conservation and specialization of eukaryotic vesicle tethering complexes. We identified unexpected ancient interactions, such as those involving EIPR1 with GARP/EARP, TMF1 and RNF20-RNF40 with COG, and TRAPPC12 with the TRAPP complex, and additionally saw evidence for flexible and lineage-specific adaptations. Given this utility for examining evolutionary conservation and diversification of LECA-associated complexes, we next applied it to shed light on a central question of LECA evolution that is in dispute.

Primordial origins of cell projection and phagocytosis

While phagocytosis is a trait widely observed across diverse groups of eukaryotes, it is debated whether LECA had the ability to recognize and engulf large particles. Contention stems from arguments concerning how the first eukaryotic

common ancestor (FECA) acquired the alpha-proteobacterial precursor of the mitochondrion, *i.e.*, whether FECA was akin to a phagocytosing archaeon or a more "simple" prokaryote that existed in protracted syntrophy with an alpha-proteobacterium [84–86]. Existing phylogenomic investigations into the origins of phagocytosis are conflicting; at least three independent studies conclude phagocytosis probably evolved independently in multiple eukaryotic lineages [87–89], while others argue that the trait was present in LECA and the absence of phagocytosis in certain eukaryotic groups is due to secondary loss as they adapted to new niches [90,91]. Recent investigations into Asgard archaea, the sister group to eukaryotes, reveal a dynamic actin-based cytoskeleton composed of F-actin assemblies, actin-related proteins (Arps), and actin-binding proteins such as profilins and gelsolins capable of modulating eukaryotic actin [92–94], suggesting that FECA may have had phagocytic capacity.

Within our LECA gene set, we find an extensive complement of LECA OGs generally associated both with cell projections (pseudopodia, lamellipodia, filopodia) and with phagocytosis (phagocytic cups and phagosomes) (**Figures 1, 5B**). Specifically, LECA appears to have had Rho GTPases such as RAC/CDC42, formins, coronins, cofilins, gelsolins, and proteins associated the ARP2/3, ENA/VASP, WASP, SCAR/WAVE, and PI3K complexes. For example, in the LECA interactome, we recovered all seven subunits of the ARP2/3 complex, responsible for the cytoskeletal rearrangement required for cell projection, clustering with other related protein complexes involved in cell protrusion in addition to a number of proteins that are critical for phagocytosis in extant eukaryotic cells (**Figure 5B**). The core ARP2/3 complex consists of the proteins ARP2, ARP3, ARPC1, ARPC2, ARPC3, ARPC4, and ARPC5. Interestingly, in our data, ARPC5 is the most peripherally associated component in the cluster and has been completely lost in all species sampled within Excavata and TSAR.

Combined with previous evidence [95–97], the presence of these genes in LECA suggests that the ancestral eukaryotic cell was almost certainly capable of pseudopod formation and projection-based motility despite the lack of UniProt annotations in species outside Amorphea (**Figure S1**). However, because cell projections and phagocytosis share underlying molecular machinery, it is less clear if the presence of these systems necessarily imply a phagocytosing LECA, and more evidence is required to conclude that phagocytosis is an ancestral trait present at the root of eukaryotes. To address this question, we therefore explored the conservation of additional interactions that might shed light on this issue.

Among peripheral interactors of the ARP2/3 complex, we observe CAPZA and CAPZB forming the heterodimeric F-actin capping complex, an essential regulator of actin nucleation that restricts elongation [98], as well as formins and coronins known to

promote elongation [95,99]. We also find interactions with WDR1, a promoter of cofilin-mediated actin severing [100] that assists both actin polymerization and depolymerization [101]. Research strongly implicates these systems in Amorphean phagocytosis: Coronins are strongly enriched in phagocytic cups and defects result in impaired phagocytosis in both *Dictyostelium* and mammalian cells [102–104]. Furthermore, we observe protein interaction evidence in at least two major eukaryotic supergroups consistent with the reported roles of AAK1 in receptor mediated endocytosis [105] and unconventional myosin in phagocytosis [106–109]. Taken together, our results strongly support a phagocytosing LECA.

The LECA interactome reveals a ciliary mechanism for *EFHC2*-associated renal failure

The conservation of protein interactions over billions of years of evolution implies that they are strongly constrained and that their malfunction is likely to be pathogenic. Thus, studying genetic variation through the lens of conserved protein interactions should clarify mechanisms of genetic disease development, tolerance, and resilience. Consequently, we expect the LECA protein interactions to offer direct insights into human disease genetics, and by similar logic, to genotype-phenotype relationships of other modern eukaryotes.

Slightly less than half of human genes date back to LECA. Where these conserved genes have been characterized, they have been shown to be responsible for a large and diverse subset of major human diseases, spanning developmental disorders, cancers, chronic respiratory diseases, neurodegenerative conditions, and motor disorders (**Figure 6A**). For example, of the ~100 human genes known to be associated with deafness, nearly ¾ were present in LECA (**Figure 6A**). While some human diseases are “new”, evolutionarily-speaking, such as deficiency of the animal-specific pituitary hormone, many other diseases, such as ciliary dyskinesia, arise nearly entirely from genes in LECA OGs. In order to test the utility of our data for illuminating the biology of extant species, we next asked if the LECA interactome could be leveraged to predict human disease mechanisms and novel gene-disease relationships.

Approximately 500 LECA OGs are related to cilia, and among the most common ciliopathies are diseases of the kidney [110]. We identified a male infant with microcephaly, seizures, polycystic kidney disease, and end-stage renal failure, and whole exome sequencing and pedigree analyses revealed a significant hemizygous, X-linked G>A variant in *EFHC2* (rs34729789, 11:44148852:G:A) (**Figures 6B-D, S2**). Essentially nothing is known of the function of *EFHC2*, though it and its paralog *EFHC1* encode proteins thought to be microtubule inner proteins (MIPs) that function

specifically in motile cilia; loss of their orthologues in *Chlamydomonas* or *Tetrahymena* leads to defective ciliary axonemes and/or ciliary beating, but do not disrupt ciliogenesis [111,112]. Because motile cilia are not present in mammalian kidneys, the link between *EFHC2* and this patient's disease was surprising. We therefore examined our LECA interactome for insights.

We first noted that the patient's missense variant altered an arginine residue at position 133 to histidine, and this residue is conserved across Archaeplastida, Excavata, TSAR, and Amorphea (**Figure 6D**). Moreover, EFHC2 was closely and exclusively linked in our LECA interactome to other proteins involved in cilia motility (**Figure 6E**). We therefore examined the protein's localization in *Xenopus* multiciliated cells, and found it to be very strongly localized to ciliary axonemes. By contrast, the disease-associated R133H variant failed to localize to cilia (**Figure 6F, G**), suggesting that defective ciliary localization of this protein contributed to ciliopathic kidney disease in the affected child.

This result then prompted us to ask if other proteins in the cluster, which are also thought to function specifically in motile cilia, might also be implicated in kidney disease. Indeed, previous genomic analyses link both *PACRG* [113] and *TPPP* [114] to chronic kidney disease, with *PACRG* specifically linked to end-stage renal disease [113]. This combination of clinical data, the LECA interactome, and specific hypothesis testing in a vertebrate model organism thus links EFHC2 ciliary function to an end-stage renal disease for which the molecular etiology was previously unknown and underscores the power of our comparative evolution strategy.

Network propagation for systematic ranking of gene-disease relationships

We next sought to score potential disease-causative proteins systematically within our conserved interactome on a disease-by-disease basis. To this end, we used cross-validated network guilt-by-association [115] to predict novel gene-disease pairs for 109 unique diseases based on clinically-validated genotype-to-phenotype relationships sourced from the OMIM database [116] (**Figure 7A**; see **Methods**). We measured the power of our approach as the areas under receiver operating characteristic curves (AUROC), and compared the predictive performance of known disease-associated gene sets versus that of random gene sets. As expected, random predictions have AUROC scores distributed around 0.5 (**Figure 7A, yellow**). In striking contrast, LECA-interactome predictions were skewed to higher scores (**Figure 7A, blue**), and using a conservative AUROC threshold of 0.7, we made strong new disease candidate predictions for almost one-third of the diseases considered (~35 Mendelian disorders). Below, we discuss the prediction and validation in animals of two such novel protein associations.

Identification and validation of *ATP6V1A* as a novel candidate for osteopetrosis

Our LECA network propagation approach implicated several vacuolar-type H^+-ATPase (V-ATPase) proteins in the molecular etiology of osteopetrosis (AUROC ~0.8), a disorder in which bones grow abnormally and become overly dense [117] (**Figure 7B**). Given the key role of V-ATPases in regulating bone homeostasis by acidifying the space between osteoclasts and bone to help dissolution of bone hydroxyapatite [118], one might assume the disruption of many V-ATPase subunits would result in increased bone density. However, only three subunits have so far been implicated in osteopetrosis in humans [116,119] or mice [120]. The remaining V-ATPase subunits instead display a remarkably broad spectrum of disease associations, including cutis laxa (loose skin) and renal tubular acidosis [121], neurodegenerative disease, deafness [122], Zimmermann-Laband syndrome [123], and even osteoporosis (bone loss) [124], highlighting the need to elucidate the discrete molecular functions of specific V-ATPase subunits.

Our LECA network propagation approach gratifyingly made precise predictions, linking three specific subunits (ATP6V1A, ATP6V1B, ATP6V0D) to osteopetrosis (**Figure 7B**). To confirm these predictions we examined heterozygous CRISPR-Cas9 knockouts of *ATP6V1A* (performed by the KOMP2 high-throughput mouse phenotyping site at the Baylor College of Medicine, see **Methods**) and found these mice showed significantly increased bone mineral content (**Figure 7C**). The effect size was much stronger in female mice (p = 0.00003) than in male mice (p = 0.00813), echoing previous observations of sexual dimorphism in the body composition of mammals [125,126]. Despite the obvious lack of bones in the single celled last eukaryotic ancestor, then, our examination of the underlying protein interaction network of that organism nonetheless identified a specific mammalian phenotype with one specific subunit from among a large repertoire of closely related genes.

Ancient interactions suggest new candidate genes for a lethal human ciliopathy

Our highest scoring disease association (AUROC ~0.98) involved short-rib thoracic dysplasia (SRTD), a severe human ciliopathy characterized by skeletal abnormalities including dysplasia of the axial skeleton that in many cases lethally impairs respiratory function [127]. The disease is strongly associated with proteins involved in Intraflagellar Transport (IFT), the system which moves cargoes into and out of cilia, and this was reflected in our LECA interactome (**Figure 7D**). Our highest scoring non-IFT protein prediction, however, was the Golgi protein GLG1 [128]. This was an interesting candidate because mouse mutants of GLG1 display defects in rib development similar to SRTD [129], yet the protein has never been implicated in any aspect of ciliary biology.

We therefore explored the function of GLG1 in *Xenopus* multiciliated cells, and found it predominantly localized to the Golgi in MCCs, as expected. We observed no apparent localization at basal bodies or in cilia (not shown). Nonetheless, knockdown of GLG1 resulted in a significant loss of cilia from MCCs, an effect that was specific since it could be rescued by expression of GLG1-FLAG (**Figure 7E**). To ask if this defect in ciliogenesis was related to IFT, we performed live imaging of GFP fusions to two components of the IFT complex. In normal cells, both markers labeled small punctae in axonemes of *Xenopus* MCCs, consistent with previous imaging of IFT in these cells [130,131]. By contrast, GLG1 knockdown cells displayed large accumulations of IFT proteins within axonemes (**Figure 7F, 7G**) that resemble those seen previously after disruption of IFT [130,131].

Thus, analysis of the LECA interactome made a single, specific prediction of ciliary function for just one among the large array of Golgi-resident proteins, and that prediction was validated by experiments in *Xenopus*. These data provided new insights into the still obscure link between IFT and the Golgi [132,133] and, moreover, identified a plausible candidate gene for SRTD.

Conclusions

Studies of ancient protein-protein interactions and their conservation across species offer valuable insights for exploring the genetic underpinnings of contemporary genetic traits and diseases in modern species. In this work, we took an integrated approach to reconstruct the macromolecular assemblies of ancient proteins that, until now, have only been sparsely described. We defined a core set of likely LECA orthogroups, finding that slightly fewer than half of human genes can be traced back to this set. We integrated those data with more than 26,000 mass spectrometry proteomics experiments, capturing hundreds of millions of unique peptide measurements for hundreds of thousands of unique proteins in species sampled from across the tree of eukaryotes. Using these data, we reconstructed a high-quality conserved LECA protein interactome. This interaction network has formed the core of eukaryotic biology for nearly two billion years, and the dataset reveals new insights into both known protein complexes and novel assemblies.

Consistent with our central premise that the most highly-conserved protein assemblies will tend to be most critical for proper cell and organism function, the LECA interactome successfully predicts mechanisms of human disease and novel gene-disease relationships. We specifically presented evidence for a ciliary mechanism in human *EFHC2*-associated renal failure, identified the V-ATPase subunit ATP6V1A in the etiology of mammalian osteopetrosis, and demonstrated a role for the Golgi protein GLG1 in trafficking IFT-A proteins into cilia as a molecular mechanism for short-rib

thoracic dysplasia. Given the intrinsic richness of these datasets, we expect this approach should similarly extend to traits and diseases in most other eukaryotic species, while providing insights into the specific molecular mechanisms involved due to being anchored in deeply conserved ancient protein activities.

SUPPLEMENTAL INFORMATION

Detailed Supplemental Methods are provided, including 6 supplemental figures and 5 supplemental tables. All raw and interpreted mass spectrometry data were deposited to the ProteomeXchange *via* the PRIDE partner repository with the identifiers provided in **Table S2**. All project data files are deposited at Zenodo (https://doi.org/10.5281/zenodo.11267529). Custom R, Python, Bash and Perl scripts used for all analyses and figure generation are available at https://github.com/marcottelab/leca-proteomics.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the generous support of the Tetrahymena stock center (Cornell University & Washington University in St. Louis), the Texas Advanced Computing Center at The University of Texas at Austin for providing high-performance computing resources, the Knockout Mouse Program (KOMP2) at the Baylor College of Medicine for *Atp6v1a* mice data, Johann Eberhart for rotifer cultures, the University of Texas UTEX Culture Collection of Algae for algal samples, and Angel Syrett and Elinor Marcotte for species illustrations. Research was funded by grants from the National Institute of General Medical Sciences (R35GM122480 to E.M.M. and F31GM143881 to R.M.C), National Institute of Diabetes and Digestive and Kidney Diseases (R01DK068306 to F.H.), National Institute of Child Health and Human Development (R00HD092613 to K.D. and R01HD085901 to J.B.W. and E.M.M.), Army Research Office (W911NF-12-1-0390 to E.M.M.), and Welch Foundation (F-1515 to E.M.M.).

AUTHOR CONTRIBUTIONS

Design and co-supervision: R.M.C., J.B.W., and E.M.M.
Proteomics experiments: O.P., aided by K.D. and J.C.L.
Data analysis: R.M.C., aided by T.G., A.M.B., M.L., K.D., C.D.M., D.Y., and D.D., and guided by E.M.M.
Clinical genetics: S.S. and F.H.
Xenopus experiments: C.L., T.G., and J.B.W.
Manuscript initial draft: R.M.C., O.P., J.B.W., and E.M.M.
All authors discussed results and contributed edits.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL METHODS

Resources for inferring the LECA gene set

Reference proteomes for 156 species (122 eukaryotes, 7 archaea, 27 bacteria; see Zenodo repository) were downloaded from the UniProt database along with the corresponding reference species tree [134] for the parsimony analysis. This species tree and set of organisms were selected because they span the tree of life and serve as the gold standards curated by the Quest for Orthologs group for benchmarking orthology inference [135]. The species tree was downloaded from SwissTree (https://swisstree.sib.swiss/cgi-bin/swisst). Analysis of UniProt database reviewed proteins annotated with subcellular localizations was performed using the standardized SL accessions (see Zenodo repository), extracted with REST API queries.

Orthology mapping

Protein sequences from each reference FASTA file were searched against the eggNOG 5.0 database [37] and mapped to orthologous groups (OGs) at the rootNOG level (taxonomic level = 1) using eggNOG-mapper v2.0.5 [136] with DIAMOND and a hit cut-off e-value of 10^{-3}. As a result, 89,955 unique OGs spanning 156 species across the tree of life were used as input to the Dollo parsimony analysis. The group of rootNOGs assigned to the LECA node as a result of the parsimony procedure were converted to euNOGs (taxonomic level = 2759) with a set of hierarchical mapping files provided by Dr. Jaime Huerta-Cepas, the author of the eggNOG algorithm, *via* personal correspondence.

Dollo parsimony

Using the Count evolutionary analysis software [137], we implemented a Dollo parsimony approach [36] across 156 organisms to obtain a conservative estimate of the LECA proteome. The Dollo parsimony model relies on the simplifying assumption that gene loss is irreversible, e.g., once a gene is lost it cannot be regained in a lineage. Thus, we determined the ancestral LECA proteome as the set of orthogroups either (a) shared by the respective outgroups (prokaryotes) and at least one of the eukaryotic species or (b) shared by two eukaryotic groups whose last common ancestor was LECA as defined by the gold standard species tree [138]. This approach has previously been shown to be effective at reconstructing likely LECA orthogroups [19].

Investigation of orthologous groups of unknown function

We attempted to assign functions, or, at the minimum, subcellular localizations, to the 25% of uncharacterized (by eggNOG) LECA OGs using the UniProt database. To this end, we downloaded 363,430 proteins from the UniProt database that were (a) “reviewed” status and (b) assigned a standardized subcellular localization (SL) ID for each compartment that we trace back to the last eukaryotic common ancestor (**Figure**

S1A). The total number of proteins assigned to a subcellular compartment varied by four orders of magnitude, where 166,296 proteins were assigned to the cytoplasm at the highest end and 116 proteins were assigned to phagocytic cups at the lowest end (**Figure S1A**). Furthermore, we investigated the diversity and magnitude of eukaryotic and prokaryotic species contributing to these annotations (**Figure S1B**) and observed an underrepresentation of clades outside Amorphea (see tree in **Figure S3** for supergroup organization). For example, the nucleus is considered a distinguishing feature of eukaryotes; UniProt proteins annotated to localize to the nucleus come predominantly from 1,032 distinct Amorphean species, with almost two orders of magnitude fewer such annotations contributed by non-amorphean species, consisting of 152 archaeplastidans, 2 cryptophytes, 21 excavates, and 62 TSAR species.

To quantify what proportion of the LECA gene set is represented in UniProt, we mapped all 363,420 reviewed UniProt proteins mentioned above to eukaryotic orthologous groups ("euNOGs"; NCBI taxonomic identifier = 2759). This resulted in 13,556 unique euNOGs, the percentage of which that trace back to LECA varies significantly by subcellular localization (**Figure S1C**). As an aside, we note that the same euNOG can often be assigned multiple UniProt SL IDs, netting a total 24,628 euNOG-SL mappings.

Of the 2,790 ciliary proteins that map to 478 unique eukaryotic orthologous groups (euNOGs), nearly ~25% of the 299 UniProt ciliary euNOGs that intersect with LECA OGs were originally assigned "unknown function" by the eggNOG functional annotation algorithm–a larger proportion than most of the other eukaryotic compartments described within the UniProt database. Similarly, of 1,317 cytoskeletal euNOGs, ~60% trace back to LECA and 111 of those were assigned the "function unknown" eggNOG category. Thus, the combination of eggNOG and Uniprot annotations provided a reasonable initial annotation set for subsequent analyses.

Challenges and limitations to defining the LECA gene set

Binning proteins into evolutionarily related orthologous groups with respect to the root of the eukaryotic tree nets a "coarse-grained" mapping of the relationships between eukaryotic genes, *i.e.*, we can not rigorously distinguish orthologs from paralogs. With that said, we are still able to draw conclusions about the properties of families of genes rather than the pairwise relationships of individual members; this approach is intuitive and convenient for large-scale systematic studies and broadly supported [139–141]. However, there is considerable disparity between OG assignment algorithms, though eggNOG has been demonstrated to have among the highest accuracies when tested on a benchmark set of manually curated orthologs [19]. In the same study, the eggNOG algorithm was also shown to perform best at detecting distant homology and properly

splitting out-paralogs, making it the best suited algorithm currently available for our goals. Some protein families and eukaryotic lineages with fast rates of evolution (e.g., transcription factors, proteins associated with the innate immune response, and in general plants that are prone to whole genome duplication [142–146]) remain a weakness to the approach. Proteins such as these are likely "under-split" with respect to their associated orthologous groups. Leucine-rich repeat proteins are a salient example: more than 100 human LRR proteins were assigned to KOG0619, an OG we traced back to LECA related to intracellular trafficking and secretion, and this trend persists across nearly all eukaryotes sampled that had proteins assigned to KOG0619. In this way, we are most likely underestimating the size of the distinct LECA gene set.

Additionally, it should be noted that the Dollo parsimony procedure we used to approximate the LECA gene set, which assumes that the probability that a trait emerges more than once is negligible [36], is the simplest form of ancestral state inference. The use of Dollo parsimony is justified and perhaps even preferable [147–149], given that (a) our goal was to determine a binary character state (the presence or absence of genes), (b) we had a consensus reference species tree in hand [134], (c) the target gene set is eukaryotic wherein independent gene losses are common and gains of multiple genes are (relatively) rare, (d) the expected influence of horizontal gene transfer is minimal (estimated to be ~1% of genes or less [150]), and (e) probabilistic ancestral state reconstruction methods, such as phylogenetic birth-death-gain models, are prohibitively slow for a data set of this size. Nonetheless, one flaw in our approach is worth noting: multiple species within Excavata host plastids or plastid-derived genes orthologous to plastid proteins in plants [151–153], even though it is widely accepted that primary plastids share a single origin [154,155] and Archaeplastida is monophyletic [156,157]. If the Archaeplastidan monophyly is to be believed [158,159], the last common ancestor of Excavata independently acquired plastids (violating Dollo's law), resulting in the inflation of our LECA gene set by ~40 plastid-associated OGs. To correct for this error, we manually removed these OGs from consideration during construction of the LECA interactome.

Resources for interactome mapping

Biological samples, mass spectrometry data sets, and software used in this analysis are summarized in **Table S2**. Proteomes for 31 eukaryotic species were sourced as summarized in **Table S4**.

Mass spectrometry

Native protein extraction and fractionation

For lysates described below protease inhibitor cocktail was cOmplete mini EDTA-free (Roche), phosphatase inhibitors were PhosSTOP EASY pack (Roche), and

all steps after addition of lysis buffers were conducted at 4°C or on ice unless otherwise indicated. Native soluble extracts were quantified by DC Protein Assay (BioRad). All protein samples were 0.45 µm filtered (Ultrafree-MC-HV Durapore PVDF, Millipore) prior to chromatography. Chromatography was performed on an HPLC system as in [31] unless otherwise stated.

Brachionus rotundiformis was collected in batches on Filter Mesh 100 Nylon (~65 µm pore) to remove feeder algae prior to flash freezing in liquid nitrogen. Frozen material (3.1 g) was ground to power in a liquid nitrogen-chilled mortar and pestle and resuspended in an equal volume of *Tetrahymena* Lysis Buffer (25 mM Tris pH7.4, 25 mM NaCl, 1 mM EDTA, 10 % glycerol, 0.2% NP40, with 1 mM DTT, 1 mM PMSF, phosphatase inhibitors, and protease inhibitor cocktail added freshly). Cells were disrupted with 10 strokes in a glass dounce fit with a tight pestle. Following centrifugation 3000 x g, 10 minutes to remove debris, the supernatant was clarified twice by centrifugation 20,000 x g 10 minutes. Size Exclusion Chromatography was performed with 2.6 mg extract in a 200 µl sample loop and mobile phase Buffer S (50 mM Tris-HCl pH 7.5, 50 mM NaCl).

Phaeodactylum tricornutum (UTEX 646) grown without silica was briefly washed by pelleting (2000 x g, 10 minutes, 21°C, no brake) and resuspended in 0.5x artificial seawater (UTEX) before collecting (3000 x g, 4°C, slow deceleration) and flash freezing. Frozen material was ground to powder and allowed to thaw before refreezing and regrinding. 1g of powdered material was resuspended in 800 µl Lysis Buffer (50 mM Tris pH 7.5, 150 mM NaCl 5 mM EGTA, 10% glycerol, 1% NP40 with 0.1mM DTT) with phosphatase inhibitors and Plant Specific Protease Inhibitors (Sigma # P9599). Material was frozen and thawed again before sonicating 6 x 10 seconds on, 20 seconds off, 70% duty cycle. Lysis was monitored by microscopy. The extract was incubated on ice with periodic gentle vortexing for 30 minutes prior to clarification twice at 14,000 x g, 10 minutes. Extract was diluted 3-fold with 50 mM NaCl prior to loading 2 mg for SEC separation as above. For separation by mixed bed ion exchange chromatography (Poly CATWAX A, PolyLC Inc.) salt was reduced by 5x dilution with 10 mM Tris pH 7.5, 5% glycerol, 0.01% NaN_3 and proteins were re-concentrated by ultrafiltration (Amicon Ultra 0.5 ml 10,000 MWCO). IEX chromatography was with 1.9 mg in a 250 µl sample loop.

Euglena gracilis (UTEX 753) was washed briefly by centrifugation (1,500 x g, 5 minutes, 21°C) and resuspension in dH_20, before collection by centrifugation and flash freezing. Material was ground as above and 4.7 g was resuspended in Lysis buffer plus both the cOmplete mini EDTA-free protease inhibitors and the Plant-Specific Protease Inhibitors (Roche). Lysate was sonicated 9 x 10 seconds on, 20 seconds off, 60% duty cycle, followed by gentle nutation 30 minutes. Debris was removed by centrifugation

1,500 x g, 10 minutes, and the supernatant was further clarified twice with 14,000 x g, 10 minute spins. Final extract was filtered through a 0.45 µm syringe filter (Durapore PVDF, Millipore) prewashed with dH_20. Extract was diluted 4-fold in Buffer S and 2 mg loaded on a 200 µl sample loop for SEC fractionation.

Two fresh pig tracheas (*Sus scrofa*) were shipped on ice from Sierra for Medical Science arriving within 24 hours of harvest. After removal of fat tissues the trachea were slit lengthwise, chopped crosswise into several pieces, and washed with multiple changes of ice cold PBS pH 7.4 to remove serum and blood cells prior to extraction with 100 ml Ca++ shock buffer as in [160] including protease and phosphatase inhibitors at 0.5x concentration and 0.1 mM PMSF. Cilia were released by vortexing and manual agitation for 10 minutes. Debris was pelleted 500 x g, 2 minutes and floating lipids were removed by aspiration. Cilia were collected by centrifugation 12,000 x g 10 minutes and washed once by resuspension and centrifugation. Ciliary pellets were resuspended in Ca++ shock buffer with 1% NP40 to extract soluble proteins and residual axonemes were removed by centrifugation twice at 12,000 x g 10 minutes. Any floating lipids were removed after each spin. Extract was flash frozen until used. Thawed extract was diluted 2-fold with 10 mM Tris pH 7.5, 5% glycerol, 0.01% NaN_3 and re-clarified 12,000 x g 10 minutes prior to ultrafiltration with 30,000 MWCO Ultracel Amicon Ultra 0.5 ml units to load 1.7 mg in a 250 µl sample loop for IEX chromatography.

Tetrahymena thermophila SB715 were grown and cilia extracts made as in [161] except that deciliation was by pH shock according to [162]. 1.5 mg cilia extract was fractionated by mixed bed IEX and 1.2 mg by SEC with SEC mobile phase Buffer S-C (50 mM Tris-HCl pH 7.4, 50 mM NaCl, 3 mM $MgSO_4$, 0.1 mM EGTA). Deciliated *Tetrahymena* “bodies” were collected by centrifugation 1,700 x g, 5 minutes, washed once by resuspension in Deciliation Medium (10 mM Tris-HCl pH 7.4, 10 mM $CaCl_2$, 50 mM sucrose), collected by centrifugation as before and flash frozen until use. *Tetrahymena* body lysate was prepared by liquid nitrogen grinding frozen material before resuspending in an equal volume *Tetrahymena* Lysis Buffer with 0.1 mM PMSF. Lysis was achieved on ice for 10 minutes by pipetting up and down. Debris was removed by centrifugation 3,000 x g, 10 minutes. Supernatant was clarified and floating lipids removed by sequential centrifugations at 40,000 x g, 10 minutes, 45,000 x g 30 minutes, 130,000 x g 1 hour, and 130,000 x g 45 minutes. Extract was diluted (final NaCl 22 mM) prior to loading 2.2 mg on a 250 µl sample loop for IEX chromatography. The remaining extract was flash frozen and thawed later for SEC chromatography. Extract was clarified 25,000 x g 10 minutes immediately after thawing, and again after dilution in Buffer S-C for loading of 1.4 mg in 200 µl sample loop. For DSSO-crosslinked samples, cilia extract was prepared using the pH shock method as above, but 20 mM HEPES pH7.4 was substituted for the 50 mM Tris of the Cilia Wash Buffer. Extract was

concentrated by ultrafiltration in an Amicon Ultra Ultracel 10k NMWL unit (UFC501096) to load 1.5 mg in a 250 µl sample loop. The final concentration of NP40 was 2.75%. Fractionation on a mixed bed IEX was performed with substitution of 10 mM HEPES pH 7.4 for Tris in the chromatography buffers A and B. For crosslinking DSSO was dissolved freshly in dry DMF to 50 mM and then diluted with 10 mM HEPES pH 7.4 to 10.5 mM before dispensing 25 µl into each 500 µl fraction. To ensure activity of the crosslinker the DSSO solution was prepared in 2 consecutive batches to treat a total of 76 column fractions. Crosslinking proceeded 1 hour at room temperature (~21°C) and was quenched by addition of Tris pH 8.0 to 28 mM.

Xenopus laevis sperm were isolated from dissected testes of five or eight J-strain *Xenopus laevis* males. Testes were perforated with a 25-gauge needle, sperm blown out using MMR (Marc's Modified Ringers). Larger debris was allowed to settle, and liquid transferred to a fresh tube. Sperm were collected by centrifugation 1,500 x g, 10 minutes. Supernatant was discarded and the sperm pellet was lysed by resuspension in an equal volume of Sperm Lysis Buffer (10 mM Tris-HCL pH7.5, 20 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 1% n-Dodecy-ß-D-Maltoside (Anatrace) with 0.5 mM DTT added freshly). Lysate was clarified by centrifugation 14,000 x g, 10 minutes. 1.2 mg was loaded for mixed bed IEX column fractionation (PolyLC Mixed-Bed WAX-WCX, PolyLC Inc. #204CTWX0510), and 3 mg for SEC fractionation (BioSep-SEC-s4000, Phenomenex).

Mus musculus (embryonic stem cells) were grown as described in [163]. Cells were harvested without trypsin by washing in ice cold phosphate buffered saline (PBS), pelleted, and placed on ice. A 250 µl cell pellet was lysed on ice (5 min) by resuspension in 500 µl of Pierce IP Lysis Buffer (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% NP-40 and 5% glycerol; Thermo Fisher) containing 1x protease inhibitor cocktail III (Calbiochem). During the 5 minutes, cells were periodically dounce homogenized with a small-clearance glass pestle (pestle B). Approximately 2 mg of total protein was loaded on either a mixed bed IEX column (PolyLC Mixed-Bed WAX-WCX, PolyLC Inc. #204CTWX0510) or a BioSep-SEC-s4000 gel filtration column (Phenomenex) equilibrated in PBS, pH 7.2. HPLC chromatography was as in [31] and collected fractions were processed as described in [163].

Lysate preparation and chromatographic separation of other species is described in [27,31,161,164].

Data acquisition and processing

All column fractions were reduced, alkylated and digested with trypsin for mass spectrometry by either method 1 or 2 of the protocols in [165]. Spectra were collected

as in [31] on either a Thermo Scientific Orbitrap Fusion Tribrid or an Orbitrap Fusion Lumos Tribrid mass spectrometer except as noted below. Euglena data were collected on a Lumos using CID (35%) and a topspeed 75 minute method as in [31]. Spectra for DSSO crosslinked *Tetrahymena* cilia IEX fractions were collected using a 2 hour DDA MS2-MS3 method as described in [161] but processed for protein identifications in this study using the MSBlender pipeline described below.

Computational analyses

Reference database construction

Protein sequences from each of the 31 reference FASTA files in **Table S4** were compared against the eggNOG 5.0 database [37] and mapped to orthologous groups (OGs) at the euNOG level (taxonomic level = 2759) using eggNOG-mapper v2.0.5 [136] with DIAMOND and a hit cut-off e-value of 10^{-3}. For each species, a reference database was constructed where proteins are binned into their respective OGs such that each FASTA entry represents a bin of proteins or protein family; this was accomplished by concatenating each sequence assigned to an OG interposed with a triple lysine sequence. Since we allow for two missed trypsin cleavages in peptide spectra assignment, this triple lysine sequence ensures that we avoid the misassignment of peptides matching a chimera of two binned sequences. The benefits of this approach are three-fold: (1) defining proteomes in terms of OGs enables cross-species comparisons, (2) OG binning recovers peptide mass spectra that otherwise could not be uniquely assigned to highly sequence-similar proteins, and as a natural extension (3) facilitates proteomic analysis of species with high ploidy, e.g., *X. laevis* (allotetraploid) [164] and *T. aestivum* (allohexaploid) [31].

Peptide mass spectra processing

Matching of mass spectra to peptides was performed with MSGF+, X!Tandem, and Comet-2013.02.0, each run with 10ppm precursor tolerance and allowing for fixed cysteine carbamidomethylation (+57.021464) and optional methionine oxidation (+15.9949). Peptide search results were integrated with MSBlender [166] as described in [27,31] with the exception that high confidence (1% FDR) peptide spectral matches were required from two out of the three peptide identification algorithms. In all, we measured 379,758,411 peptides that were uniquely assigned to 259,732 unique proteins and orthogroups across all fractions. These results were filtered such that we only retain orthogroups that (a) were determined to trace back to the last eukaryotic common ancestor and (b) were strongly observed such that the sum total peptide spectral matches (PSMs) across all fractionations was >= 150.

Feature curation for protein-protein interactions

For each orthogroup found in each MS fractionation for each species sample, an elution vector was constructed by concatenating the peptide spectral counts for each orthogroup in each fraction. Four measures were used to compare all pairwise elution vectors: the Pearson correlation coefficient, Spearman's correlation coefficient, Euclidean distance, Bray-Curtis dissimilarity. These measures were computed as described in [165] and were generated for: (1) vectors for 149 individual fractionations, (2) concatenated vectors that include all samples within the Amorphea eukaryotic supergroup, (3) concatenated vectors that include all samples within the Excavate eukaryotic supergroup, (4) concatenated vectors that include all samples within the TSAR eukaryotic supergroup, (5) concatenated vectors that include all samples within the Archaeplastida eukaryotic supergroup, and (6) concatenated vectors that include all eukaryotic samples, netting 616 CFMS features.

In order to specifically target conserved pan-eukaryotic protein interactions, we required elution vectors for each protein-protein interaction (PPI) to have a minimum Pearson *r* of 0.3 and be observed in at least two of the four eukaryotic supergroups, *i.e.*, Amorphea, Excavata, TSAR, and/or Archaeplastida (**Figure 2B**). This reduced the size of our input data from 17,895,154 pairwise protein comparisons to a curated set of 4,491,719 highly conserved PPIs. Finally, we integrated the intersection of these conserved PPIs with 47 pairwise features generated from an orthogonal collection of ~15,000 mass spectrometry proteomic experiments [28] that include APMS [50–52], proximity labeling [53,54], and RNA-pulldown data [55] to attain our final PPI feature matrix, resulting in a total of 663 features for each of 4,491,719 highly conserved potential pairwise PPIs.

Assembly of gold standard protein complexes

Gold standard protein interactions were downloaded from the CORUM [56] (http://mips.helmholtz-muenchen.de/corum) and Complex Portal [57,58] (https://www.ebi.ac.uk/complexportal) databases. Both databases include protein-protein interactions for multiple species, spanning multiple mammals in CORUM (human, rat, mouse, cow, pig) and many eukaryotes in Complex Portal (human, rat, mouse, cow, pig, yeast, *Arabidopsis*, worm, fly, chicken, snake, fish, frog, rabbit). Redundant complexes were merged, and (to reduce representational bias) any complex with >30 subunits was removed from the gold standard complex set. Finally, UniProt IDs were matched to euNOG IDs and the gold standard complexes were pruned to only include those in the LECA proteome as determined by the ancestral state reconstruction described above.

Machine learning for protein interactions

All gold standard PPIs observed in our filtered data set were labeled as positive interactions. Negative interactions were defined as interactions between proteins in different gold standard complexes (e.g., given two gold standard heterotrimers A-B-C and X-Y-Z, data corresponding to an A-X protein pair would be labeled as a negative interaction). To mitigate class imbalance, the total number of negative labels was limited to 3X the observed number of positive PPIs in our data, resulting in 6,629 total positive PPI and 19,887 total negative PPI labels in our feature matrix.

Positive PPIs that participate in multiple complexes are a potential source of representation bias in the truth set, which can lead to under or overfitting during model training. To overcome this, we implemented a data stratification approach (**Figure S4**). First, all gold standard protein complexes are given a unique numeric ID. All protein pairs within a complex inherit that ID. Negative PPIs receive group labels by randomly sampling the distribution of positive PPI group IDs with replacement. If a protein pair participates in >1 complex, that pair will be labeled with a list of IDs. These ID lists represent networks of overlapping complexes. We implemented transitive closure of the networks by recursively merging ID lists that overlap with each other, netting a fully stratified “supergroup” label for each gold standard PPI in the data set.

A group-based split method (scikit-learn’s “GroupShuffleSplit” class) was used to generate 5 sets of test and training data, where 75% of the labeled data was used for training and 25% for testing. Each of the 5 training sets were used as input into TPOT [167], an automated machine learning pipeline built on top of scikit-learn, to find the best classification method, pre-processing steps, and parameters for our data. While TPOT generates an internal cross-validation score to evaluate the performance of different models, the pipeline is agnostic to strata within the training set and is thus subject to overfitting. In each instance, the 25% hold-out set was used as a true test of the optimized models produced by TPOT. These results are reported in **Table S5**. In the majority of cases, TPOT reports the extremely randomized trees (ExtraTrees) algorithm with slightly varying parameters and pre-processing steps as the best model for our data. However, both a stochastic gradient descent and linear support vector classification (SGDClassifier and LinearSVC, respectively) pipeline scored comparably to the ExtraTrees method, so we moved forward with feature selection and model assessment for those pipelines as well.

To further reduce the risk of overfitting, recursive feature elimination with cross-validation (RFECV) was used to obtain feature importances and an optimal feature set for each of the chosen models (see Zenodo data repository). Feature importance was evaluated using either the Gini index (in the case of ExtraTrees) or the absolute value of the coefficients of the linear model (in the case of LinearSVC and

SGDClassifier). While RFECV generates an internal cross-validated test score to determine an optimal feature set, the module is agnostic to stratified data and is thus subject to sampling bias. Again, we employed a custom group-based split approach (scikit-learn's "GroupKFold" class, illustrated in **Figure S5**) to generate 5 sets of test and training data. Since GroupKFold generates test/train splits such that every protein complex supergroup is included in the test data at least once, this allows us to use hold-out test sets to gauge bias in the "best" feature sets output by RFECV while also evaluating feature importance stability.

RFECV determines an optimal feature set following these steps: (1) the input training data is used to fit a given model, resulting in either a Gini index value or coefficient for each feature in the data set; (2) the least important feature(s) are recursively removed and the mean test accuracy is computed using cross-validation across the input training set, (3) the optimal number of features is determined such that mean test accuracy is maximized. Then, we use the holdout test set to evaluate the true performance of the final model as output by RFECV method. True feature importance was assessed by aggregating the RFECV results of each split, and, for each feature, counting the number of times it appears in the "optimal" set output and computing the mean and relative standard deviation of its Gini index/coefficient. The "best" features are those that maximize the number of appearances in the final "optimal" set across GroupKFolds splits, maximize the absolute value of the Gini index/coefficient and minimize relative standard deviation (*i.e*., low RSD indicates the feature importance is stable across GroupKFold train/test splits).

We measured precision and recall for each model using the top 5, 10, 25, 50, 100, and 250 highest ranked features (determined on a model-by-model basis, in other words, the "top" features for the LinearSVC classifier are different than the "top" features for the ExtraTreesClassifier) and all 663 features (**Figure S6A**), choosing the feature set that had the highest number of PPIs and unique proteins within a 10% FDR threshold (**Figure S6B,C**).

Community detection of protein complexes

Communities of interacting proteins were identified using the walktrap algorithm *via* the igraph library interface in Python. Briefly, the walktrap algorithm detects community structure by executing a user-defined number of random walks from each vertex in a graph. Random walks tend to become "trapped" in strongly connected sub-communities of the graph [59], a behavior that we reinforce by weighting graph edges with the probability scores output by our three PPI models (LinearSVC, ExtraTrees, and SGDClassifier). Then, individual protein complex clusters are partitioned such that the modularity of the network is maximized [168], netting an

"optimal" number of protein complex communities given the input graph structure. The walktrap algorithm performed best with the PPI scored output by the LinearSVC model, and was thus chosen as our final interactome. The features used in the final model are reported in the Zenodo repository.

Validation of the protein interaction model with external data

External interaction datasets were sourced from [60–62]. The likelihood of LECA PPIs (I_L) agreeing with external interaction networks (I_E) was calculated with a formula analogous to an odds ratio, described by the equations below.

$$A = I_E \cap I_L \qquad \text{Eq. 1}$$
$$B = I_E \setminus I_L \qquad \text{Eq. 2}$$
$$C = I_L \setminus I_E \qquad \text{Eq. 3}$$
$$D = I \notin (I_E \cup I_L) \qquad \text{Eq. 4}$$
$$A + B + C + D = \frac{N_i(N_i - 1)}{2} \qquad \text{Eq. 5}$$

		LECA interaction (I_L)	
		+	−
External interaction (I_E)	+	***A***	***B***
	−	***C***	***D***

$$\frac{P(I_{L+} \mid I_{E+})}{P(I_{L+} \mid I_{E-})} = \frac{A/(A+B)}{C/(C+D)} = \frac{A(C+D)}{C(A+B)} \qquad \text{Eq.6}$$

Challenges and limitation to mapping the LECA protein interaction set

It is important to note that the data used in this study strongly favors humans and mammals in general. Most of the CFMS and APMS experiments in this study are sourced from humans. Of the 25,000 experiments included in this study, approximately 18,000 are derived from human cells. Gold standard protein complexes obtained from the CORUM database are exclusively mammalian, which motivated us to also incorporate gold standard PPIs from the ComplexPortal database. However, though ComplexPortal is more diverse and includes *Arabidopsis* assemblies, the majority of the data is still Amorphean. We took a number of steps to ensure we target pan-eukaryotic proteins and protein interactions; for example, we filtered the MS data to only include proteins that we trace back to LECA as well as require every PPI be observed and reasonably correlated by CFMS in at least 2 of the 4 eukaryotic supergroups (as defined in **Figure 2B**) prior to entry into the machine learning pipeline.

Co-fractionation mass spectrometry identifies stable complexes that survive biochemical fractionation. Stochastic sampling across a large number of species and

fractionations allows us to recover some transient interactions. CFMS is also biased towards abundant and soluble proteins, though we typically employ detergents to improve coverage of membrane proteins. We measure approximately 60% of the 10,092 OGs we trace to LECA (probably due to the above biases) and high precision PPIs for half of these, indicating a high false negative rate. Inclusion of the APMS data sets increases the power of our model for PPI detection but only for systems conserved within Amorphea. Binning proteins into evolutionarily related protein families (orthogroups) prior to peptide identification and assignment results in loss of resolution for different isoforms and some paralogs. For the most part, we are able to draw conclusions about the properties of families of genes rather than individual members. In some cases, we can retroactively disentangle which variant(s) or paralog(s) participate in a specific interaction by examining the specific peptides identified by mass spectrometry.

Resources for disease analyses

We downloaded 17,019 gene-disease relationships from the Online Mendelian Inheritance in Man database (omim.org) [116].

Curation of a gene-disease data set

Ensembl accessions from the OMIM data set were first mapped to human UniProt identifiers and then subsequently matched with their corresponding eggNOG orthologous groups (OGs) at the root of eukaryotes (NCBI taxonomy level = 2759). Gene-disease assignments in the raw OMIM data did not follow a standardized schema and required a combination of programmatic and manual cleaning. For example, AKT1, PTEN, KLLN, and SEC23B are respectively assigned to Cowden syndrome 6, Cowden syndrome 1, Cowden syndrome 4, and Cowden syndrome 7 in the original data set. After cleaning, these genes are grouped under a common “Cowden syndrome” label. We filtered the data set to contain only LECA OGs, netting 5,761 genotype-to-phenotype associations for the network propagation of gene-disease relationships in the conserved eukaryotic interactome. In total, we curated 1,683 unique disease labels for 2,262 highly conserved human genes.

Network propagation

We used a cross-validated network propagation approach to systematically assign disease predictions to proteins in the interactome. If a protein in the network is known to be associated with a disease, then each connecting node receives the score of the connecting edge. This process is repeated for each unique disease label in our OMIM data set given that the disease has at least five mapped associations with genes that also map to an orthologous group in the last eukaryotic common ancestor. To assess the quality of this propagation for each disease, we iteratively leave out true

positive nodes and query how well the propagation recapitulates known gene-disease relationships (*i.e.*, leave-one-out cross validation). We calculate true and false positive rates to construct a receiver operating characteristic (ROC) curve as a function of propagated score. Then, we use the area under the ROC curve (AUROC) as a measure of performance. Additionally, for each disease, we repeat propagation from randomly selected nodes from the network to evaluate the statistical strength of the gene-disease network versus randomly assigned gene-disease relationships.

Experimental analyses of candidate disease genes

Genetic knock outs in Mus musculus

We sourced *Atp6v1a* knockout data (with permission) from the Knockout Mouse Program (KOMP2) sited at the Baylor College of Medicine, which resides under the umbrella of the International Mouse Phenotyping Consortium (IMPC). The IMPC aims to systematically phenotype mice that are homozygous for a single-gene knockout or heterozygous when homozygotes are lethal or sub-viable [169]. Within the IMPC, KOMP2 production centers use the high-throughput and rigorously standardized IMPReSS pipeline to generate and phenotype single-gene knockout mice. Methods for generating single-gene null alleles are described in [170] and phenotype data collection procedures can be accessed in detail at https://www.mousephenotype.org/impress/index.

Atp6v1a mutant alleles were generated by KOMP using CRISPR/Cas9 to introduce a critical exon deletion in a murine C57BL/6N background. Phenotypes were measured from postnatal mice following the embryonic and early adult IMPC pipelines. In the case of *Atp6v1a*, homozygous knockouts resulted in complete penetrance of pre-weaning lethality. As a result, heterozygous knockouts were generated for 8 female mice and 8 male mice. At 14 weeks, bone mineral content and density was measured for each *Atp6v1a*$^{em1(IMPC)Bay/+}$ mutant using dual-energy X-ray absorptiometry (DEXA). The IMPC uses the PhenStat R package to identify abnormal phenotypes from high-throughput pipelines [171]; for the *Atp6v1a*$^{em1(IMPC)Bay/+}$ mutants, a linear mixed model factoring in the effects of sex and body weight was implemented in PhenStat to assess significance of differential bone mineral content.

EFHC2 patient genetics

Individual A4237-22 was a male of Egyptian origin who was diagnosed with small kidneys, increased echogenicity, cortical and medullary cysts, and microcephaly. To identify a potential genetic cause for the individual's phenotype, authors S.S. and F.H. performed whole exome sequencing (WES) analysis on individual A4237-22. Given the parents' unaffected status regarding their renal phenotype, a recessive mode of inheritance was hypothesized.

Homozygosity mapping revealed only 4.4 Mb of homozygosity, confirming the non-consanguinity of the parents (**Figure S2A**). We detected a hemizygous X-linked missense variant in A4237-22 (c.398G>A; p.Arg133His) (**Figure 6D, S2B**). The variant has not been reported as homozygously or hemizygously in the gnomAD database in 166,211 control individuals. The p.Arg133His amino acid resides in the DM10 domain (**Figure 6E**).

The patient's DNA was also screened for potentially deleterious variants in all genes known to cause kidney disease without results.

Protein localization and knockdown experiments in Xenopus laevis

Xenopus embryo manipulations were performed as in [172–174]. Briefly, female adult *Xenopus* were ovulated by injection of hCG (human chorionic gonadotropin). In vitro fertilization was carried out by homogenizing a small fraction of a testis in 1X Marc's Modified Ringer's (MMR). Embryos were dejellied in 1/3X MMR with 2.5% cysteine (pH 7.8) at the two-cell stage. For microinjections, embryos were placed in a 2% Ficoll and 1/3X MMR solution, injected with mRNA using forceps and an Oxford universal micromanipulator, and washed with 1/3X MMR after 2 hours.

The full length sequences of *Xenopus EFHC2* and *GLG1* were downloaded from Xenbase [175]. The DNAs corresponding to the open reading frames (ORFs) of *EFHC2* and *GLG1* were amplified from *Xenopus* cDNA and were cloned into a pCS10R MCC vector containing an N-terminal GFP or a C-terminal FLAG tag driven by an MCC specific alpha tubulin promoter, respectively. The pCS10R MCC *GFP-EFHC2* R133H construct was generated by site-directed mutagenesis (NEB, #E0554S) from pCS10R MCC *GFP-EFHC2*. Capped mRNAs were synthesized using the mMESSAGE mMACHINE SP6 transcription kit (Invitrogen Ambion, #AM1340). A morpholino antisense oligonucleotide (MO) against *GLG1* was designed to block translation (GeneTools). The MO sequence is 5'-CCATCTTGGGAAGTGCTAGTCAAG-3'.

mRNA and MO were injected into two ventral blastomeres of 4-cell stage Xenopus embryos in 2% Ficoll (w/v) in 1/3 X MMR and the injected doses of mRNAs or MO per cell are as follows: *GFP-EFHC2* and *GFP-EFHC2* R133H (78 pg), *GFP-IFT56* and *GFP-IFT80* (100 pg) [131], membraneRFP(50 pg), *GLG1-FLAG* for rescue experiment (700 pg), and *GLG1* MO (30 ng) for the knockdown experiment. Live images were captured at stage 23 or stage 25 with LSM700 inverted confocal microscope (Carl Zeiss) with a Plan-APOCHROMAT 63×/1.4 oil immersion objective or Nikon eclipse Ti confocal microscope with 60×/1.4 oil immersion objective. Imaging analysis was

performed using Fiji. Bonferroni-adjusted *p*-values were calculated in R using the base stats package.

FIGURE LEGENDS

Figure 1. Inferred subcellular organization in LECA, the last eukaryotic common ancestor, based on its estimated gene content. Cell illustration adapted from multiple graphics sourced from SwissBioPics [41].

Figure 2. Overview of experimental and computational methods. (A) Schematic representation of a co-fractionation mass spectrometry experiment. **(B)** Proteomics data used to construct the LECA interactome included eukaryotes spanning ~1.8 billion years of evolution. Tree structure is based on [26]. Branch lengths are not drawn to scale. **(C)** Schematic overview of the approach for computing protein-protein interaction (PPI) features based on CFMS (1) and APMS (2) datasets, scoring conserved PPIs based on these features (3), and clustering scored PPIs into complexes (4).

Figure 3. Determining the LECA protein interactome. Co-elution matrix and results of the protein interaction machine learning pipeline. **(A)** Heat map of the filtered elution matrix for 5,989 strongly observed LECA OGs across 10,481 CFMS mass spectrometry fractions (left) and a blow-up of elution vectors for the COPI, 20S proteasome, and eukaryotic initiation factor 3 complexes for a select subset of species (right). **(B)** Precision-recall performance of three classifiers trained with increasingly larger sets of ranked features. **(C)** Precision-recall curves for the reconstruction of known protein complexes defined by a walktrap algorithm, where pairwise PPI scores from each classifier are used as input. Points are labeled with the total number of protein clusters (complexes) constructed at each point in the hierarchy. **(D)** The likelihood that PPIs in our network are present in externally defined protein-protein or mRNA coexpression networks as a function of our model's PPI score. As PPI scores increase, our model becomes increasingly likely to agree with external studies.

Figure 4. Visualizing hierarchical clustering of protein complexes for a subset of the conserved eukaryotic interactome. The circles of the smallest diameter correspond to individual proteins, where their colors correspond to whether the proteins within each cluster are characterized to interact with each other in the literature (red), whether a novel protein is interacting with a known complex (blue), or whether all the associations within a cluster are uncharacterized (yellow).

Figure 5. Notable LECA systems related to vesicle tethering and cell projection. (A) Node colors for each vesicle tethering complex correspond to their primary subcellular localization: endoplasmic reticulum (light green), Golgi apparatus (dark green), digestive vesicles (orange), or endosomes (yellow). **(B)** Dark and light blue nodes depict core and peripheral cell projection components. In both **(A)** and **(B)**, edges

between proteins are colored by the number of eukaryotic supergroups in which the interaction is observed: red for interactions observed in all supergroups considered, orange for interactions observed in three of the four eukaryotic supergroups, and yellow for interactions only observed in half of the supergroups. The four supergroups considered are Amorphea, Excavata, TSAR and Archaeplastida (see **Figure 2**).

Figure 6. LECA protein interactions suggest mechanisms of genetic disease, as for end stage renal disease gene *EFHC2*, identified by whole exome sequencing and confirmed to have a ciliary etiology. (A) Causal genes for human diseases are frequently ancient, as shown by plotting gene-disease relationships obtained from OMIM, with each point representing a unique disease group with an associated number of genes (x-axis) and age, determined as the percentage of genes in LECA OGs (y-axis). **(B)** Pedigree of the index family A4237. Squares represent males, circles females, black shading the affected proband individual A4237-22 included in whole-exome sequencing (WES), and white shading the unaffected parents and siblings. **(C)** Summary of the phenotype and recessive disease-causing R133H *EFHC2* variant identified by WES. **(D)** Location of Arginine 133 in relation to *EFHC2* exon/intron (black/white) structure and DM10 protein domains (purple), and its deep evolutionary conservation. **(E)** EFHC2-containing ciliary complex uncovered in the LECA interactome. **(F)** Localization of GFP-EFHC2 to axonemes in *Xenopus* motile cilia. **(G)** Introduction of the R133H mutation results in loss of ciliary localization of GFP-EFHC2, confirmed by co-labeling with membrane-RFP. Scale bar = 10 μm

Figure 7. Guilt-by-association in the LECA interactome identifies *ATP6V1A* as causative for osteopetrosis and *GLG1* for short-rib thoracic dysplasia (SRTD). (A) Guilt-by-association in the LECA PPI network correctly associates genes to human diseases for roughly a third of the 109 diseases tested, measured as the areas under receiver operating characteristic curves (AUROCs) of leave-one-out cross-validated predictions of known disease genes (light blue) versus random associations (yellow). **(B)** PPI network of genes clinically linked to osteopetrosis (black half-discs; 3 additional genes lie outside this cluster), the highest-ranking new candidates (purple), and their interactions with other V-ATPase subunits that were not indicated for osteopetrosis (orange). **(C)** For the top-scoring gene *ATP6V1A*, the bone mineral content is plotted for knockout (KO) mice with a heterozygous exon deletion in *ATP6V1A* (n=8 for each sex, n=16 total) compared to healthy control mice (female n=834, male n=780). Null mice show significantly increased bone density, consistent with the clinical manifestation of osteopetrosis. **(D)** The PPI network of genes clinically linked to SRTD (black half-discs) implicates *GLG1* (yellow) and suggests a ciliary role, based on interactions with intraflagellar trafficking IFT-A (blue) and IFT-B (purple) complexes, cytoplasmic dyneins and dynactins (green), and other interactors (gray). **(E)** Morpholino knockdown (KD) of

GLG1 significantly reduced the number of cilia in *X. laevis* multi-ciliated cells (Bonferroni adjusted t-test $p < 10^{-16}$, $n = 60$ control cells, 79 knockdown cells, and 76 rescue cells, 9 embryos per condition over 3 injection replicates) compared to uninjected control animals; rescue by co-injection with a non-targeted *GLG1* allele confirmed specificity. **(F)** In control *Xenopus* multi-ciliated cells, IFT56-GFP and IFT80-GFP, two subunits of IFT-B, are distributed as particles along the ciliary axonemes. However, MO knockdown of *GLG1* leads to the accumulation of IFT-B proteins in the proximal region of axonemes. Scale bar = 10 µm. **(G)** This effect is quantified for IFT80-GFP for 3 cilia per cell for all cells analyzed in panel **(E)**.

Figure 1.

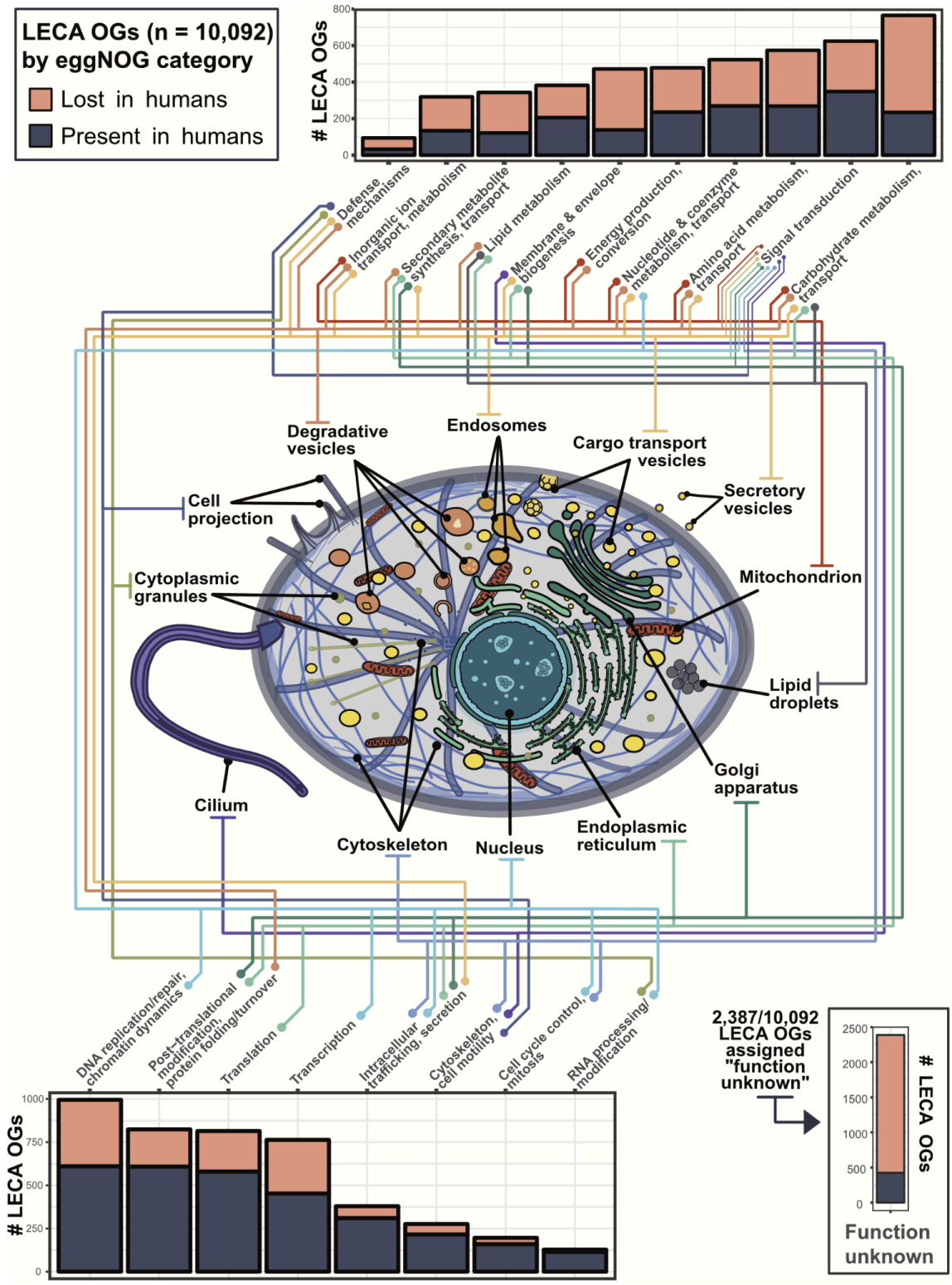

LECA OGs (n = 10,092) by eggNOG category
Lost in humans
Present in humans
LECA OGs
Defense mechanisms
Inorganic ion transport, metabolism
Secondary metabolite synthesis, transport
Lipid metabolism
Membrane & envelope biogenesis
Energy production, conversion
Nucleotide & coenzyme metabolism, transport
Amino acid metabolism, transport
Signal transduction
Carbohydrate metabolism, transport
Degradative vesicles
Endosomes
Cargo transport vesicles
Secretory vesicles
Cell projection
Cytoplasmic granules
Mitochondrion
Lipid droplets
Golgi apparatus
Cilium
Cytoskeleton
Nucleus
Endoplasmic reticulum
DNA replication/repair, chromatin dynamics
Post-translational modification, protein folding/turnover
Translation
Transcription
Intracellular trafficking, secretion
Cytoskeleton, cell motility
Cell cycle control, mitosis
RNA processing/ modification
2,387/10,092 LECA OGs assigned "function unknown"
Function unknown

Figure 2.

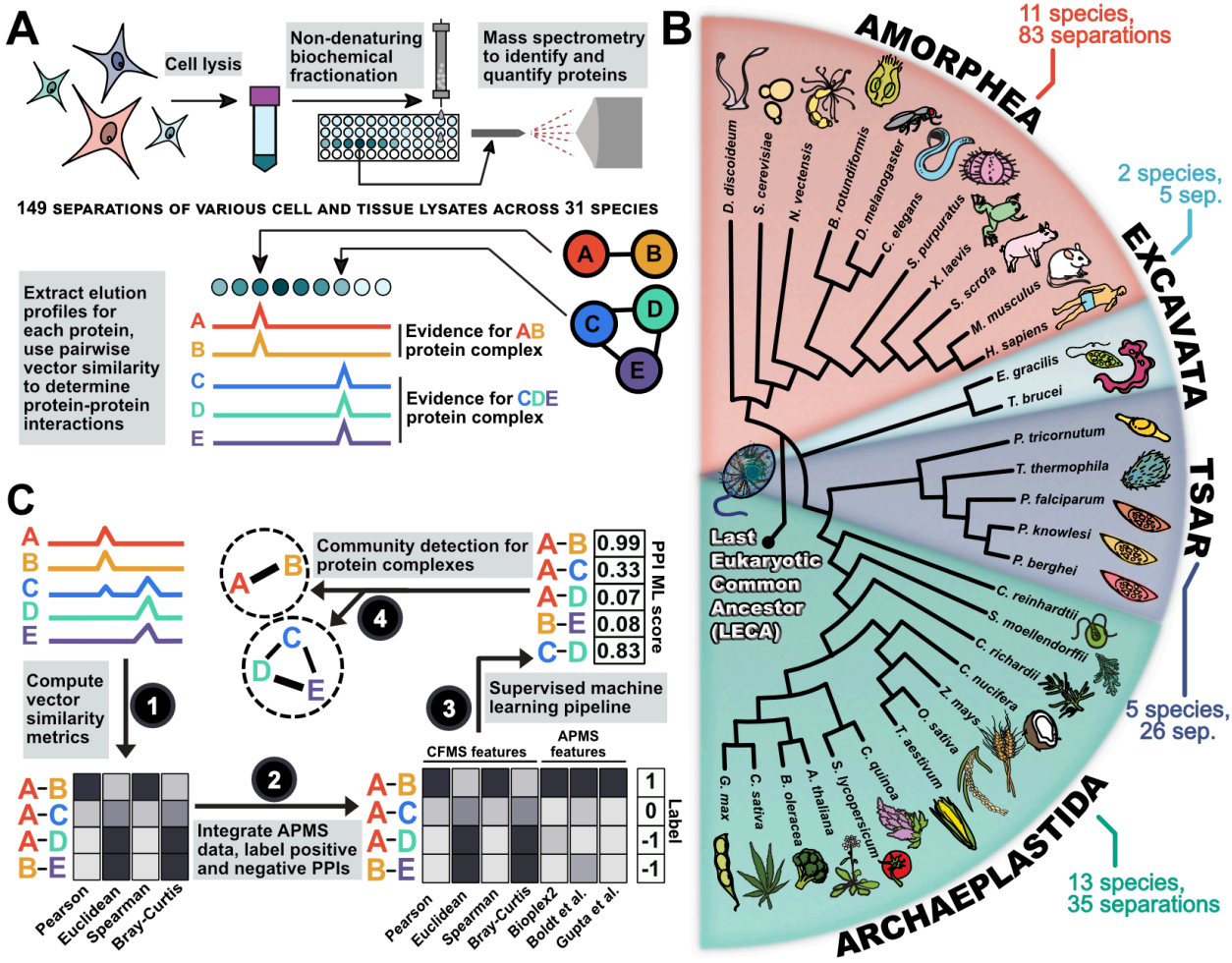

A
Cell lysis
Non-denaturing biochemical fractionation
Mass spectrometry to identify and quantify proteins
149 SEPARATIONS OF VARIOUS CELL AND TISSUE LYSATES ACROSS 31 SPECIES
Extract elution profiles for each protein, use pairwise vector similarity to determine protein-protein interactions
Evidence for AB protein complex
Evidence for CDE protein complex
B
AMORPHEA
11 species, 83 separations
EXCAVATA
2 species, 5 sep.
TSAR
5 species, 26 sep.
ARCHAEPLASTIDA
13 species, 35 separations
Last Eukaryotic Common Ancestor (LECA)
D. discoideum
S. cerevisiae
N. vectensis
B. rotundiformis
D. melanogaster
C. elegans
S. purpuratus
X. laevis
S. scrofa
M. musculus
H. sapiens
E. gracilis
T. brucei
P. tricornutum
T. thermophila
P. falciparum
P. knowlesi
P. berghei
C. reinhardtii
S. moellendorffii
C. richardii
C. nucifera
Z. mays
O. sativa
T. aestivum
C. quinoa
S. lycopersicum
A. thaliana
B. oleracea
C. sativa
G. max
C
Compute vector similarity metrics
Pearson
Euclidean
Spearman
Bray-Curtis
Integrate APMS data, label positive and negative PPIs
CFMS features
APMS features
Bioplex2
Boldt et al.
Gupta et al.
Label
Supervised machine learning pipeline
Community detection for protein complexes
PPI ML score
A-B 0.99
A-C 0.33
A-D 0.07
B-E 0.08
C-D 0.83

Figure 3.

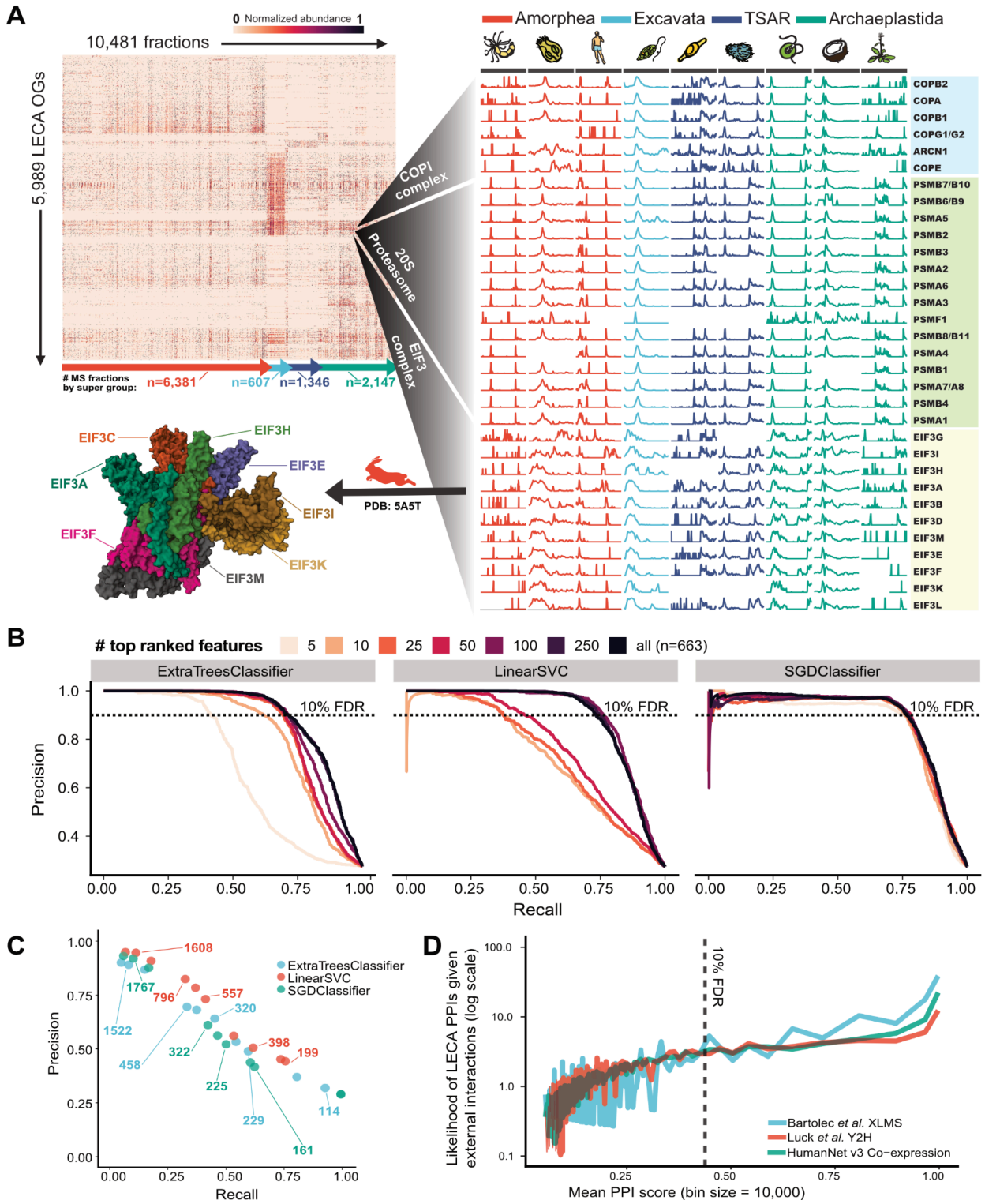

A
0 Normalized abundance 1
10,481 fractions
5,989 LECA OGs
# MS fractions by super group:
n=6,381
n=607
n=1,346
n=2,147
COPI complex
20S Proteasome
EIF3 complex
Amorphea
Excavata
TSAR
Archaeplastida
COPB2
COPA
COPB1
COPG1/G2
ARCN1
COPE
PSMB7/B10
PSMB6/B9
PSMA5
PSMB2
PSMB3
PSMA2
PSMA6
PSMA3
PSMF1
PSMB8/B11
PSMA4
PSMB1
PSMA7/A8
PSMB4
PSMA1
EIF3G
EIF3I
EIF3H
EIF3A
EIF3B
EIF3D
EIF3M
EIF3E
EIF3F
EIF3K
EIF3L
EIF3C
EIF3H
EIF3E
EIF3A
EIF3I
EIF3F
EIF3K
EIF3M
PDB: 5A5T
B
top ranked features
5
10
25
50
100
250
all (n=663)
ExtraTreesClassifier
LinearSVC
SGDClassifier
10% FDR
Precision
Recall
C
ExtraTreesClassifier
LinearSVC
SGDClassifier
Precision
Recall
D
10% FDR
Likelihood of LECA PPIs given external interactions (log scale)
Mean PPI score (bin size = 10,000)
Bartolec et al. XLMS
Luck et al. Y2H
HumanNet v3 Co-expression

Figure 4.

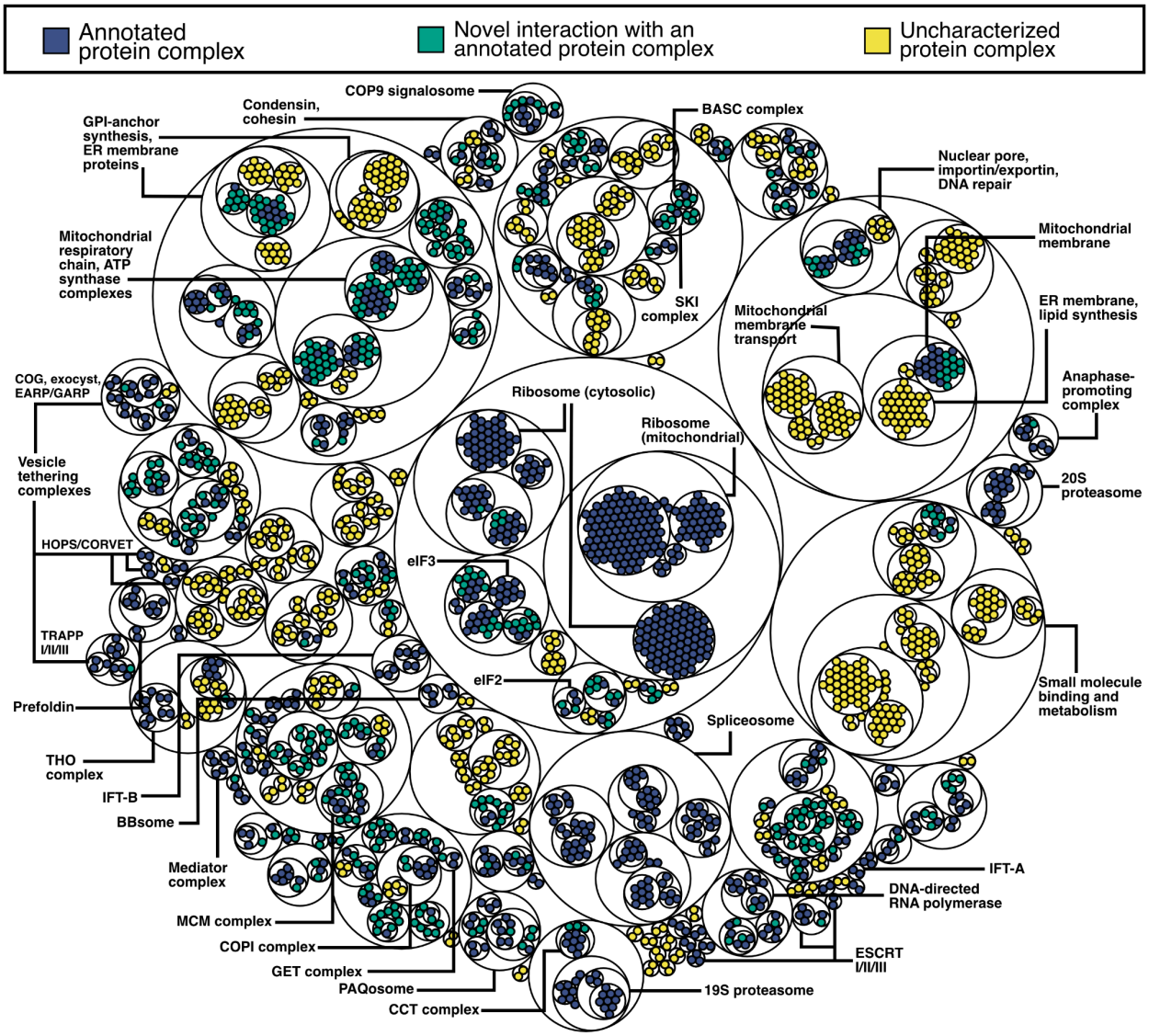

Annotated protein complex
Novel interaction with an annotated protein complex
Uncharacterized protein complex
COP9 signalosome
Condensin, cohesin
GPI-anchor synthesis, ER membrane proteins
BASC complex
Nuclear pore, importin/exportin, DNA repair
Mitochondrial respiratory chain, ATP synthase complexes
Mitochondrial membrane
SKI complex
Mitochondrial membrane transport
ER membrane, lipid synthesis
COG, exocyst, EARP/GARP
Ribosome (cytosolic)
Anaphase-promoting complex
Ribosome (mitochondrial)
Vesicle tethering complexes
20S proteasome
HOPS/CORVET
eIF3
TRAPP I/II/III
eIF2
Small molecule binding and metabolism
Prefoldin
Spliceosome
THO complex
IFT-B
BBsome
Mediator complex
IFT-A
DNA-directed RNA polymerase
MCM complex
COPI complex
ESCRT I/II/III
GET complex
PAQosome
19S proteasome
CCT complex

Figure 5.

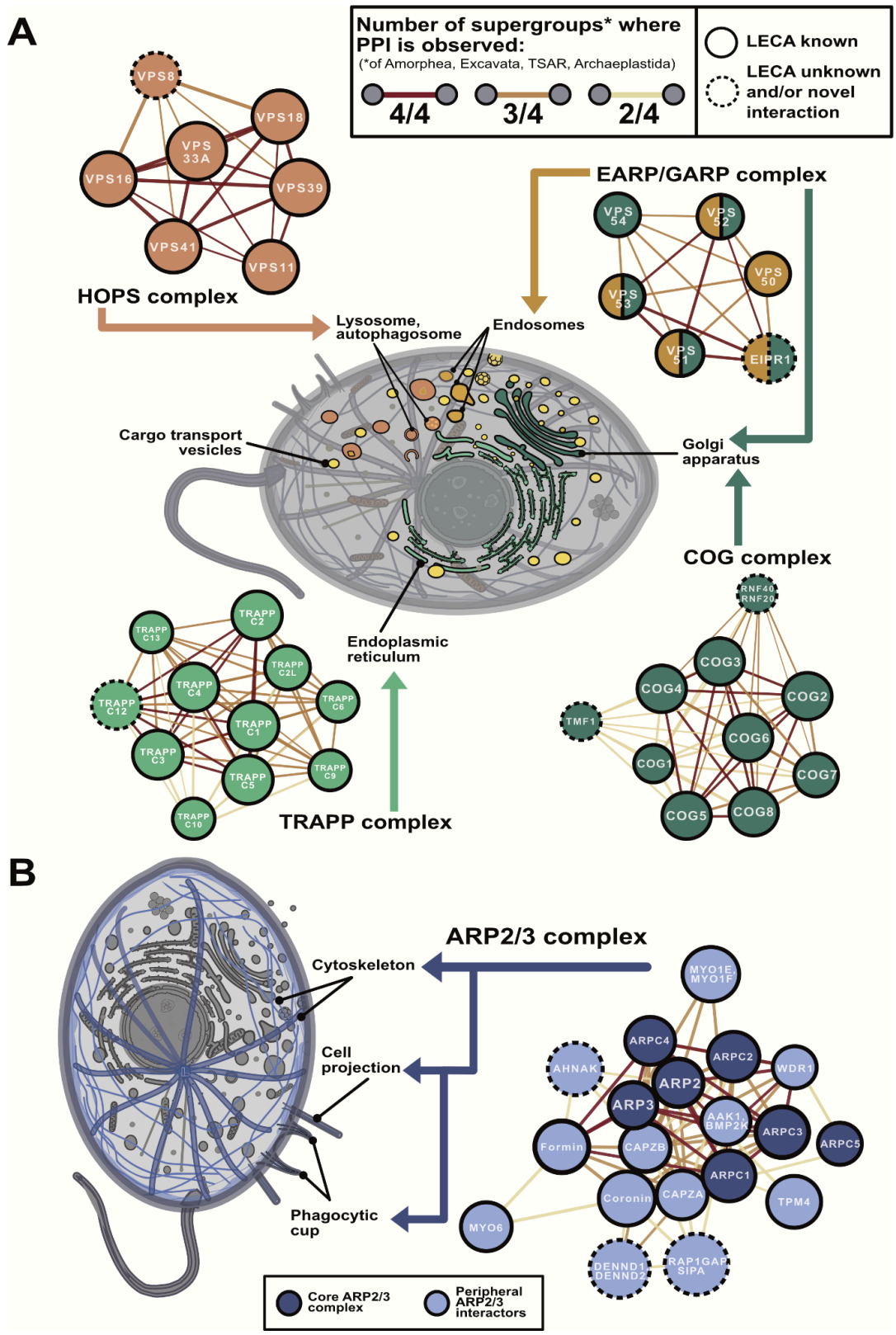

A
Number of supergroups* where PPI is observed:
(*of Amorphea, Excavata, TSAR, Archaeplastida)
4/4
3/4
2/4
LECA known
LECA unknown and/or novel interaction
VPS8
VPS18
VPS 33A
VPS16
VPS39
VPS41
VPS11
HOPS complex
EARP/GARP complex
VPS 54
VPS 52
VPS 50
VPS 53
VPS 51
EIPR1
Lysosome, autophagosome
Endosomes
Cargo transport vesicles
Golgi apparatus
COG complex
RNF40 RNF20
COG3
COG4
COG2
TMF1
COG6
COG1
COG7
COG5
COG8
Endoplasmic reticulum
TRAPP C2
TRAPP C13
TRAPP C2L
TRAPP C4
TRAPP C12
TRAPP C6
TRAPP C1
TRAPP C3
TRAPP C9
TRAPP C5
TRAPP C10
TRAPP complex
B
ARP2/3 complex
Cytoskeleton
Cell projection
Phagocytic cup
MYO1E, MYO1F
ARPC4
ARPC2
WDR1
AHNAK
ARP2
ARP3
AAK1, BMP2K
ARPC3
ARPC5
Formin
CAPZB
ARPC1
Coronin
CAPZA
TPM4
MYO6
DENND1 DENND2
RAP1GAP SIPA
Core ARP2/3 complex
Peripheral ARP2/3 interactors

Figure 6.

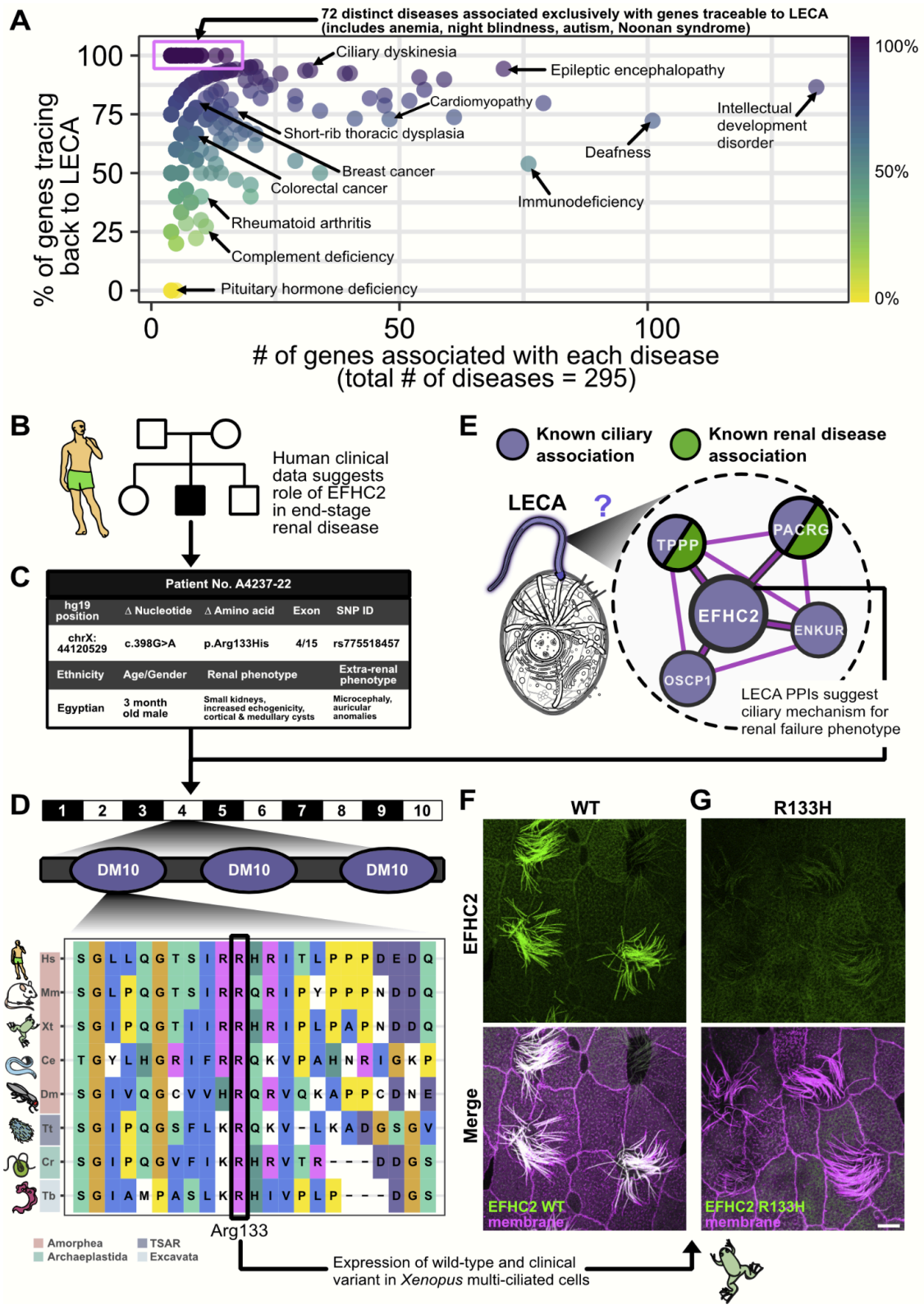
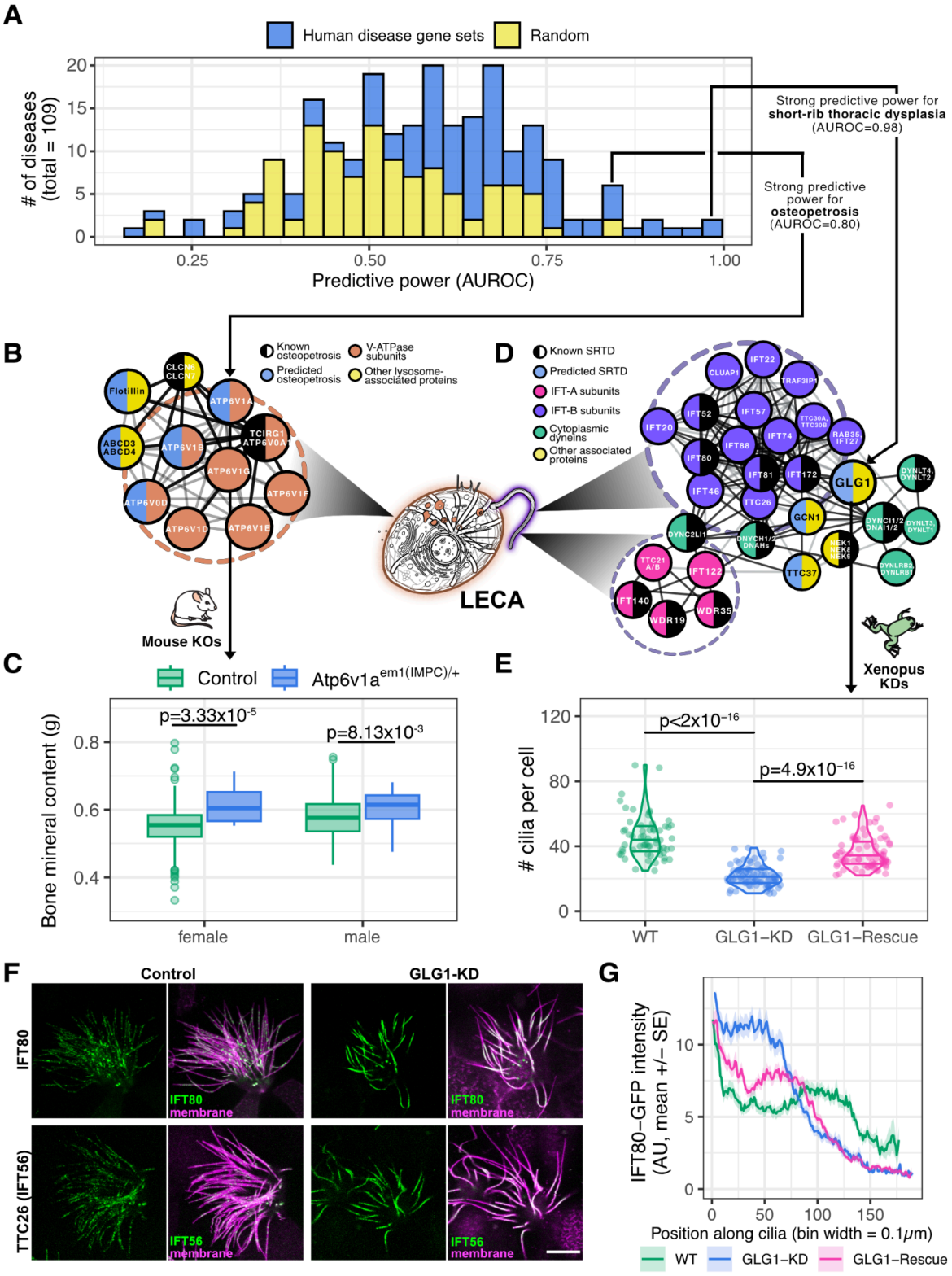

A
72 distinct diseases associated exclusively with genes traceable to LECA (includes anemia, night blindness, autism, Noonan syndrome)
Ciliary dyskinesia
Epileptic encephalopathy
Cardiomyopathy
Intellectual development disorder
Short-rib thoracic dysplasia
Deafness
Breast cancer
Colorectal cancer
Immunodeficiency
Rheumatoid arthritis
Complement deficiency
Pituitary hormone deficiency
% of genes tracing back to LECA
# of genes associated with each disease (total # of diseases = 295)
100%
50%
0%
B
Human clinical data suggests role of EFHC2 in end-stage renal disease
C
Patient No. A4237-22
hg19 position | Δ Nucleotide | Δ Amino acid | Exon | SNP ID
chrX: 44120529 | c.398G>A | p.Arg133His | 4/15 | rs775518457
Ethnicity | Age/Gender | Renal phenotype | Extra-renal phenotype
Egyptian | 3 month old male | Small kidneys, increased echogenicity, cortical & medullary cysts | Microcephaly, auricular anomalies
D
DM10
Arg133
Amorphea
Archaeplastida
TSAR
Excavata
E
Known ciliary association
Known renal disease association
LECA
TPPP
PACRG
EFHC2
ENKUR
OSCP1
LECA PPIs suggest ciliary mechanism for renal failure phenotype
F
WT
G
R133H
EFHC2
Merge
EFHC2 WT membrane
EFHC2 R133H membrane
Expression of wild-type and clinical variant in Xenopus multi-ciliated cells

Figure 7.


A
Human disease gene sets
Random
# of diseases (total = 109)
Predictive power (AUROC)
Strong predictive power for short-rib thoracic dysplasia (AUROC=0.98)
Strong predictive power for osteopetrosis (AUROC=0.80)
B
Known osteopetrosis
Predicted osteopetrosis
V-ATPase subunits
Other lysosome-associated proteins
Mouse KOs
LECA
D
Known SRTD
Predicted SRTD
IFT-A subunits
IFT-B subunits
Cytoplasmic dyneins
Other associated proteins
Xenopus KDs
C
Control
Atp6v1a em1(IMPC)/+
Bone mineral content (g)
p=3.33x10-5
p=8.13x10-3
female
male
E
cilia per cell
p<2x10−16
p=4.9x10−16
WT
GLG1−KD
GLG1−Rescue
F
Control
GLG1-KD
IFT80
TTC26 (IFT56)
IFT80 membrane
IFT56 membrane
G
IFT80−GFP intensity (AU, mean +/− SE)
Position along cilia (bin width = 0.1μm)
WT
GLG1−KD
GLG1−Rescue

SUPPLEMENTAL FIGURES AND LEGENDS

Figure S1. Phylogenetic analysis of the reviewed UniProt database by subcellular localization. Limitations to available annotations are evident in an analysis of the UniProt protein database across species, where reviewed proteins have assigned subcellular localizations likely present in the last eukaryotic common ancestor. **(A)** Light gray, total number of reviewed UniProt proteins by UniProt SL term; dark gray, total number of unique eukaryotic OGs assigned to UniProt proteins by UniProt SL term. **(B)** Phylogenetic representation of the proteins sourced from UniProt by UniProt SL term. **(C)** The percentage of eukaryotic orthologous groups (euNOGs) that trace back to LECA by UniProt SL term.

Figure S2. **Homozygosity mapping and verification of EFHC2 mutation in individual A4237-22. (A)** Homozygosity mapping depicts a homozygosity of 4.4 Mb and confirms the reported non-consanguinity of the parents. **(B)** Chromatograms obtained by direct sequencing of PCR products reveal a homozygous substitution of C for T in exon 4 of the *EFHC2* gene in A4237-22.

Figure S3. Reference species tree illustration generated by the Interactive Tree of Life [96] for most of the Quest for Orthologs benchmark species (147/156) used in the Dollo parsimony analysis. Branch lengths are not to scale. Major supergroups are highlighted across the tree. Prokaryotic groups include Bacteria (gray) and Archaea (yellow). Eukaryotic groups include Excavata (light blue), Archaeplastida (green), TSAR (purple), and Amorphea (red).

Figure S4. Illustration of transitive closure for grouping gold standard protein complexes into supergroups.

Figure S5. Illustration of group-based k-fold (in this example, k=3) cross-validation for protein-protein interactions.

Figure S6. Model selection and optimization. (A) Precision-recall curves for three different algorithms, varying the number of “top” most important features used as input (duplicate panel to Figure 3B). Feature importance is defined per algorithm, ranked by either the absolute value of coefficients for linear models (LinearSVC, SGDClassifier) or the Gini index (ExtraTreesClassifier). (B) The number of pairwise protein-protein interactions (PPIs) within a 10% FDR threshold for each model. Black stars (★) denote the final models used as input to a community detection algorithm to define protein complexes. (C) The number of unique proteins that have at least one interaction scored within a 10% FDR threshold for each model.

Figure S1.

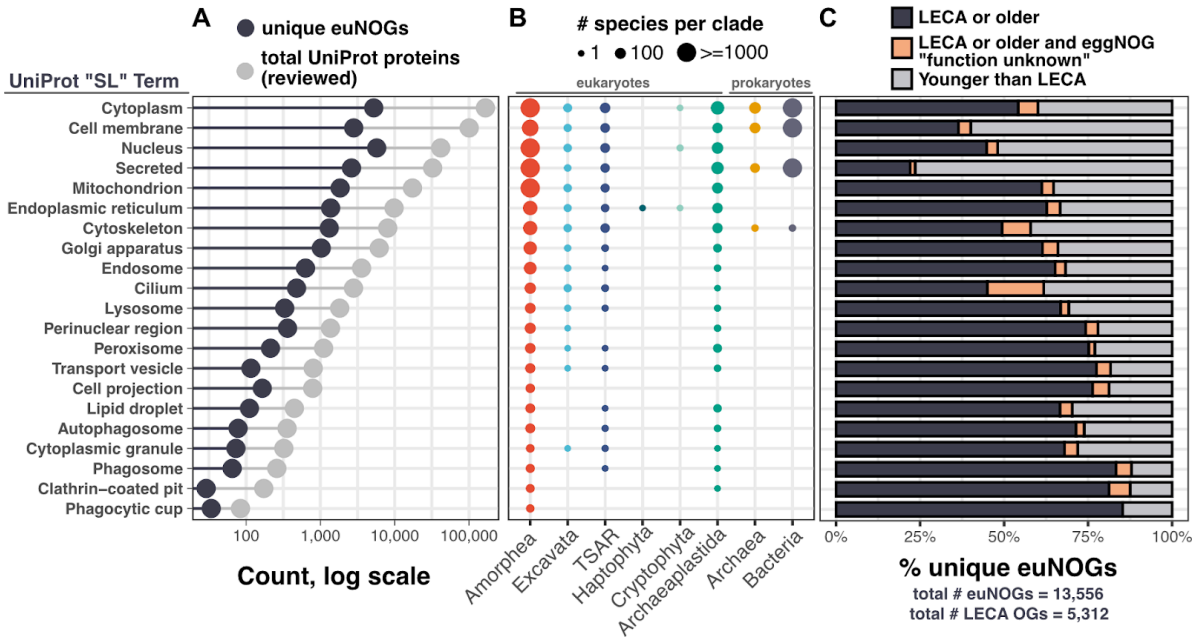

A
unique euNOGs
total UniProt proteins (reviewed)
UniProt "SL" Term
Cytoplasm
Cell membrane
Nucleus
Secreted
Mitochondrion
Endoplasmic reticulum
Cytoskeleton
Golgi apparatus
Endosome
Cilium
Lysosome
Perinuclear region
Peroxisome
Transport vesicle
Cell projection
Lipid droplet
Autophagosome
Cytoplasmic granule
Phagosome
Clathrin–coated pit
Phagocytic cup
100
1,000
10,000
100,000
Count, log scale
B
species per clade
1
100
>=1000
eukaryotes
prokaryotes
Amorphea
Excavata
TSAR
Haptophyta
Cryptophyta
Archaeaplastida
Archaea
Bacteria
C
LECA or older
LECA or older and eggNOG "function unknown"
Younger than LECA
0%
25%
50%
75%
100%
% unique euNOGs
total # euNOGs = 13,556
total # LECA OGs = 5,312

Figure S2.

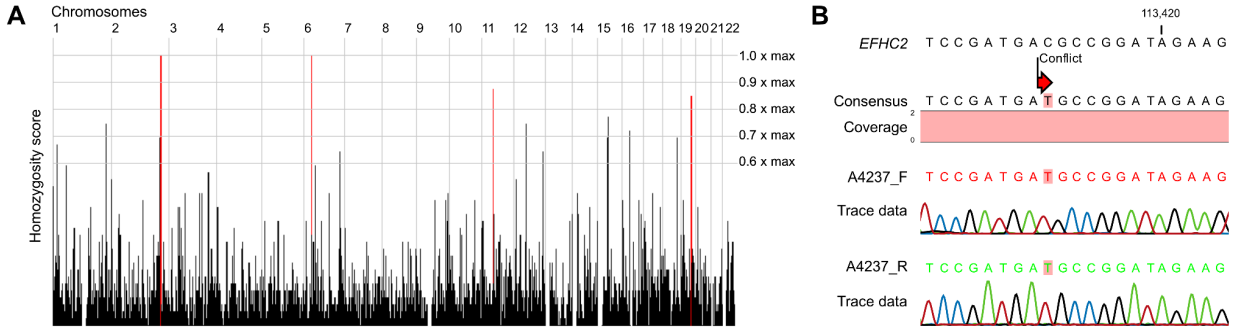

A
Chromosomes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
1.0 x max
0.9 x max
0.8 x max
0.7 x max
0.6 x max
Homozygosity score
B
113,420
EFHC2
T C C G A T G A C G C C G G A T A G A A G
Conflict
Consensus
T C C G A T G A T G C C G G A T A G A A G
Coverage
A4237_F
T C C G A T G A T G C C G G A T A G A A G
Trace data
A4237_R
T C C G A T G A T G C C G G A T A G A A G
Trace data

Figure S3.

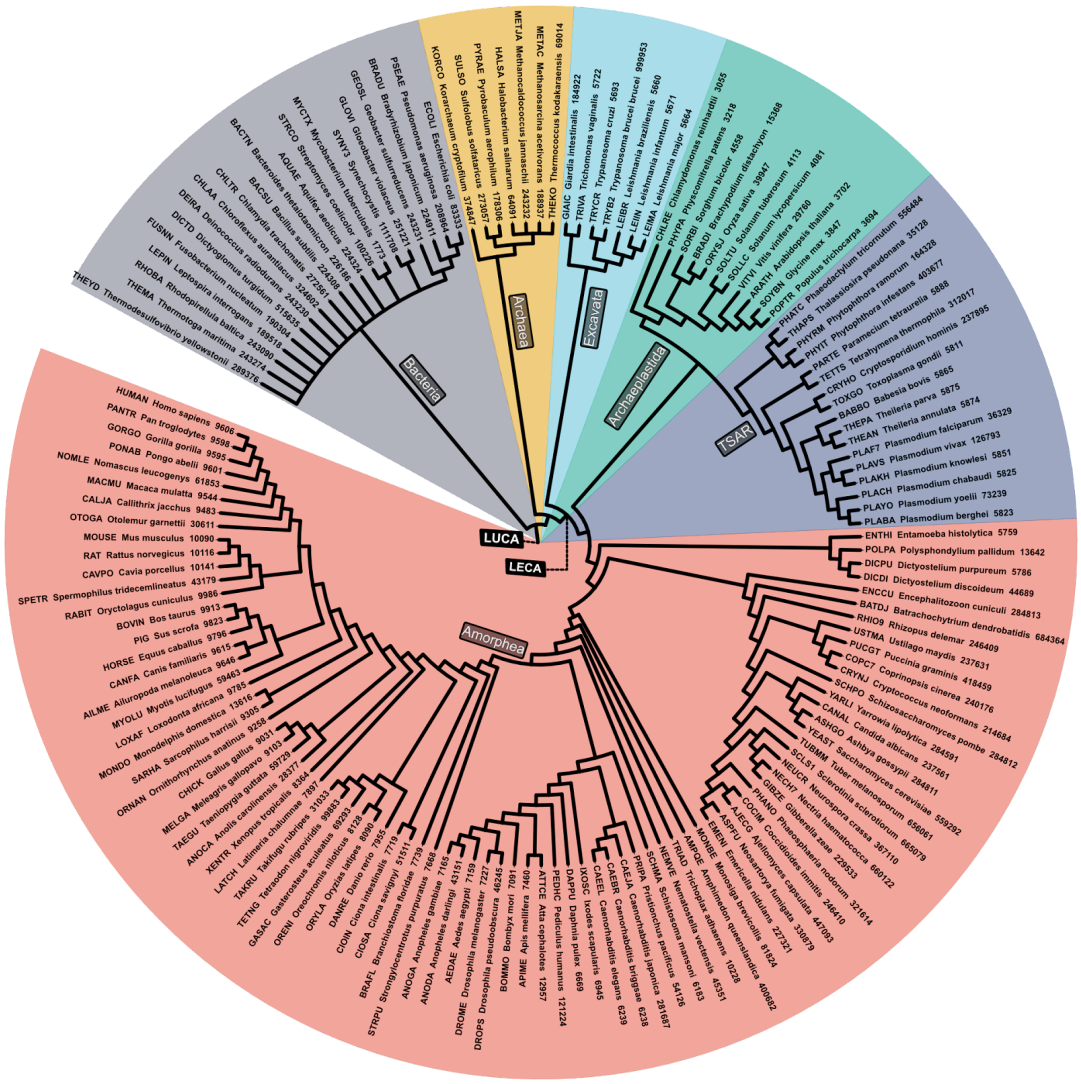

Bacteria
Archaea
Excavata
Archaeplastida
TSAR
Amorphea
LUCA
LECA
HUMAN Homo sapiens 9606
PANTR Pan troglodytes 9598
GORGO Gorilla gorilla 9595
PONAB Pongo abelii 9601
NOMLE Nomascus leucogenys 61853
MACMU Macaca mulatta 9544
CALJA Callithrix jacchus 9483
OTOGA Otolemur garnettii 30611
MOUSE Mus musculus 10090
RAT Rattus norvegicus 10116
CAVPO Cavia porcellus 10141
SPETR Spermophilus tridecemlineatus 43179
RABIT Oryctolagus cuniculus 9986
BOVIN Bos taurus 9913
PIG Sus scrofa 9823
HORSE Equus caballus 9796
CANFA Canis familiaris 9615
AILME Ailuropoda melanoleuca 9646
MYOLU Myotis lucifugus 59463
LOXAF Loxodonta africana 9785
MONDO Monodelphis domestica 13616
SARHA Sarcophilus harrisii 9305
ORNAN Ornithorhynchus anatinus 9258
CHICK Gallus gallus 9031
MELGA Meleagris gallopavo 9103
TAEGU Taeniopygia guttata 59729
ANOCA Anolis carolinensis 28377
XENTR Xenopus tropicalis 8364
LATCH Latimeria chalumnae 7897
TAKRU Takifugu rubripes 31033
TETNG Tetraodon nigroviridis 99883
GASAC Gasterosteus aculeatus 69293
ORENI Oreochromis niloticus 8128
ORYLA Oryzias latipes 8090
DANRE Danio rerio 7955
CIONI Ciona intestinalis 7719
CIOSA Ciona savignyi 51511
BRAFL Branchiostoma floridae 7739
STRPU Strongylocentrotus purpuratus 7668
ANOGA Anopheles gambiae 7165
AEDAE Aedes aegypti 43151
DROME Drosophila melanogaster 7227
DROPS Drosophila pseudoobscura 46245
BOMMO Bombyx mori 7091
APIME Apis mellifera 7460
ATTCE Atta cephalotes 12957
PEDHC Pediculus humanus 121224
DAPPU Daphnia pulex 6669
IXOSC Ixodes scapularis 6945
CAEEL Caenorhabditis elegans 6239
CAEBR Caenorhabditis briggsae 6238
PRIPA Pristionchus pacificus 54126
CAEJA Caenorhabditis japonica 281687
SCHMA Schistosoma mansoni 6183
TRIAD Trichoplax adhaerens 10228
HEMVE Hematodella vectensis 45351
MNEMI Mnemiopsis leidyi 27923
AMPQE Amphimedon queenslandica 400682
EMENI Emericella nidulans 227321
ASPFU Neosartorya fumigata 330879
COCIM Coccidioides immitis 246410
LJECG Aphanomyces euteiches 447093
PHANO Phaeosphaeria nodorum 321614
GIBZE Gibberella zeae 229533
NECHT Nectria haematococca 660122
NEUCR Neurospora crassa 367110
SCLS1 Sclerotinia sclerotiorum 665079
TUBMM Tuber melanosporum 656061
YEAST Saccharomyces cerevisiae 559292
ASHGO Ashbya gossypii 284811
CANAL Candida albicans 237561
YARLI Yarrowia lipolytica 284591
SCHPO Schizosaccharomyces pombe 284812
CRYNJ Cryptococcus neoformans 214684
COPC7 Coprinopsis cinerea 240176
PUCGT Puccinia graminis 418459
USTMA Ustilago maydis 237631
RHIO9 Rhizopus delemar 246409
BATDJ Batrachochytrium dendrobatidis 684364
ENCCU Encephalitozoon cuniculi 284813
DICDI Dictyostelium discoideum 44689
DICPU Dictyostelium purpureum 5786
POLPA Polysphondylium pallidum 13642
ENTHI Entamoeba histolytica 5759
PLABA Plasmodium berghei 5823
PLAYO Plasmodium yoelii 73239
PLACH Plasmodium chabaudi 5825
PLAKH Plasmodium knowlesi 5851
PLAVS Plasmodium vivax 126793
PLAF7 Plasmodium falciparum 36329
THEAN Theileria annulata 5874
THEPA Theileria parva 5875
BABBO Babesia bovis 5865
TOXGO Toxoplasma gondii 5811
CRYHO Cryptosporidium hominis 237895
TETTS Tetrahymena thermophila 312017
PARTE Paramecium tetraurelia 5888
PHYIT Phytophthora infestans 403677
PHYRM Phytophthora ramorum 164328
PHATG Phaeodactylum tricornutum 556484
THAPS Thalassiosira pseudonana 35128
POPTR Populus trichocarpa 3694
SOYBN Glycine max 3847
ARATH Arabidopsis thaliana 3702
VITVI Vitis vinifera 29760
SOLLC Solanum lycopersicum 4081
SOLTU Solanum tuberosum 4113
ORYSJ Oryza sativa 39947
BRADI Brachypodium distachyon 15368
SORBI Sorghum bicolor 4558
PHYPA Physcomitrella patens 3218
CHLRE Chlamydomonas reinhardtii 3055
LEINA Leishmania major 5664
LEIBR Leishmania brazilienses 5660
TRYB2 Trypanosoma brucei brucei 999953
TRYCR Trypanosoma cruzi 5693
TRIVA Trichomonas vaginalis 5722
GIAIC Giardia intestinalis 184922
THEKO Thermococcus kodakaraensis
METAC Methanosarcina acetivorans 188937
METJA Methanocaldococcus jannaschii 243232
HALSA Halobacterium salinarum 64091
PYRAE Pyrobaculum aerophilum 178306
SULSO Sulfolobus solfataricus 273057
KORCO Korarchaeum cryptofilum 374847
ECOLI Escherichia coli 83333
PSEAE Pseudomonas aeruginosa 208964
BRADU Bradyrhizobium japonicum 224911
GEOSL Geobacter sulfurreducens 243231
GLOVI Gloeobacter violaceus 251221
STNY3 Synechocystis 1111708
MYCTX Mycobacterium tuberculosis 1773
STRCO Streptomyces coelicolor 100226
AQUAE Aquifex aeolicus 224324
BACTN Bacteroides thetaiotaomicron 226186
BACSU Bacillus subtilis 224308
CHLTR Chlamydia trachomatis 272561
CHLAA Chloroflexus aurantiacus 324602
DEIRA Deinococcus radiodurans 243230
DICTD Dictyoglomus turgidum 515635
FUSNN Fusobacterium nucleatum 190304
LEPIN Leptospira interrogans 189518
RHOBA Rhodopirellula baltica 243090
THEMA Thermotoga maritima 243274
THEYD Thermodesulfovibrio yellowstonii 289376

Figure S4.

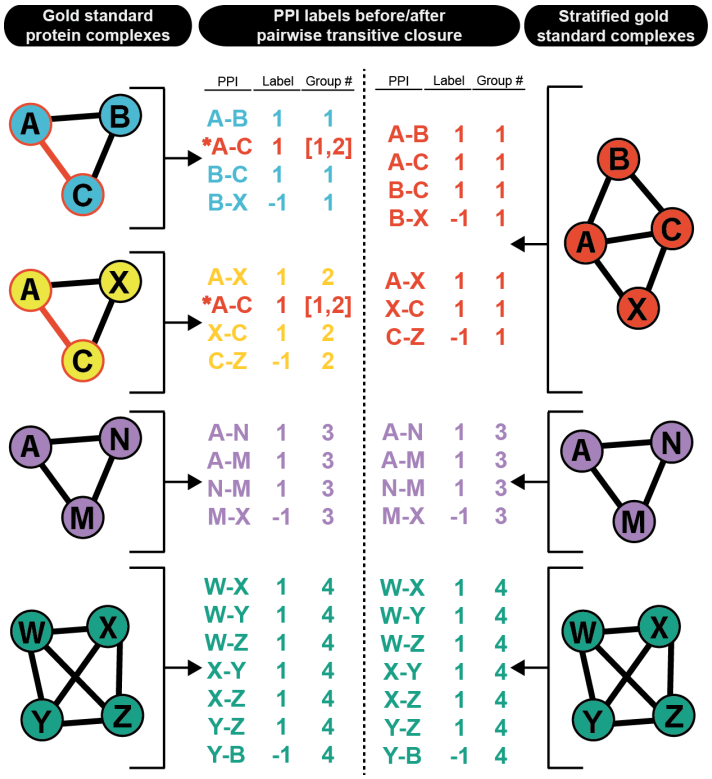

Gold standard protein complexes
PPI labels before/after pairwise transitive closure
Stratified gold standard complexes
PPI Label Group #
A-B 1 1
*A-C 1 [1,2]
B-C 1 1
B-X -1 1
A-X 1 2
*A-C 1 [1,2]
X-C 1 2
C-Z -1 2
A-N 1 3
A-M 1 3
N-M 1 3
M-X -1 3
W-X 1 4
W-Y 1 4
W-Z 1 4
X-Y 1 4
X-Z 1 4
Y-Z 1 4
Y-B -1 4
PPI Label Group #
A-B 1 1
A-C 1 1
B-C 1 1
B-X -1 1
A-X 1 1
X-C 1 1
C-Z -1 1
A-N 1 3
A-M 1 3
N-M 1 3
M-X -1 3
W-X 1 4
W-Y 1 4
W-Z 1 4
X-Y 1 4
X-Z 1 4
Y-Z 1 4
Y-B -1 4

Figure S5.

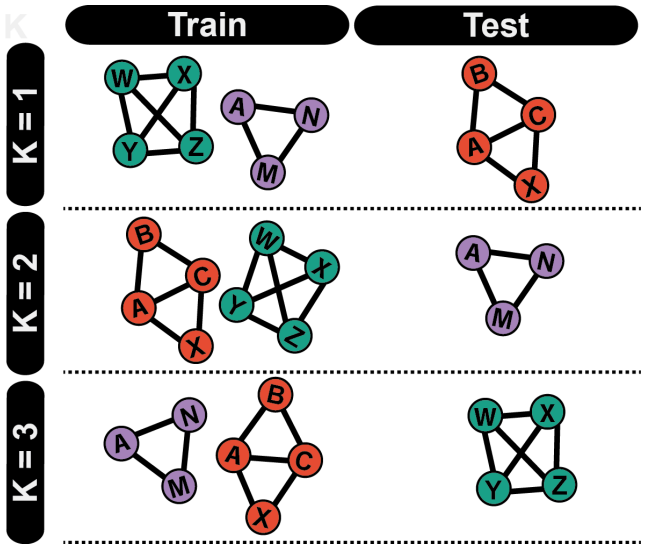
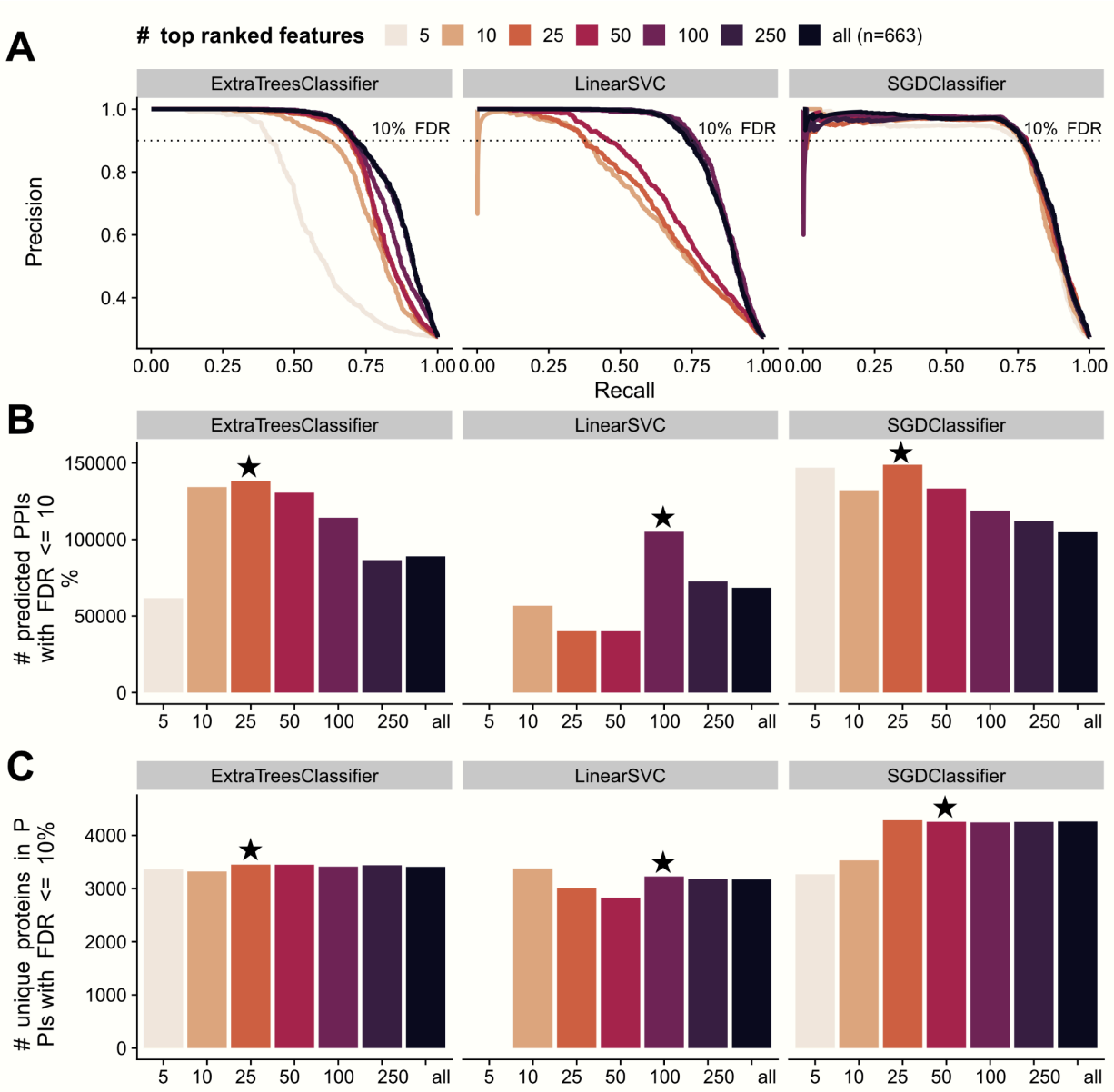


Figure S6.


A
top ranked features
5
10
25
50
100
250
all (n=663)
ExtraTreesClassifier
LinearSVC
SGDClassifier
10% FDR
Precision
Recall
B
predicted PPIs with FDR <= 10 %
C
unique proteins in PPIs with FDR <= 10%

Supplemental Tables

Table S1. 10,092 orthogroups estimated by Dollo parsimony to have been present in LECA. 12MB file, available on Zenodo repository.

Table S2. Summary of biological samples, data sets, software, and algorithms used to derive the conserved eukaryotic interactome.

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
BIOLOGICAL SAMPLES		
Euglena gracilis	UTEX Culture Collection of Algae	UTEX 753
Tetrahymena thermophila	Tetrahymena Stock Center	SB175 (SD01508)
Pig trachea (*Sus scrofa*)	Sierra for Medical Science sierra-medical.com	N/A
Diatom (*Phaeodactylum tricornutum*)	UTEX Culture Collection of Algae	UTEX 646
Rotifer (*Brachionus rotundiformis*)	Eberhart lab, UT Austin	S-type
MASS SPECTROMETRY DATA		
Euglena gracilis	This work	**PRIDE:** PXD050669
Tetrahymena thermophila	This work	**PRIDE:** PXD050671, PXD050672, PXD050674
Pig (*Sus scrofa*)	This work	**PRIDE:** PXD041980
Diatom (*Phaeodactylum tricornutum*)	This work	**PRIDE:** PXD050670
Rotifer (*Brachionus rotundiformis*)	This work	**PRIDE:** PXD050673
Mouse (*Mus musculus*)	This work	**PRIDE:** PXD041915
	Wan et al., 2015	**PRIDE:** PXD002323
Arabidopsis thaliana	McWhite et al., 2020	**PRIDE:** PXD013264, PXD013321, PXD014617
Broccoli (*Brassica oleracea*)	McWhite et al., 2020	**PRIDE:** PXD013281, PXD013322, PXD013282
Chlamydomonas reinhardtii	McWhite et al., 2020	**PRIDE:** PXD013369, PXD013735
Coconut (*Cocos nucifera*)	McWhite et al., 2020	**PRIDE:** PXD012865
Fern (*Ceratopteris richardii*)	McWhite et al., 2020	**PRIDE:** PXD013320
Hemp (*Cannabis sativa*)	McWhite et al., 2020	**PRIDE:** PXD012969
Maize (*Zea mays*)	McWhite et al., 2020	**PRIDE:** PXD012810
Quinoa (*Chenopodium quinoa*)	McWhite et al., 2020	**PRIDE:** PXD013080
Rice (*Oryza sativa*)	McWhite et al., 2020	**PRIDE:** PXD013213
Selaginella moellendorffii	McWhite et al., 2020	**PRIDE:** PXD013093
Soy (*Glycine max*)	McWhite et al., 2020	**PRIDE:** PXD013198, PXD013704
Tomato (*Solanum lycopersicum*)	McWhite et al., 2020	**PRIDE:** PXD013004

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Wheat (*Triticum aestivum*)	McWhite et al., 2020	**PRIDE:** PXD013214, PXD013280, PXD013300
Plasmodium berghei	Hillier et al., 2019	**PRIDE:** PXD009039
Plasmodium falciparum	Hillier et al., 2019	**PRIDE:** PXD009039
Plasmodium knowlesi	Hillier et al., 2019	**PRIDE:** PXD009039
Trypanosoma brucei	Crozier et al., 2017	**PRIDE:** PXD005968
Human (*Homo sapiens*)	Wan et al., 2015	**PRIDE:** PXD002322, PXD002328
Worm (*Caenorhabditis elegans*)	Wan et al., 2015	**PRIDE:** PXD002319
Slime mold (*Dictyostelium discoideum*)	Wan et al., 2015	**PRIDE:** PXD002320
Fly (*Drosophila melanogaster*)	Wan et al., 2015	**PRIDE:** PXD002321
Sea anemone (*Nematostella vectensis*)	Wan et al., 2015	**PRIDE:** PXD002324
Sea urchin (*Strongylocentrotus purpuratus*)	Wan et al., 2015	**PRIDE:** PXD002325
African clawed frog (*Xenopus laevis*)	Drew et al., 2020; Wan et al., 2015	**PRIDE:** PXD017650, PXD017659; PXD002326
Yeast (*Saccharomyces cerevisiae*)	Wan et al., 2015	**PRIDE:** PXD002327
SOFTWARE AND ALGORITHMS		
Orthogroup inference	eggNOG v2.0.5	https://github.com/eggnogdb/eggnog-mapper/releases
Reference database construction	This paper	https://github.com/marcottelab/leca-proteomics/scripts/concat_ortho_proteins.py
Peptide identification	MSblender	https://github.com/marcottelab/MSblender
Ancestral genome inference	Count	http://www.iro.umontreal.ca/~csuros/gene_content/count.html
Feature extraction (CFMS data)	McWhite et al., 2020	https://github.com/marcottelab/protein_complex_maps/tree/master/protein_complex_maps/features/ExtractFeatures/canned_scripts/extract_features.py
Feature integration (APMS data)	hu.MAP 2.0	http://humap2.proteincomplexes.org/static/downloads/humap2/
	This paper	https://github.com/marcottelab/leca-proteomics/notebooks/map_entrez_to_eggnog.ipynb
Pairwise protein interaction labels	This paper	https://github.com/marcottelab/leca-proteomics/scripts/label_featmat.py
Model optimization	TPOT 0.11.7	http://epistasislab.github.io/tpot/
	This paper	https://github.com/marcottelab/leca-proteomics/scripts/run_tpot.py

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Feature/model selection	This paper	https://github.com/marcottelab/leca-proteomics/scripts/select_features.py
Final model generation/assessment	This paper	https://github.com/marcottelab/leca-proteomics/scripts/predict_ppis.py
Protein complex detection	This paper	https://github.com/marcottelab/leca-proteomics/scripts/detect_communities.py

Table S3. 3,193 LECA orthogroups organized hierarchically into 2,013 protein assemblies. 1.3 MB file, available on Zenodo repository.

Table S4. Reference proteomes sourced for co-fractionation mass spectrometry data processing.

Species Code	Species Name	Proteome Source	Date Accessed
ARATH	*Arabidopsis thaliana* (Mouse-ear cress)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000006548	2/5/2021
BRAOL	*Brassica oleracea* (broccoli)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000032141	2/5/2021
BRART	*Brachionus* rotundiformis (rotifer)	Transcriptome downloaded from https://www.ncbi.nlm.nih.gov/Traces/wgs/GINZ01?display=contigs	2/5/2021
CAEEL	*Caenorhabditis elegans*	Proteome downloaded from https://www.uniprot.org/proteomes/UP000001940	2/5/2021
CANSA	*Cannabis sativa* (hemp)	Proteome downloaded from http://genome.ccbr.utoronto.ca/downloads.html	2/5/2021
CERRI	*Ceratopteris richardii* (fern)	Proteome downloaded from https://zenodo.org/record/3467771#.YB2hB-hKguU	2/5/2021
CHEQI	*Chenopodium quinoa* (quinoa)	Proteome downloaded from https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Cquinoa	2/5/2021
CHLRE	*Chlamydomonas reinhardtii* (*Chlamydomonas smithii*)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000006906	2/5/2021
COCNU	*Cocos nucifera* (coconut)	Proteome downloaded from https://www.ncbi.nlm.nih.gov/assembly/GCA_008124465.1/	2/5/2021
DICDI	*Dictyostelium discoideum* (Slime mold)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000002195	2/5/2021
DROME	*Drosophila melanogaster* (Fruit fly)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000000803	2/5/2021
EUGGR	*Euglena gracilis* (algae)	Proteome downloaded from ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2019/01/PXD009998	2/5/2021
HUMAN	*Homo sapiens* (Human)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000005640	2/5/2021
MAIZE	*Zea mays* (Maize)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000007305	2/5/2021
MOUSE	*Mus musculus* (Mouse)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000000589	2/5/2021
NEMVE	*Nematostella vectensis* (Starlet sea anemone)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000001593	2/5/2021
ORYSJ	*Oryza sativa* subsp. *japonica* (Rice)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000059680	2/5/2021

Species Code	Species Name	Proteome Source	Date Accessed
PHATC	*Phaeodactylum tricornutum* (diatom, strain CCAP 1055/1)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000000759	2/5/2021
PIG	*Sus scrofa* (wild boar)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000008227	2/5/2021
PLABA	*Plasmodium berghei* (strain Anka)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000074855	2/8/2021
PLAF7	*Plasmodium falciparum* (isolate 3D7)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000001450	2/5/2021
PLAKH	*Plasmodium knowlesi* (strain H)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000031513	2/8/2021
SELML	*Selaginella moellendorffii* (spikemoss)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000001514	2/5/2021
SOLLC	*Solanum lycopersicum* (tomato)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000004994	2/5/2021
SOYBN	*Glycine max* (soybean)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000008827	2/5/2021
STRPU	*Strongylocentrotus purpuratus* (purple sea urchin)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000007110	2/5/2021
TETTS	*Tetrahymena thermophila* (ciliate, strain SB210)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000009168	2/5/2021
TRYB2	*Trypanosoma brucei brucei* (strain 927/4 GUTat10.1)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000008524	2/22/2021
WHEAT	*Triticum aestivum* (wheat)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000019116	2/5/2021
XENLA	*Xenopus laevis* (African clawed frog)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000186698	2/5/2021
YEAST	*Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast)	Proteome downloaded from https://www.uniprot.org/proteomes/UP000002311	2/5/2021

Table S5. Summary of the top scoring algorithms, parameters and pre-processing steps found by TPOT. Models marked with a star (★) were selected for further evaluation.

Model	Pre-processing Steps	Parameters	TPOT CV Score	True Test Score
★ LinearSVC	RobustScaler(), ZeroCount(), VarianceThreshold (threshold=0.01)	C=0.01, dual=False, loss="squared_hinge", penalty="l1", tol=0.01	0.869	0.878
ExtraTreesClassifier	OneHotEncoder (minimum_fraction= 0.15, sparse=False, threshold=10)	bootstrap=False, criterion="gini", max_features=0.6000000000000001, min_samples_leaf=11, min_samples_split=12, n_estimators=100	0.911	0.848
★ ExtraTreesClassifier	None	bootstrap=True, criterion="gini", max_features=0.6000000000000001, min_samples_leaf=15, min_samples_split=2, n_estimators=100	0.87	0.891
ExtraTreesClassifier	Normalizer (norm="max")	bootstrap=False, criterion="entropy", max_features=0.35000000000000003, min_samples_leaf=15, min_samples_split=3, n_estimators=100	0.867	0.854
★ SGDClassifier	Normalizer (norm="l1"), StandardScaler()	alpha=0.01, eta0=0.01, fit_intercept=False, l1_ratio=1.0, learning_rate="invscaling", loss="modified_huber", penalty="elasticnet", power_t=0.5	0.87	0.897

REFERENCES

[1] Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol 2018;2:1556–62. https://doi.org/10.1038/s41559-018-0644-x.

[2] Brocks JJ, Nettersheim BJ, Adam P, Schaeffer P, Jarrett AJM, Güneli N, et al. Lost world of complex life and the late rise of the eukaryotic crown. Nature 2023;618:767–73. https://doi.org/10.1038/s41586-023-06170-w.

[3] Skejo J, Garg SG, Gould SB, Hendriksen M, Tria FDK, Bremer N, et al. Evidence for a Syncytial Origin of Eukaryotes from Ancestral State Reconstruction. Genome Biol Evol 2021;13:evab096. https://doi.org/10.1093/gbe/evab096.

[4] Bremer N, Tria FDK, Skejo J, Martin WF. The Ancestral Mitotic State: Closed Orthomitosis With Intranuclear Spindles in the Syncytial Last Eukaryotic Common Ancestor. Genome Biol Evol 2023;15:evad016. https://doi.org/10.1093/gbe/evad016.

[5] Tromer EC, van Hooff JJE, Kops GJPL, Snel B. Mosaic origin of the eukaryotic kinetochore. Proc Natl Acad Sci 2019;116:12873–82. https://doi.org/10.1073/pnas.1821945116.

[6] Kontou A, Herman EK, Field MC, Dacks JB, Koumandou VL. Evolution of factors shaping the endoplasmic reticulum. Traffic 2022;23:462–73. https://doi.org/10.1111/tra.12863.

[7] Klute MJ, Melançon P, Dacks JB. Evolution and Diversity of the Golgi. Cold Spring Harb Perspect Biol 2011;3:a007849. https://doi.org/10.1101/cshperspect.a007849.

[8] Field MC, Dacks JB. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. Curr Opin Cell Biol 2009;21:4–13. https://doi.org/10.1016/j.ceb.2008.12.004.

[9] Mans B, Anantharaman V, Aravind L, Koonin EV. Comparative Genomics, Evolution and Origins of the Nuclear Envelope and Nuclear Pore Complex. Cell Cycle 2004;3:1625–50. https://doi.org/10.4161/cc.3.12.1316.

[10] Wickstead B, Gull K. The evolution of the cytoskeleton. J Cell Biol 2011;194:513–25. https://doi.org/10.1083/jcb.201102065.

[11] Richards TA, Cavalier-Smith T. Myosin domain evolution and the primary divergence of eukaryotes. Nature 2005;436:1113–8. https://doi.org/10.1038/nature03949.

[12] Hodges ME, Scheumann N, Wickstead B, Langdale JA, Gull K. Reconstructing the evolutionary history of the centriole from protein components. J Cell Sci 2010;123:1407–13. https://doi.org/10.1242/jcs.064873.

[13] Mitchell DR. Evolution of Cilia. Cold Spring Harb Perspect Biol 2017;9:a028290. https://doi.org/10.1101/cshperspect.a028290.

[14] Dacks JB, Field MC. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. J Cell Sci 2007;120:2977–85. https://doi.org/10.1242/jcs.013250.

[15] Gabaldón T. Evolution of the Peroxisomal Proteome. In: del Río LA, Schrader M, editors. Proteomics Peroxisomes Identifying Nov. Funct. Regul. Netw., Singapore: Springer; 2018, p. 221–33. https://doi.org/10.1007/978-981-13-2233-4_9.

[16] Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, et al. Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. Microbiol Mol Biol Rev 2012;76:444–95. https://doi.org/10.1128/mmbr.05024-11.

[17] Koonin EV. The Incredible Expanding Ancestor of Eukaryotes. Cell 2010;140:606–8. https://doi.org/10.1016/j.cell.2010.02.022.

[18] Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit Rev Biochem Mol Biol 2013;48:373–96. https://doi.org/10.3109/10409238.2013.821444.

[19] Deutekom ES, Snel B, van Dam TJP. Benchmarking orthology methods using

phylogenetic patterns defined at the base of Eukaryotes. Brief Bioinform 2021;22:bbaa206. https://doi.org/10.1093/bib/bbaa206.
[20] O'Malley MA, Leger MM, Wideman JG, Ruiz-Trillo I. Concepts of the last eukaryotic common ancestor. Nat Ecol Evol 2019;3:338–44. https://doi.org/10.1038/s41559-019-0796-3.
[21] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47–52. https://doi.org/10.1038/35011540.
[22] Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature 2001;409:853–5. https://doi.org/10.1038/35057050.
[23] Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, et al. A map of the interactome network of the metazoan C. elegans. Science 2004;303:540–3. https://doi.org/10.1126/science.1091403.
[24] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006;440:637–43. https://doi.org/10.1038/nature04670.
[25] Guruharsha KG, Rual J-F, Zhai B, Mintseris J, Vaidya P, Vaidya N, et al. A protein complex network of Drosophila melanogaster. Cell 2011;147:690–703. https://doi.org/10.1016/j.cell.2011.08.047.
[26] Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, et al. A census of human soluble protein complexes. Cell 2012;150:1068–81. https://doi.org/10.1016/j.cell.2012.08.011.
[27] Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, et al. Panorama of ancient metazoan macromolecular complexes. Nature 2015;525:339–44. https://doi.org/10.1038/nature14877.
[28] Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol 2021;17:e10016. https://doi.org/10.15252/msb.202010016.
[29] Sae-Lee W, McCafferty CL, Verbeke EJ, Havugimana PC, Papoulas O, McWhite CD, et al. The protein organization of a red blood cell. Cell Rep 2022;40:111103. https://doi.org/10.1016/j.celrep.2022.111103.
[30] Bassel GW, Gaudinier A, Brady SM, Hennig L, Rhee SY, De Smet I. Systems Analysis of Plant Functional, Transcriptional, Physical Interaction, and Metabolic Networks. Plant Cell 2012;24:3859–75. https://doi.org/10.1105/tpc.112.100776.
[31] McWhite CD, Papoulas O, Drew K, Cox RM, June V, Dong OX, et al. A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies. Cell 2020;181:460-474.e14. https://doi.org/10.1016/j.cell.2020.02.049.
[32] Pan J, Li L-P, You Z-H, Yu C-Q, Ren Z-H, Guan Y-J. Prediction of Protein–Protein Interactions in Arabidopsis, Maize, and Rice by Combining Deep Neural Network With Discrete Hilbert Transform. Front Genet 2021;12.
[33] Crozier TWM, Tinti M, Larance M, Lamond AI, Ferguson MAJ. Prediction of Protein Complexes in Trypanosoma brucei by Protein Correlation Profiling Mass Spectrometry and Machine Learning. Mol Cell Proteomics 2017;16:2254–67. https://doi.org/10.1074/mcp.O117.068122.
[34] Hillier C, Pardo M, Yu L, Bushell E, Sanderson T, Metcalf T, et al. Landscape of the Plasmodium Interactome Reveals Both Conserved and Species-Specific Functionality. Cell Rep 2019;28:1635-1647.e5. https://doi.org/10.1016/j.celrep.2019.07.019.
[35] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci 2010;107:6544–9. https://doi.org/10.1073/pnas.0910200107.
[36] Farris JS. Phylogenetic Analysis Under Dollo's Law. Syst Biol 1977;26:77–88. https://doi.org/10.1093/sysbio/26.1.77.

[37] Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309–14. https://doi.org/10.1093/nar/gky1085.
[38] Martin W, Koonin EV. Introns and the origin of nucleus–cytosol compartmentalization. Nature 2006;440:41–5. https://doi.org/10.1038/nature04531.
[39] Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, et al. The molecular architecture of the nuclear pore complex. Nature 2007;450:695–701. https://doi.org/10.1038/nature06405.
[40] Gabaldón T, Pittis AA. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. Biochimie 2015;119:262–8. https://doi.org/10.1016/j.biochi.2015.03.021.
[41] Ferrada E, Superti-Furga G. A structure and evolutionary-based classification of solute carriers. iScience 2022;25:105096. https://doi.org/10.1016/j.isci.2022.105096.
[42] Leipe DD, Wolf YI, Koonin EV, Aravind L. Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 2002;317:41–72. https://doi.org/10.1006/jmbi.2001.5378.
[43] Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, et al. Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell 2004;117:541–52. https://doi.org/10.1016/S0092-8674(04)00450-7.
[44] Sigg MA, Menchen T, Lee C, Johnson J, Jungnickel MK, Choksi SP, et al. Evolutionary Proteomics Uncovers Ancient Associations of Cilia with Signaling Pathways. Dev Cell 2017;43:744-762.e11. https://doi.org/10.1016/j.devcel.2017.11.014.
[45] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[46] Kondrashov FA, Koonin EV, Morgunov IG, Finogenova TV, Kondrashova MN. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. Biol Direct 2006;1:31. https://doi.org/10.1186/1745-6150-1-31.
[47] Dunn MF, Ramírez-Trujillo JA, Hernández-Lucas I. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. Microbiology 2009;155:3166–75. https://doi.org/10.1099/mic.0.030858-0.
[48] Salido E, Pey AL, Rodriguez R, Lorenzo V. Primary hyperoxalurias: Disorders of glyoxylate detoxification. Biochim Biophys Acta BBA - Mol Basis Dis 2012;1822:1453–64. https://doi.org/10.1016/j.bbadis.2012.03.004.
[49] Kristensen AR, Gsponer J, Foster LJ. A high-throughput approach for measuring temporal changes in the interactome. Nat Methods 2012;9:907–9. https://doi.org/10.1038/nmeth.2131.
[50] Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 2015;163:712–23. https://doi.org/10.1016/j.cell.2015.09.053.
[51] Boldt K, van Reeuwijk J, Lu Q, Koutroumpas K, Nguyen T-MT, Texier Y, et al. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. Nat Commun 2016;7:11491. https://doi.org/10.1038/ncomms11491.
[52] Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature 2017;545:505–9. https://doi.org/10.1038/nature22366.
[53] Gupta GD, Coyaud É, Gonçalves J, Mojarad BA, Liu Y, Wu Q, et al. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. Cell 2015;163:1484–99. https://doi.org/10.1016/j.cell.2015.10.065.
[54] Youn J-Y, Dunham WH, Hong SJ, Knight JDR, Bashkurov M, Chen GI, et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. Mol Cell 2018;69:517-532.e11. https://doi.org/10.1016/j.molcel.2017.12.020.

[55] Treiber T, Treiber N, Plessmann U, Harlander S, Daiß J-L, Eichner N, et al. A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. Mol Cell 2017;66:270-284.e13. https://doi.org/10.1016/j.molcel.2017.03.014.
[56] Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucleic Acids Res 2019;47:D559–63. https://doi.org/10.1093/nar/gky973.
[57] Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, et al. The complex portal - an encyclopaedia of macromolecular complexes. Nucleic Acids Res 2015;43:D479–84. https://doi.org/10.1093/nar/gku975.
[58] Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horáčková A, Melicher F, et al. Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res 2019;47:D550–8. https://doi.org/10.1093/nar/gky1001.
[59] Pons P, Latapy M. Computing communities in large networks using random walks (long version) 2005. https://doi.org/10.48550/arXiv.physics/0512106.
[60] Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x.
[61] Kim CY, Baek S, Cha J, Yang S, Kim E, Marcotte EM, et al. HumanNet v3: an improved database of human gene networks for disease research. Nucleic Acids Res 2022;50:D632–9. https://doi.org/10.1093/nar/gkab1048.
[62] Bartolec TK, Vázquez-Campos X, Norman A, Luong C, Johnson M, Payne RJ, et al. Cross-linking mass spectrometry discovers, evaluates, and corroborates structures and protein–protein interactions in the human cell. Proc Natl Acad Sci 2023;120:e2219418120. https://doi.org/10.1073/pnas.2219418120.
[63] Koumandou VL, Dacks JB, Coulson RM, Field MC. Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. BMC Evol Biol 2007;7:29. https://doi.org/10.1186/1471-2148-7-29.
[64] Carvalho-Santos Z, Azimzadeh J, Pereira-Leal JoséB, Bettencourt-Dias M. Tracing the origins of centrioles, cilia, and flagella. J Cell Biol 2011;194:165–75. https://doi.org/10.1083/jcb.201011152.
[65] Mani J, Meisinger C, Schneider A. Peeping at TOMs—Diverse Entry Gates to Mitochondria Provide Insights into the Evolution of Eukaryotes. Mol Biol Evol 2016;33:337–51. https://doi.org/10.1093/molbev/msv219.
[66] D'Souza Z, Taher FS, Lupashin VV. Golgi inCOGnito: From vesicle tethering to human disease. Biochim Biophys Acta BBA - Gen Subj 2020;1864:129694. https://doi.org/10.1016/j.bbagen.2020.129694.
[67] Morava É, Guillard M, Lefeber DJ, Wevers RA. Autosomal recessive cutis laxa syndrome revisited. Eur J Hum Genet 2009;17:1099–110. https://doi.org/10.1038/ejhg.2009.22.
[68] Makhoul C, Gosavi P, Gleeson PA. Golgi Dynamics: The Morphology of the Mammalian Golgi Apparatus in Health and Disease. Front Cell Dev Biol 2019;7.
[69] Bonifacino JS, Glick BS. The Mechanisms of Vesicle Budding and Fusion. Cell 2004;116:153–66. https://doi.org/10.1016/S0092-8674(03)01079-1.
[70] Gillingham AK, Munro S. Transport carrier tethering – how vesicles are captured by organelles. Curr Opin Cell Biol 2019;59:140–6. https://doi.org/10.1016/j.ceb.2019.04.010.
[71] Schindler C, Chen Y, Pu J, Guo X, Bonifacino JS. EARP is a multisubunit tethering complex involved in endocytic recycling. Nat Cell Biol 2015;17:639–50. https://doi.org/10.1038/ncb3129.
[72] Gershlick DC, Schindler C, Chen Y, Bonifacino JS. TSSC1 is novel component of the endosomal retrieval machinery. Mol Biol Cell 2016;27:2867–78. https://doi.org/10.1091/mbc.E16-04-0209.

[73] Balderhaar HJ kleine, Ungermann C. CORVET and HOPS tethering complexes – coordinators of endosome and lysosome fusion. J Cell Sci 2013;126:1307–16. https://doi.org/10.1242/jcs.107805.
[74] Kim JJ, Lipatova Z, Segev N. TRAPP Complexes in Secretion and Autophagy. Front Cell Dev Biol 2016;4.
[75] Wendler F, Gillingham AK, Sinka R, Rosa-Ferreira C, Gordon DE, Franch-Marro X, et al. A genome-wide RNA interference screen identifies two novel components of the metazoan secretory pathway. EMBO J 2010;29:304–14. https://doi.org/10.1038/emboj.2009.350.
[76] Scrivens PJ, Noueihed B, Shahrzad N, Hul S, Brunet S, Sacher M. C4orf41 and TTC-15 are mammalian TRAPP components with a role at an early stage in ER-to-Golgi trafficking. Mol Biol Cell 2011;22:2083–93. https://doi.org/10.1091/mbc.e10-11-0873.
[77] Ungar D, Oka T, Vasile E, Krieger M, Hughson FM. Subunit Architecture of the Conserved Oligomeric Golgi Complex. J Biol Chem 2005;280:32729–35. https://doi.org/10.1074/jbc.M504590200.
[78] Miller VJ, Sharma P, Kudlyk TA, Frost L, Rofe AP, Watson IJ, et al. Molecular Insights into Vesicle Tethering at the Golgi by the Conserved Oligomeric Golgi (COG) Complex and the Golgin TATA Element Modulatory Factor (TMF). J Biol Chem 2013;288:4229–40. https://doi.org/10.1074/jbc.M112.426767.
[79] Sethi G, Shanmugam MK, Arfuso F, Kumar AP. Role of RNF20 in cancer development and progression – a comprehensive review. Biosci Rep 2018;38:BSR20171287. https://doi.org/10.1042/BSR20171287.
[80] Fu J, Liao L, Balaji KS, Wei C, Kim J, Peng J. Epigenetic modification and a role for the E3 ligase RNF40 in cancer development and metastasis. Oncogene 2021;40:465–74. https://doi.org/10.1038/s41388-020-01556-w.
[81] Zhang F, Yu X. WAC, a Functional Partner of RNF20/40, Regulates Histone H2B Ubiquitination and Gene Transcription. Mol Cell 2011;41:384–97. https://doi.org/10.1016/j.molcel.2011.01.024.
[82] Totsukawa G, Kaneko Y, Uchiyama K, Toh H, Tamura K, Kondo H. VCIP135 deubiquitinase and its binding protein, WAC, in p97ATPase-mediated membrane fusion. EMBO J 2011;30:3581–93. https://doi.org/10.1038/emboj.2011.260.
[83] Joachim J, Jefferies HBJ, Razi M, Frith D, Snijders AP, Chakravarty P, et al. Activation of ULK Kinase and Autophagy by GABARAP Trafficking from the Centrosome Is Regulated by WAC and GM130. Mol Cell 2015;60:899–913. https://doi.org/10.1016/j.molcel.2015.11.018.
[84] Sagan L. On the origin of mitosing cells. J Theor Biol 1967;14:225-IN6. https://doi.org/10.1016/0022-5193(67)90079-3.
[85] Martin W, Müller M. The hydrogen hypothesis for the first eukaryote. Nature 1998;392:37–41. https://doi.org/10.1038/32096.
[86] Mills DB, Boyle RA, Daines SJ, Sperling EA, Pisani D, Donoghue PCJ, et al. Eukaryogenesis and oxygen in Earth history. Nat Ecol Evol 2022;6:520–32. https://doi.org/10.1038/s41559-022-01733-y.
[87] Yutin N, Wolf MY, Wolf YI, Koonin EV. The origins of phagocytosis and eukaryogenesis. Biol Direct 2009;4:9. https://doi.org/10.1186/1745-6150-4-9.
[88] Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. The Physiology of Phagocytosis in the Context of Mitochondrial Origin. Microbiol Mol Biol Rev 2017;81:10.1128/mmbr.00008-17. https://doi.org/10.1128/mmbr.00008-17.
[89] Bremer N, Tria FDK, Skejo J, Garg SG, Martin WF. Ancestral State Reconstructions Trace Mitochondria But Not Phagocytosis to the Last Eukaryotic Common Ancestor. Genome Biol Evol 2022;14:evac079. https://doi.org/10.1093/gbe/evac079.
[90] Maruyama S, Kim E. A Modern Descendant of Early Green Algal Phagotrophs. Curr Biol

2013;23:1081–4. https://doi.org/10.1016/j.cub.2013.04.063.
[91] Burns JA, Pittis AA, Kim E. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. Nat Ecol Evol 2018;2:697–704. https://doi.org/10.1038/s41559-018-0477-7.
[92] Akıl C, Robinson RC. Genomes of Asgard archaea encode profilins that regulate actin. Nature 2018;562:439–43. https://doi.org/10.1038/s41586-018-0548-6.
[93] Akıl C, Tran LT, Orhant-Prioux M, Baskaran Y, Manser E, Blanchoin L, et al. Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. Proc Natl Acad Sci 2020;117:19904–13. https://doi.org/10.1073/pnas.2009167117.
[94] Rodrigues-Oliveira T, Wollweber F, Ponce-Toledo RI, Xu J, Rittmann SK-MR, Klingl A, et al. Actin cytoskeleton and complex cell architecture in an Asgard archaeon. Nature 2023;613:332–9. https://doi.org/10.1038/s41586-022-05550-y.
[95] Pollard TD. Regulation of Actin Filament Assembly by Arp2/3 Complex and Formins. Annu Rev Biophys Biomol Struct 2007;36:451–77. https://doi.org/10.1146/annurev.biophys.35.040405.101936.
[96] Davidson AJ, Amato C, Thomason PA, Insall RH. WASP family proteins and formins compete in pseudopod- and bleb-based migration. J Cell Biol 2018;217:701–14. https://doi.org/10.1083/jcb.201705160.
[97] Damiano-Guercio J, Kurzawa L, Mueller J, Dimchev G, Schaks M, Nemethova M, et al. Loss of Ena/VASP interferes with lamellipodium architecture, motility and integrin-dependent adhesion. eLife 2020;9:e55351. https://doi.org/10.7554/eLife.55351.
[98] Edwards M, Zwolak A, Schafer DA, Sept D, Dominguez R, Cooper JA. Capping protein regulators fine-tune actin assembly dynamics. Nat Rev Mol Cell Biol 2014;15:677–89. https://doi.org/10.1038/nrm3869.
[99] Gandhi M, Goode BL. Coronin: The Double-Edged Sword of Actin Dynamics. Madame Curie Biosci. Database Internet, Landes Bioscience; 2013.
[100] Bolger-Munro M, Choi K, Cheung F, Liu YT, Dang-Lawson M, Deretic N, et al. The Wdr1-LIMK-Cofilin Axis Controls B Cell Antigen Receptor-Induced Actin Remodeling and Signaling at the Immune Synapse. Front Cell Dev Biol 2021;9.
[101] Pavlov D, Muhlrad A, Cooper J, Wear M, Reisler E. Actin Filament Severing by Cofilin. J Mol Biol 2007;365:1350–8. https://doi.org/10.1016/j.jmb.2006.10.102.
[102] Maniak M, Rauchenberger R, Albrecht R, Murphy J, Gerisch G. Coronin involved in phagocytosis: Dynamics of particle-induced relocalization visualized by a green fluorescent protein tag. Cell 1995;83:915–24. https://doi.org/10.1016/0092-8674(95)90207-4.
[103] Yan M, Collins RF, Grinstein S, Trimble WS. Coronin-1 Function Is Required for Phagosome Formation. Mol Biol Cell 2005;16:3077–87. https://doi.org/10.1091/mbc.e04-11-0989.
[104] Yan M, Di Ciano-Oliveira C, Grinstein S, Trimble WS. Coronin Function Is Required for Chemotaxis and Phagocytosis in Human Neutrophils1. J Immunol 2007;178:5769–78. https://doi.org/10.4049/jimmunol.178.9.5769.
[105] Toshima J, Toshima JY, Martin AC, Drubin DG. Phosphoregulation of Arp2/3-dependent actin assembly during receptor-mediated endocytosis. Nat Cell Biol 2005;7:246–54. https://doi.org/10.1038/ncb1229.
[106] Yu B, Egbejimi A, Dharmat R, Xu P, Zhao Z, Long B, et al. Phagocytosed photoreceptor outer segments activate mTORC1 in the retinal pigment epithelium. Sci Signal 2018;11:eaag3315. https://doi.org/10.1126/scisignal.aag3315.
[107] Dürrwang U, Fujita-Becker S, Erent M, Kull FJ, Tsiavaliaris G, Geeves MA, et al. Dictyostelium myosin-IE is a fast molecular motor involved in phagocytosis. J Cell Sci 2006;119:550–8. https://doi.org/10.1242/jcs.02774.

[108] Barger SR, Reilly NS, Shutova MS, Li Q, Maiuri P, Heddleston JM, et al. Membrane-cytoskeletal crosstalk mediated by myosin-I regulates adhesion turnover during phagocytosis. Nat Commun 2019;10:1249. https://doi.org/10.1038/s41467-019-09104-1.
[109] Fili N, Toseland CP. Unconventional Myosins: How Regulation Meets Function. Int J Mol Sci 2020;21:67. https://doi.org/10.3390/ijms21010067.
[110] Bergmann C, Guay-Woodford LM, Harris PC, Horie S, Peters DJM, Torres VE. Polycystic kidney disease. Nat Rev Dis Primer 2018;4:1–24. https://doi.org/10.1038/s41572-018-0047-y.
[111] Ikeda K, Brown JA, Yagi T, Norrander JM, Hirono M, Eccleston E, et al. Rib72, a Conserved Protein Associated with the Ribbon Compartment of Flagellar A-microtubules and Potentially Involved in the Linkage between Outer Doublet Microtubules. J Biol Chem 2003;278:7725–34. https://doi.org/10.1074/jbc.M210751200.
[112] Stoddard D, Zhao Y, Bayless BA, Gui L, Louka P, Dave D, et al. Tetrahymena RIB72A and RIB72B are microtubule inner proteins in the ciliary doublet microtubules. Mol Biol Cell 2018;29:2566–77. https://doi.org/10.1091/mbc.E18-06-0405.
[113] Swan EJ, Salem RM, Sandholm N, Tarnow L, Rossing P, Lajer M, et al. Genetic risk factors affecting mitochondrial function are associated with kidney disease in people with Type 1 diabetes. Diabet Med 2015;32:1104–9. https://doi.org/10.1111/dme.12763.
[114] Han M, Moon S, Lee S, Kim K, An WJ, Ryu H, et al. Novel Genetic Variants Associated with Chronic Kidney Disease Progression. J Am Soc Nephrol 2023;34:857. https://doi.org/10.1681/ASN.0000000000000066.
[115] Wang PI, Marcotte EM. It's the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics 2010;73:2277–89. https://doi.org/10.1016/j.jprot.2010.07.005.
[116] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015;43:D789–98. https://doi.org/10.1093/nar/gku1205.
[117] Tolar J, Teitelbaum SL, Orchard PJ. Osteopetrosis. N Engl J Med 2004;351:2839–49. https://doi.org/10.1056/NEJMra040952.
[118] Qin A, Cheng TS, Pavlos NJ, Lin Z, Dai KR, Zheng MH. V-ATPases in osteoclasts: Structure, function and potential inhibitors of bone resorption. Int J Biochem Cell Biol 2012;44:1422–35. https://doi.org/10.1016/j.biocel.2012.05.014.
[119] Duan X, Yang S, Zhang L, Yang T. V-ATPases and osteoclasts: ambiguous future of V-ATPases inhibitors in osteoporosis. Theranostics 2018;8:5379–99. https://doi.org/10.7150/thno.28391.
[120] Nishi T, Kawasaki-Nishi S, Forgac M. Expression and function of the mouse V-ATPase d subunit isoforms. J Biol Chem 2003;278:46396–402. https://doi.org/10.1074/jbc.M303924200.
[121] Esmail S, Kartner N, Yao Y, Kim JW, Reithmeier RAF, Manolson MF. Molecular mechanisms of cutis laxa– and distal renal tubular acidosis–causing mutations in V-ATPase a subunits, ATP6V0A2 and ATP6V0A4. J Biol Chem 2018;293:2787–800. https://doi.org/10.1074/jbc.M117.818872.
[122] Yuan Y, Zhang J, Chang Q, Zeng J, Xin F, Wang J, et al. De novo mutation in ATP6V1B2 impairs lysosome acidification and causes dominant deafness-onychodystrophy syndrome. Cell Res 2014;24:1370–3. https://doi.org/10.1038/cr.2014.77.
[123] Kortüm F, Caputo V, Bauer CK, Stella L, Ciolfi A, Alawi M, et al. Mutations in KCNH1 and ATP6V1B2 cause Zimmermann-Laband syndrome. Nat Genet 2015;47:661–7. https://doi.org/10.1038/ng.3282.
[124] Duan X, Liu J, Zheng X, Wang Z, Zhang Y, Hao Y, et al. Deficiency of *ATP6V1H* Causes Bone Loss by Inhibiting Bone Resorption and Bone Formation through the TGF-β1

Pathway. Theranostics 2016;6:2183–95. https://doi.org/10.7150/thno.17140.
[125] Öz OK, Zerwekh JE, Fisher C, Graves K, Nanu L, Millsaps R, et al. Bone Has a Sexually Dimorphic Response to Aromatase Deficiency. J Bone Miner Res 2000;15:507–14. https://doi.org/10.1359/jbmr.2000.15.3.507.
[126] Wells JCK. Sexual dimorphism of body composition. Best Pract Res Clin Endocrinol Metab 2007;21:415–30. https://doi.org/10.1016/j.beem.2007.04.007.
[127] Lai B, Jiang H, Gao Y, Zhou X. Skeletal ciliopathy: pathogenesis and related signaling pathways. Mol Cell Biochem 2023. https://doi.org/10.1007/s11010-023-04765-5.
[128] Steegmaier M, Borges E, Berger J, Schwarz H, Vestweber D. The E-selectin-ligand ESL-1 is located in the Golgi as well as on microvilli on the cell surface. J Cell Sci 1997;110:687–94. https://doi.org/10.1242/jcs.110.6.687.
[129] Yang T, Mendoza-Londono R, Lu H, Tao J, Li K, Keller B, et al. E-selectin ligand–1 regulates growth plate homeostasis in mice by inhibiting the intracellular processing and secretion of mature TGF-β. J Clin Invest 2010;120:2474–85. https://doi.org/10.1172/JCI42150.
[130] Brooks ER, Wallingford JB. Control of vertebrate intraflagellar transport by the planar cell polarity effector Fuz. J Cell Biol 2012;198:37–45. https://doi.org/10.1083/jcb.201204072.
[131] Toriyama M, Lee C, Taylor SP, Duran I, Cohn DH, Bruel A-L, et al. The ciliopathy-associated CPLANE proteins direct basal body recruitment of intraflagellar transport machinery. Nat Genet 2016;48:648–56. https://doi.org/10.1038/ng.3558.
[132] Follit JA, Tuft RA, Fogarty KE, Pazour GJ. The intraflagellar transport protein IFT20 is associated with the Golgi complex and is required for cilia assembly. Mol Biol Cell 2006;17:3781–92. https://doi.org/10.1091/mbc.e06-02-0133.
[133] Quidwai T, Wang J, Hall EA, Petriman NA, Leng W, Kiesel P, et al. A WDR35-dependent coat protein complex transports ciliary membrane cargo vesicles to cilia. eLife 2021;10:e69786. https://doi.org/10.7554/eLife.69786.
[134] Boeckmann B, Marcet-Houben M, Rees JA, Forslund K, Huerta-Cepas J, Muffato M, et al. Quest for Orthologs Entails Quest for Tree of Life: In Search of the Gene Stream. Genome Biol Evol 2015;7:1988–99. https://doi.org/10.1093/gbe/evv121.
[135] Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernández-Plaza A, et al. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Res 2020;48:W538–45. https://doi.org/10.1093/nar/gkaa308.
[136] Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293.
[137] Csűös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 2010;26:1910–2. https://doi.org/10.1093/bioinformatics/btq315.
[138] Boeckmann B, Dylus D, Moretti S, Altenhoff A, Train C-M, Kriventseva E, et al. Taxon sampling unequally affects individual nodes in a phylogenetic tree: consequences for model gene tree construction in SwissTree 2017:181966. https://doi.org/10.1101/181966.
[139] Altenhoff AM, Dessimoz C. Inferring orthology and paralogy. Methods Mol Biol Clifton NJ 2012;855:259–79. https://doi.org/10.1007/978-1-61779-582-4_9.
[140] Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, et al. Standardized benchmarking in the quest for orthologs. Nat Methods 2016;13:425–30. https://doi.org/10.1038/nmeth.3830.
[141] Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, et al. The Quest for Orthologs orthology benchmark service in 2022. Nucleic Acids Res 2022;50:W623–32. https://doi.org/10.1093/nar/gkac330.
[142] Daugherty MD, Malik HS. Rules of Engagement: Molecular Insights from Host-Virus Arms Races. Annu Rev Genet 2012;46:677–700.

https://doi.org/10.1146/annurev-genet-110711-155522.
[143] Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 2006;440:242–5. https://doi.org/10.1038/nature04559.
[144] de Mendoza A, Sebé-Pedrós A. Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev 2019;58–59:25–32. https://doi.org/10.1016/j.gde.2019.07.010.
[145] Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, et al. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLOS Biol 2012;10:e1001241. https://doi.org/10.1371/journal.pbio.1001241.
[146] Luo A, Ho SYW. The molecular clock and evolutionary timescales. Biochem Soc Trans 2018;46:1183–90. https://doi.org/10.1042/BST20180186.
[147] Gould SJ. Dollo on Dollo's law: Irreversibility and the status of evolutionary laws. J Hist Biol 1970;3:189–212. https://doi.org/10.1007/BF00137351.
[148] Rogozin IB, Wolf YI, Babenko VN, Koonin EV. Dollo parsimony and the reconstruction of genome evolution. In: Albert VA, editor. Parsimony Phylogeny Genomics, Oxford University Press; 2006, p. 0. https://doi.org/10.1093/acprof:oso/9780199297306.003.0011.
[149] Goldberg EE, Igić B. ON PHYLOGENETIC TESTS OF IRREVERSIBLE EVOLUTION. Evolution 2008;62:2727–41. https://doi.org/10.1111/j.1558-5646.2008.00505.x.
[150] Etten JV, Bhattacharya D. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends Genet 2020;36:915–25. https://doi.org/10.1016/j.tig.2020.08.006.
[151] Martin W, Borst P. Secondary loss of chloroplasts in trypanosomes. Proc Natl Acad Sci 2003;100:765–7. https://doi.org/10.1073/pnas.0437776100.
[152] Hannaert V, Saavedra E, Duffieux F, Szikora J-P, Rigden DJ, Michels PAM, et al. Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci U S A 2003;100:1067–71. https://doi.org/10.1073/pnas.0335769100.
[153] Maruyama S, Matsuzaki M, Misawa K, Nozaki H. Cyanobacterial contribution to the genomes of the plastid-lacking protists. BMC Evol Biol 2009;9:197. https://doi.org/10.1186/1471-2148-9-197.
[154] Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, Miyagishima S, et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature 2004;428:653–7. https://doi.org/10.1038/nature02398.
[155] Reyes-Prieto A, Bhattacharya D. Phylogeny of Calvin cycle enzymes supports Plantae monophyly. Mol Phylogenet Evol 2007;45:384–91. https://doi.org/10.1016/j.ympev.2007.02.026.
[156] Bhattacharya D, Yoon HS, Hackett JD. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. BioEssays 2004;26:50–60. https://doi.org/10.1002/bies.10376.
[157] Burki F, Roger AJ, Brown MW, Simpson AGB. The New Tree of Eukaryotes. Trends Ecol Evol 2020;35:43–55. https://doi.org/10.1016/j.tree.2019.08.008.
[158] Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, et al. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups." Proc Natl Acad Sci 2009;106:3859–64. https://doi.org/10.1073/pnas.0807880106.
[159] Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, et al. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. Syst Biol 2010;59:518–33. https://doi.org/10.1093/sysbio/syq037.
[160] Raychowdhury MK, McLaughlin M, Ramos AJ, Montalbetti N, Bouley R, Ausiello DA, et al. Characterization of Single Channel Currents from Primary Cilia of Renal Epithelial Cells. J Biol Chem 2005;280:34718–22. https://doi.org/10.1074/jbc.M507793200.
[161] McCafferty CL, Papoulas O, Jordan MA, Hoogerbrugge G, Nichols C, Pigino G, et al.

Integrative modeling reveals the molecular architecture of the intraflagellar transport A (IFT-A) complex. eLife 2022;11:e81977. https://doi.org/10.7554/eLife.81977.
[162] Dentler WL. Chapter 3 Isolation of Cilia from Tetrahymena thermophila. Methods Cell Biol., vol. 47, Elsevier; 1995, p. 13–5. https://doi.org/10.1016/S0091-679X(08)60784-0.
[163] Mallam AL, Marcotte EM. Systems-wide Studies Uncover Commander, a Multiprotein Complex Essential to Human Development. Cell Syst 2017;4:483–94. https://doi.org/10.1016/j.cels.2017.04.006.
[164] Drew K, Lee C, Cox RM, Dang V, Devitt CC, McWhite CD, et al. A systematic, label-free method for identifying RNA-associated proteins in vivo provides insights into vertebrate ciliary beating machinery. Dev Biol 2020;467:108–17. https://doi.org/10.1016/j.ydbio.2020.08.008.
[165] McWhite CD, Papoulas O, Drew K, Dang V, Leggere JC, Sae-Lee W, et al. Co-fractionation/mass spectrometry to identify protein complexes. STAR Protoc 2021;2:100370. https://doi.org/10.1016/j.xpro.2021.100370.
[166] Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. J Proteome Res 2011;10:2949–58. https://doi.org/10.1021/pr2002116.
[167] Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. Proc. Genet. Evol. Comput. Conf. 2016, Denver Colorado USA: ACM; 2016, p. 485–92. https://doi.org/10.1145/2908812.2908918.
[168] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;69:026113. https://doi.org/10.1103/PhysRevE.69.026113.
[169] Groza T, Gomez FL, Mashhadi HH, Muñoz-Fuentes V, Gunes O, Wilson R, et al. The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. Nucleic Acids Res 2023;51:D1038–45. https://doi.org/10.1093/nar/gkac972.
[170] Birling M-C, Yoshiki A, Adams DJ, Ayabe S, Beaudet AL, Bottomley J, et al. A resource of targeted mutant mouse lines for 5,061 genes. Nat Genet 2021;53:416–9. https://doi.org/10.1038/s41588-021-00825-y.
[171] Kurbatova N, Mason JC, Morgan H, Meehan TF, Karp NA. PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data. PLOS ONE 2015;10:e0131274. https://doi.org/10.1371/journal.pone.0131274.
[172] Park TJ, Gray RS, Sato A, Habas R, Wallingford JB. Subcellular Localization and Signaling Properties of Dishevelled in Developing Vertebrate Embryos. Curr Biol 2005;15:1039–44. https://doi.org/10.1016/j.cub.2005.04.062.
[173] Huizar RL, Lee C, Boulgakov AA, Horani A, Tu F, Marcotte EM, et al. A liquid-like organelle at the root of motile ciliopathy. eLife 2018;7:e38497. https://doi.org/10.7554/eLife.38497.
[174] Lee C, Cox RM, Papoulas O, Horani A, Drew K, Devitt CC, et al. Functional partitioning of a liquid-like organelle during assembly of axonemal dyneins. eLife 2020;9:e58662. https://doi.org/10.7554/eLife.58662.
[175] Fisher M, James-Zorn C, Ponferrada V, Bell AJ, Sundararaj N, Segerdell E, et al. Xenbase: key features and resources of the Xenopus model organism knowledgebase. Genetics 2023;224:iyad018. https://doi.org/10.1093/genetics/iyad018.