The prokaryotic origins of the COMMD protein family involved in eukaryotic membrane trafficking

Meihan Liu[1], Edmund R.R Moody[2,3] , Farrah Blades[4], Caroline Puente-Lelievre[5], Kai-En Chen[1], Ella J. Stephens[1], Katharine A. Michie[6], Rosemary J. Cater[4], Tom A. Williams[3,8], Brett M. Collins[1*], Michael D. Healy[1*]

[1]Centre for Cell Biology of Chronic Disease, Institute for Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Australia.

[2]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain

[3]Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[4]Centre for Chemistry and Drug Discovery, Institute for Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Australia.

[5] School of Biological Sciences, University of Auckland, Auckland 1010, New Zealand.

[6] School of Biotechnology & Biomolecular Sciences and Mark Wainwright analytical Centre Structural Biology Facility, The University of New South Wales, Sydney, NSW 2052 Australia.

[7] Department of Life Sciences, University of Bath, Bath BA2 7AX, UK.

***Co-corresponding authors**: Brett Collins b.collins@imb.uq.edu.au; Michael Healy michael.healy@uq.edu.au

Running title: Prokaryotic origins of the COMMD protein family.

Keywords: Commander, endosome, COMMD, Type IV secretion system.

ABSTRACT

The ten eukaryotic COMMD proteins are core components of the Commander complex, with central roles in endosomal membrane trafficking and signalling. Each protein has an α-helical N-terminal (HN) domain with a C-terminal copper metabolism gene MURR1 (COMM) domain. These ten family members assemble into a heterodecameric ring composed of five specific heterodimers. In this work we have combined structural homology searches with genome-wide predicted structures to identify ancestral COMMD-like proteins that exist as single genes in Bacteria and Archaea. Although there is limited sequence similarity to the eukaryotic proteins the bacterial and archaeal COMMD-like proteins are predicted to form homomeric ring-shaped assemblies like their eukaryotic counterparts. Our biophysical studies, crystal and cryo-EM structures confirm COMMD-like proteins readily form homooligomeric rings composed of eight or ten subunits assembled from core dimeric building blocks and inter-dimer interactions that are analogous to the heterodecameric core structure of the eukaryotic Commander complex. Phylogenetic analyses using amino acid sequences and FoldSeek structural alphabet (3Di) infer that the closest identified relatives to the eukaryotic COMMD proteins are found in Myxococcota bacteria. These findings indicate that COMMD genes emerged early in eukaryotic evolution through multiple rounds of duplication from a single ancestral gene likely acquired from bacteria.

INTRODUCTION

The trafficking of transmembrane proteins and lipids through the endolysosomal system is critical to all eukaryotic life. The proteins and protein complexes that control this transport are generally conserved across the eukaryote domain from single celled protists to multicellular plants and animals[1]. Eukaryotes evolved from archaeal ancestors that formed a symbiotic partnership with a proto-mitochondrial bacterium[2,3]. The discovery of the Asgard archaea prompted further refinement of this model, with the Asgard lineage possessing numerous eukaryotic signature genes (ESGs)[4,5], including many that encode proteins related to eukaryotic membrane trafficking and cytoskeletal organisation; such as the ubiquitin-ESCRT machinery[6], RAB and ARF-like small GTPases[7,8], SNAREs involved in membrane fusion[9], as well as microtubules, actin and profilins[10-12].

As eukaryotic endosomes mature to lysosomes, several large protein complexes direct the recycling of transmembrane proteins to other compartments such as the Golgi or plasma membrane to avoid degradation[13-15]. One such complex is Commander, made up of two sub-assemblies called the COMMD/CCDC22/CCDC93 (CCC) complex and the Retriever complex, the latter of which is structurally related to the endosomal Retromer machinery[16-18]. Mutations in multiple Commander subunits are causative of X-linked intellectual disability (XLID) and Ritscher-Schinzel syndrome (RSS), a multi-system developmental disorder characterised by abnormal craniofacial features, cerebellar hypoplasia, and stunted cardiovascular development[19]. Commander is composed of sixteen subunits divided into two subassemblies: the Retriever subassembly, which is composed of VPS29, VPS35L and VPS26C, and the CCC sub-assembly, which is comprised of ten copper metabolism gene MuRR1 domain (COMMD) proteins, two coil-coiled domain containing (CCDC) proteins CCDC22 and CCDC93 and a differentially expressed in neoplastic versus normal cells (DENN) domain protein DENND10. Together these sixteen components form a stably assembled complex that is localised to endosomal membranes, regulates transmembrane cargo transport.

At the heart of the Commander complex is a heterodecameric structure containing a single copy of each of the ten COMMD proteins COMMD1 to COMMD10[16-18]. A notable feature of this complex is that the ten proteins form a closed pseudo-symmetric ring, which is itself composed of five specific heterodimers. This ring is formed due to the unique structural features of the COMMD proteins which are composed of two domains; a helical N-terminal domain (HN domain) with a

high degree of sequence divergence, and a more highly conserved C-terminal COMM domain. This COMM domain contains three β-strands followed by an α-helix, the unique arrangement of these elements leads to the formation of a hydrophobic core that results in the formation of an obligate dimer[20]. The decameric ring is then assembled from five of these dimers, where the COMM domains form a closed ring via β-sheet interactions between adjacent dimers. The N-terminal HN domains then extend peripherally on each side of this central core.

A number of studies have used amino acid sequence-based approaches to study the evolution of the Commander protein machinery[18,21-25]. All sixteen subunits are broadly conserved across the eukaryotic domain, including animals and their unicellular relatives such as choanoflagellates and amoebozoans. The complex has also been lost secondarily in some lineages, including in fungi[24] (such as yeast) and flowering plants. A conservation pattern strongly suggesting that Commander subunits have co-evolved alongside the endosomal actin-nucleating Wiskott-Aldrich syndrome protein and SCAR homolog (WASH) complex[26]. Previous analyses have shown that all ten COMMD genes were likely present in the last eukaryotic common ancestor (LECA)[18], with the series of gene duplications giving rise to each subfamily likely occurring along the eukaryotic stem (that is, prior to LECA). While search methods based on sequence similarity (such as BLAST and HMMer) did not find homologues of COMMD in prokaryotes[18,24,26], a recent preprint used structure-based methods to identified COMMD-like proteins in the Asgard archaea, the closest prokaryotic relatives of eukaryotes[27].

Here we use sequence and structure-based homology searches, evolutionary analyses, and biophysical assays to identify the origins and investigate the structure and function of prokaryotic COMMD-like proteins. We find that while CCDC22, CCDC93, DENND10, VPS26C and VPS35L are restricted to the eukaryotic lineage, ancestral COMMD proteins are present in both Bacteria and Archaea. These prokaryotic COMMD-like proteins have very low sequence identity (<20%) to any eukaryotic COMMD family members but share a conserved and highly similar fold including both HN and COMM domains. Phylogenetic analyses further indicate a close evolutionary relationship between eukaryotic COMMD proteins and a clade of bacterial sequences largely comprising representatives from the Myxococcota, bacteria that also appear to have contributed some other membrane associated genes to the nascent eukaryotic lineage[28]. We purified a selected set of bacterial and archaeal COMMD-like proteins and used mass photometry, X-ray crystallography and cryo-electron microscopy (cryo-EM) to demonstrate that most of these

assemble into analogous dimeric structures to the human COMMD proteins, that then form oligomeric rings consisting of eight to ten subunits, which are homo-oligomeric rather than hetero-oligomeric as in eukaryotes. Our results indicate that the COMMD family of proteins evolved from a single prokaryotic ancestral protein-coding gene sharing key structural features, consistent with the early presence of endomembrane components during eukaryogenesis.

RESULTS

Ancestral COMMD-like proteins are found in Bacteria and Archaea

Previous work by us and others identified clear orthologues of each of the sixteen components of Commander in many eukaryotes, including choanoflagellates such as *Salpingoeca rosetta*[18,21,24,26]. To explore whether more distant structural homologues could be found in prokaryotes we used FoldSeek[29] which allows for fast and accurate matching of structurally similar proteins, including structures of proteins encoded by bacteria and archaea available in the AlphaFold2 database[30-32]. These searches were performed initially using structural templates of Commander subunits from a variety of eukaryotic species, and then iteratively searching for additional prokaryotic proteins using identified hits (**Fig. 1A**). Using this approach, we confirmed the presence of clear orthologues of all sixteen subunits throughout the eukaryotic lineage and identified several previously unidentified homologues in *Reticulomyxa filosa* and *Anaeramoeba flamelloide* – single-celled organisms that are only distantly related to model eukaryotes (**Fig. 1B**). Turning to prokaryotes we could not identify any archaeal or bacterial proteins with significant structural similarity to CCDC22, CCDC93, VPS26C or VPS35L. Proteins with folds related to VPS29 (the shared subunit between Retriever and Retromer complexes) and DENND10 were identified in Archaea and Bacteria. Due to the abundant distribution of the metallophosphatase and DENN domain folds across all the tree of life we have not pursued these further. Using FoldSeek we were able to identify clear structural homologues of the COMMD proteins (**Fig. 1B and C**) that we have termed COMMD-like proteins. They have less than 20% sequence identity to any eukaryotic COMMD family members, explaining why they were not readily detected using sequence-based methods.

From initial searches using the ten human COMMD proteins against the UniRef100 AlphaFold2 database we identified 8,238 eukaryotic proteins with a COMMD-related protein fold. These were then used to screen the predicted proteins of all available prokaryotic genomes, with identified homologues used in subsequent iterative searches. In total we identified 250 bacterial COMMD-like proteins and 111 archaeal COMMD-like proteins broadly distributed across the diversity of prokaryotes (**Supplementary Dataset S1**). From Bacteria this included sequences from both Gracilicutes and Terrabacteria, including Proteobacteria, Chlamydia, Cyanobacteria and Actinobacteria amongst others. In the case of Archaea, we find examples across the evolutionary tree including in Methanobacteriota, Halobacteriales, Thermococcus and Asgard species. As

discussed below, most prokaryotic genomes encode at most a single COMMD-like gene, with the exception of *Candidatus Sulfopaludibacter sp. SbA4* and a few species from the *Myxococcus* clade that encode two. A comparison of human COMMD1 and examples of COMMD-like proteins is shown **Fig. 1C**. The ten examples shown from Bacteria and Archaea are all predicted to possess N-terminal HN domains that shares the same α-helical topology as their eukaryotic homologues, and a C-terminal COMM domain comprised of a three-stranded anti-parallel β-sheet and extended α-helix.

Structures and oligomeric properties of the prokaryotic COMMD-like proteins

We next characterised the structure of the ten exemplar COMMD-like proteins shown in **Fig. 1C** using both computational modelling and experimental approaches. A defining feature of the COMMD proteins in eukaryotes is that they form oligomeric complexes. The structure of the COMM domain requires the formation of an obligate dimer to bury their hydrophobic β-sheets[20], the five specific heterodimers then assemble into a closed ring structure of ten subunits[16-18]. Although the eukaryotic COMMD proteins form heteromeric complexes *in vivo*, they are able to form homomeric assemblies when recombinantly expressed individually[20]. We first tested the propensity of the prokaryotic COMMD-like proteins to form homodimers using AlphaFold2 and found that they were universally predicted to form homodimers with high confidence metrics (**Figs. S1-S3**). We then modelled each of the ten COMMD-like proteins as increasingly larger oligomers (again using AlphaFold2), focusing on the octameric and decameric assemblies. Apart from A4J5H3 and A0A558G9V7 (for the sake of precise identification proteins have been labelled by either a Uniprot ID or Uniparc ID if Uniprot ID was unavailable) all proteins are predicted, with high confidence, to form closed ring structures using either eight or ten subunits (**Figs. S1-S3**). The predicted ring structures are assembled via similar interfaces between dimers and are structurally analogous to the known heterodecameric structures of the human COMMD1-10 complex[16-18]. A0A558G9V7 is predicted to form a decameric complex, however the geometry of its COMM domain does not allow for the octameric ring to close. A4J5H3 is not predicted to form a complex however it is worth noting that less than 10 sequences were identified which likely limits the ability of AlphaFold2 to make accurate predictions. These results suggest that the prokaryotic COMMD-like proteins may assemble into structures like their eukaryotic homologues.

We next aimed to experimentally validate the structures of these proteins and their oligomeric properties. Each of the ten proteins were expressed in *E. coli* and purified using His-tag affinity chromatography. By size exclusion chromatography all proteins have relatively homogenous elution peaks and Coomassie stained SDS-PAGE gels show that these proteins can be isolated with purity sufficient to obtain macromolecular crystals (**Fig. 2A** and **Fig. S4**).

The retention volumes of the COMMD-like proteins by gel filtration are larger than expected for monomeric species suggesting they exist in homo-oligomeric complexes (**Fig. 2B**). To more precisely characterise these, we used multiangle laser light scattering (MALLS). Based on AlphaFold predictions and what is known about eukaryotic COMM domains, it is expected that all proteins form complexes that are composed of a dimeric building block. As indicated by MALLS, seven COMMD-like proteins form large complexes that consist of at least eight protomers based on the expected monomeric molecular weight (**Fig 3A; Figs. S5A and C**). The rapid decrease in molecular weight across the MALLS peaks for most proteins suggests that they are not fully monodisperse. To explore this further we used an orthogonal technique called mass photometry, which can provide an estimate of the heterogeneity within a sample and the molecular mass of each species (**Fig 3B-3D; Fig. S5B**). As an example, MBK5214510 shows evidence for the presence of dimeric, tetrameric, hexameric, octameric and decameric species. The formation of the oligomeric structure may be dependent on protein concentration (note MALLs in performed at ~100x higher concentrations). A similar distribution of different oligomeric states can also be seen with A0A2U3LUU8, A0A2U3L5I1 and A4J5H3 (**Fig. S5B**). S4XVP2, S4Y171 A0A7C1Q7X0, UYP43922 and A0A558G9V7 form stable octamers, whereas A0A6A7KI92, seemed to form relatively stable decameric assembly. (**Fig 3C; Fig. S5B**). Interestingly, the distribution of oligomeric states is only seen in some of the bacterial homologues while archaeal homologues appear to be less dependent on concentration and tend towards existing in more stable oligomeric states as a result. As an exception the *myxococcota* (bacterial) homologues S4XVP2 and S4Y171 form stable homogenous octameric complexes in both MALLS and Mass photometry.

One outstanding question is why we observe this variation in oligomeric assemblies in bacterial COMMD-like proteins. Broadly speaking there are two likely reasons: (i) the concentration at which the mass photometry measurements was taken may be too low for the larger oligomers to remain stably assembled, and/or (ii) additional components could be required to stabilise the assemblies as seen with the human COMMD proteins within the Commander

complex[18]. To examine the stability of the COMMD-like protein oligomers, nano-DSF was used to determine the T_m^{50} (the temperature at which 50% of the protein was unfolded). Unfortunately, several of the proteins had low, or no tryptophan residues required for fluorescence measurement which limited the assessment to five out of the ten. Of these, tested three were highly stable, with melting temperatures of ~60ºC and two were less stable having melting temperatures of ~40ºC. This did not appear to correlate with the oligomeric state of the COMMD-like protein or prokaryotic lineage (**Fig. S5D**).

Prokaryotic COMMD-like proteins form analogous complexes to the COMMD family

We next used Xray crystallography and cryo-EM structure determination to experimentally confirm our predicted structures of these COMMD-like proteins. Crystallisation screens of all ten COMMD-like proteins resulted in successful crystal growth, however most showed poor X-ray diffraction apart from crystals of MBK5214510 from *Flavobacteriaceae bacterium* (**Fig. 4A; Table S1; Supplementary Dataset S2**). We were able to solve the crystal structure at 2.6 Å resolution by molecular replacement using the AlphaFold2 model as a template. The asymmetric unit contains a dimer of MBK5214510 that resembles the human homo and heterodimeric structures solved previously[16,18,20] (**Fig. 4A**). As expected, the dimer is formed by the intertwining of the two C-terminal COMM domains, each comprised of three β-strands and one α-helix that creates a large hydrophobic surface buried within the dimer interface. The HN domain contains six α-helices that form a tightly packed globular domain, the orientation with respect to the core COMM domain dimer results in the HN domains been arranged opposite and at 180° from each other (**Fig. 4A**). Inspection of the crystallographic lattice shows that the MBK5214510 dimer is forming an octameric ring structure along the 4-fold symmetry axis that is the same as the AlphaFold2 predicted octamer (**Fig. 4A**).

In parallel to the crystallographic studies, we also determined the structure of MBK5214510 by cryo-EM and single particle analysis (**Fig. 4B-4C; Fig. S6C-S6D; S7; Table S2**). Within the 2D classes there were clear examples of top views of both octameric and decameric ring structures. Filtering of these two different classes allowed us to resolve the octameric and decameric structures at ~4.2 Å and ~4.8 Å resolution respectively (**Fig 4B-4C; Fig. S7**). The octameric particles resembled the octameric ring we observed in the crystal structure, while the decameric particles were assembled in the same way as the octameric rings, but were composed

of five identical homodimers rather than four. As mentioned above, AlphaFold2 predictions were able to produce confident models of closed ring structures using either 8 or 10 monomers (predictions using more or fewer subunits produce either fragments of rings (e.g A0A558G9V7) or physically unreasonable structures) (**Figs. S1-S3**). The modelling, biophysical data (**Fig. 3; Fig. S5**) and cryo-EM data (**Fig. 4B-4C**) thus supports the idea that the bacterial COMMD-like proteins have a capacity to form closed ring structures with some plasticity as to the number of dimeric building blocks that are incorporated. Comparisons of the experimental MBK5214510 oligomers to the AlphaFold2 predictions shows that the octameric crystal structure has the closest structural alignment with the predicted octamer (**Fig. S9**). The octamer determined by cryo-EM is highly similar to the prediction but shows greater flexibility than in the crystal structure. The decamer determined by cryo-EM is more divergent from the predicted decamer, with a significantly larger central pore, and is also wider than the cryo-EM structure of the human COMMD1-10 heterodecameric complex[17,18] (**Fig. S9-S10**). Overall, even though there is only an average sequence conservation of less than 13% these results show that bacterial COMMD-like proteins are able to form homomeric rings that are analogous to the eukaryotic heterodecameric complexes, formed by a central ring of interlocked COMM domains that tessellate together and HN domains that extend peripherally.

To confirm that the octameric structure was conserved across species we also investigated those from the *myxococcota - sorangium cellulosum so0157 2*. This species contained two unique sequences (S4Y171 and S4XVP2) unlike many other species we identified. Using Cryo-electron microscopy we confirmed that these proteins also form octameric structures (**Fig. 4D-4E; Fig. S6E-S6F; Fig. S8; Table S2**) with a remarkably conserved architecture to both MBK5214510 and the human heterodecameric COMMD ring (**Fig. 4F**).

In the case of A0A6A7KI92 a low resolution cryo-EM dataset was collected that shows 2D classes consistent with an ability to form a homodecameric ring. Orientation bias prevented us resolving a high resolution structure (**Fig. S6**).

Eukaryotic COMMD proteins are most closely related to COMMD-like proteins from the Myxococcota lineage.

The evidence presented here suggests that the eukaryotic COMMD proteins originated via duplication and expansion of a single ancestral protein-coding gene. To assess the phylogenetic

relationships between the Eukaryotic COMMD and the prokaryotic COMMD-like proteins, we performed independent phylogenetic analyses using amino acid sequences and 3Di structure data (**Figs. 5 and S11A**). These analyses did not support a close relationship between eukaryotic COMMDs and the COMMD-like proteins of Asgard archaea (the closest relatives of the eukaryotic nuclear lineage) or of the alphaproteobacterial relatives of the mitochondrion (**Fig. 5**), indicating that eukaryotes did not inherit COMMD from either of their immediate prokaryotic ancestors. However, robustly establishing the closest prokaryotic clade – that is, the extant relatives of the prokaryotic donor lineage - proved challenging because of the low level of similarity between sequences and substantial phylogenetic uncertainty in both sequence- and structure-based analyses, which were performed under the best-fit site-heterogeneous model (sequence analysis), and using a recently developed substitution model for 3Di characters (structure analysis[33,34]). Intriguingly, in the maximum likelihood sequence-based tree, eukaryotes branch with a clade predominantly comprising Myxococcota, but also including individual sequences from Desulfobacterota, Gemmatimonadota and Planctomycetota (bootstrap support 98%). This is intriguing because several other genes involved in eukaryotic membrane biology, including genes for sterol biosynthesis, appear to have been acquired from this clade[35]. A close and highly-supported relationship between eukaryotic COMMD and Myxococcota was also observed in the structure-based tree (**Fig. S11A)**.

Myxococcota *COMMD-like genes cluster with secretion system components*

Although the prokaryotic COMMD-like proteins clearly share a conserved structure with their eukaryotic relatives, their functions in either Bacteria or Archaea are unknown. In prokaryotes, genes that function in related pathways often tend to cluster together, for example components of hydrogenase complexes[36] and the ancestral ESCRT and ubiquitination machinery in Archaea[37-39]. We therefore investigated the genomic context of the COMMD-like genes with a particular emphasis on the *Myxococcota* bacteria. An example genomic region focused on S4XVP2 from *Sorangium cellulosum* So0157-2 is shown in **Fig. 6A**. Interestingly, by investigating the adjacent genes, we noticed similarities between this region and the VgrG1B region of *Pseudomonas aeruginosa* PA01[40] (**Fig. 6A**), although we note *P. aeruginosa* PA01 does not have an identified COMMD-like protein in the genome. The VgrG1B region is known to be critical for *P. aeruginosa*

PA01 virulence and contains numerous components related to the Type VI secretions system (T6SS) (**Fig. 6B**).

To understand if this was an isolated case or common to other COMMD-like proteins from *Myxcoccotta* we next performed syntenic analyses. Several COMMD-like homologues from *Myxococcota* species are available in the NCBI database that have not yet been catalogued in Uniprot or the AlphaFold2 database. In total, 55 full-length COMMD-like sequences were identified (compared to 32 from Uniprot and 92 in NCBI with shorter reads) with a surrounding region of at least 20,000 bp available (+/-10,000 bp). This region was exported and a clustered alignment with CLINKER[41] was performed. This gave some hints of functionally related gene clustering at the sequence level (**Fig. S12**), but potential relationships became much more apparent at the level of predicted tertiary (and quaternary) structures. We then used AlphaFold2 to predict the structure of all the proteins in the extracted genomic regions (~700 proteins) and the FoldSeek "easy-cluster" module to cluster these proteins[34]. From the ~700 proteins 216 "clusters" were identified, of which 17 contained proteins that appeared within ± 10,000 bp of the COMMD-like gene at least 10% of the time. **Fig. 6C** shows an example protein structure from each of the top ten structural clusters coloured by the occupancy in relation to the 55 COMMD-like proteins. While the functions of many of these proteins are uncharacterised, those with the highest correlation are known virulence factors including OB-fold proteins (60%), VRG1 proteins (60%) and PAAR domain (47%) proteins. As mentioned above this region resembles the *P. aeruginosa* vgrG1b gene cluster and contains a protein with the DUF2169 fold (63%) (**Fig. 6B**). For these proteins and those with less clear structural identities FoldSeek was used to search the protein data bank for any related "ground truth" structures (**Table S4**). In addition to those proteins already mentioned this included a Serine-Threonine kinase[42] (PDB: 4EQM), components of the *Mycobacterium tuberculosis* initial transcription complex[43] (PDB: 6KON) attached to a tetratricopeptide repeat protein[44] (PDB: 2PL2), HasB (PDB: 2M2K) and a thermal stable Old Yellow Enzyme[45] (PDB: 3KRU).

Lastly, we performed iterative multimeric AlphaFold predictions of the proteins within the COMMD-like gene cluster to determine if COMMD-like proteins potentially interact with any of the proteins that were encoded nearby, but were unable to identify any putative COMMD-like-containing complexes (not shown). However, we did observe predicted interactions among several proteins that appeared to be structurally related to bacterial Type VI secretion system (T6SS)

components. An example of such a complex from *Polyangium fumosum* is shown in **Fig. 6D**, composed of a homotrimer of TKD03570 and single copies each of TKD03571, TKD03569, TKD03568. The complex shares strong similarities to components of the T6SS baseplate machinery[46].

Other COMMD related proteins in Chlamydia, E. coli and Halobacteria

In addition to the Myxococcota bacteria, we also noted an expansion of COMMD-like proteins in *Chlamydia* and *Halobacteria*. The exact reason for this expansion is unclear, and a potential caveat is that due to their respective importance in disease and Archaeal evolution respectively, the *Chlamydia* and *Halobacteria* clades have many representative genome sequences compared with other species we identified. It is also worth noting that while the overall structure of the *Chlamydia* proteins conforms to the classical COMMD-fold, containing a helical HN domain and a COMM domain, they appear to function in a unique manner. This is seen in two previously determined crystal structures (PDB: 3Q9D and 4MLK[47]) in which they form a hexameric assembly coordinated by interactions between the HN domain as opposed to the classical β-sheet augmentation seen in the COMMD and COMMD-like structures presented here and elsewhere. Regardless we investigated the genomic context of these proteins to see if there was continuity with *Myxococcota*. This showed that these COMMD-like proteins clearly exist within species level clusters (**Fig S12-14**). In **Table S5** the clusters are summarised for an exemplar COMMD-like protein: O84588 for *Chlamydia trachomatis serovar D (strain ATCC VR-885 / DSM 19411 / UW-3/Cx)* and G0HSR7 for *Haloarcula hispanica (strain ATCC 33960 / DSM 4426 / JCM 8911 / NBRC 102182 / NCIMB 2187 / VKM B-1755)*. As before a combination of AlphaFold2 and FoldSeek was used to understand the type of genes these cluster with and if any complexes were forming with nearby encoded proteins. In both cases the COMMD-like protein was not predicted by AlphaFold to make direct interactions with any of the proteins within the respective gene cluster.

DISCUSSION

This work identifies archaeal and bacterial proteins that are structurally homologous to the eukaryotic COMMD protein family. Our analyses show that many species of bacteria and archaea encode either one or two COMMD-like proteins compared to the ten distinct genes that are conserved across the eukaryotic lineage. Of note, a recent preprint also reported the existence of COMMD-like proteins in Asgard archaea, supporting these findings[27]. While the evolutionary path from these single prokaryotic COMMD-like genes to the expanded eukaryotic repertoire is not entirely clear, our phylogenetic analyses suggest that eukaryotes did not acquire COMMD genes directly from their Asgard or alphaproteobacterial ancestors, but likely from another bacterial lineage. Among extant bacteria, Myxococcota appear to encode the sequences most closely related to eukaryotic COMMD proteins, although this inference may be limited by the resolving power of current phylogenetic methods on distantly related sequences.

Our structural and biophysical analyses confirm that the prokaryotic COMMD-like proteins are driven to form homodimers using the same structural principles as the heterodimers formed by specific eukaryotic COMMD family members. Furthermore, these can then assemble into homo-oligomeric ring complexes that are structurally analogous to the hetero-decameric rings formed by the ten eukaryotic proteins. Several other protein families in prokaryotes have some structural similarities to the COMMD proteins. These include the CT584/Cpn0803 proteins that are exclusively found in Chlamydiales[47,48], and the Whirly/Pur-α family of DNA binding proteins that are found in species across the three domains of life[49-54]. Despite superficial similarities, these are not strictly COMMD-like proteins. The CT584/Cpn0803 proteins have C-terminal domains that form dimers in a similar way to the COMM domain, and a helical N-terminal domain that has a related topology to the HN domain, but they assemble into very different hexamers that are unrelated to the homo and heterooligomeric rings formed by the COMMD protein family[47,48]. The Whirly/Pur-α proteins lack any structures related to the COMMD HN domain, but have small, repeated domains composed of a four-stranded β-sheet and a C-terminal α-helix that can form dimeric structures in a similar way to the COMM domain[49-54]. Unlike members of the COMMD or COMMD-like protein family Pur-α proteins have a highly charged surface allowing for DNA/RNA binding. Although it would appear COMMD-like proteins lack the surface residues required to directly bind DNA/RNA it is notable that in both Chlamydiales and Haloarchaea that genes for DNA modulation machinery are found in close proximity to the COMMD-like proteins.

Altogether these proteins do not share any sequence homology, lack key structural elements and assemble in very different ways to the COMMD proteins, so it is unclear if they share an evolutionary history or are examples of convergent evolution to superficially similar structural folds.

Many different protein families have evolved the ability to form multimeric ring structures, which most commonly involves homomeric symmetric interactions with one or a few repeated protein subunits[55-58]. Heteromeric pseudo-symmetric structures composed of multiple paralogous proteins, as seen for the eukaryotic COMMD decamer[16-18], are less common and typically arise from a series of gene duplication events converting homomeric into heteromeric complexes. Some prominent examples include the eukaryotic 20S proteosome[59], the TRiC/CCT chaperonin complex[60-62], the Sm/Lsm RNA-binding proteins[63-65], the CMG helicase[66-68] and the origin recognition complex[69,70] involved in initiating DNA replication. In the TRiC/CCT chaperonin complex for example, there is ample evidence of these duplication events and subsequent sequence divergence that creates eukaryotic heteromeric assemblies with increasing complexity that are structurally analogous to the homomeric counterparts in archaeal species[71]. In the case of the Sm/Lsm rings, much like the eukaryotic COMMD family, there is a clear distinction between eukaryotes, which possess a full complement of fourteen Sm/Lsm genes, and prokaryotes that have only a single Sm/Lsm paralogue, suggesting the duplication events occurred very early in the eukaryotic branch. It has been argued that the recruitment into the complex spliceosomal machinery is at least partly responsible for creating such a large paralogous family in eukaryotes derived from the single gene precursors in Archaea and Bacteria[72,73], and we similarly propose that the high degree of functional divergence and specialisation of the eukaryotic COMMD proteins from the prokaryotic COMMD-like proteins reflects their incorporation into the Commander and co-evolving WASH assemblies.

Unlike the COMMD proteins we did not uncover any prokaryotic proteins with clear similarity to the other subunits of Commander with the exception of VPS29 which has a typical metallophosphatase fold found in numerous proteins across the three domains of life and DENND10 which similarly contains the common DENN domain fold. The Commander complex is widely but not universally conserved in eukaryotes; for example, it has become secondarily simplified in some plants and is completely absent in fungi, although most eukaryotic lineages retain all ten subunits that date back to LECA[18]. During eukaryotic evolution, these subunits have

co-evolved with the endosomal actin remodelling WASH complex[21,24,26]. Thus, in eukaryotes, the COMMD proteins appear to be essential core components of the Commander machinery that regulate endosomal trafficking and cytoskeletal organisation. By contrast, the function(s) of the bacterial and archaeal COMMD-like proteins is currently unclear.

The origin of eukaryotes involved symbiosis between an Asgard archaeon and an Alphaproteobacterium[74]. The bacterium evolved to become the mitochondria of modern eukaryotes, while the Asgard contributed key nuclear and cytosolic machinery including information processing (that is, DNA replication, transcription and translation), membrane trafficking and cytoskeletal organisation[6-9,11,12,37,75]. Although COMMD-like proteins are present in Asgards[27], there do not appear to be any other components of the Commander or WASH machinery in Archaea that suggest a nascent role in membrane or cytoskeletal regulation. Indeed, phylogenetic analysis instead suggests that the eukaryotic COMMD family is most closely related to sequences from modern *Myxococcota*. Several other core eukaryotic genes with roles in membrane biology show phylogenetic affinities with *Myxococcota*, including key genes of sterol biosynthesis[35,74]. One possibility is that these were acquired by horizontal gene transfer during the origin of eukaryotes, potentially mediated by viruses[76]. An alternative explanation is provided by the syntrophy hypothesis for the origin of eukaryotes, in which the symbiotic interactions between Asgard and Alphaproteobacterium occurred within a delta-proteobacterial (potentially myxococcotal) host[28]. However, it is currently unclear whether an additional cell biological partner is needed to explain the acquisition of these genes from *Myxococcota* during early eukaryotic evolution, particularly given evidence for additional genetic contributions from other lineages[77,78]. We were unable to conclusively determine the function of the closest relatives of COMMD in modern *Myxococcota*, although our synteny analyses suggest a potential association with a Type VI related secretion system in the *Myxococcota* clade.

In summary, the combination of genome-wide structure prediction, and structural homology searches and phylogenetics have provided a powerful and comprehensive approach for identifying remote protein homologies that are difficult - if not impossible - using sequence-based methods alone. Eukaryotic COMMD proteins have a unique structural organisation that drives heteromeric assembly into stable ring-shaped complexes, and both archaea and bacteria possess ancestral COMMD-like proteins that form analogous complexes with only a single repeated subunit. This supports the idea that the eukaryotic genes have evolved from a single ancestral

prokaryotic gene and have undergone repeated duplication and evolution to allow for a highly specific set of heteromeric interactions to occur. While the functions of the prokaryotic COMMD-like proteins remain obscure, and the evolutionary process that promoted COMMD duplication and co-evolution with other Commander subunits is currently unknown, these questions will present a rich avenue for future investigation on the origins of eukaryotes. Furthermore, studies of the simpler prokaryotic COMMD-like proteins may provide new insights into the functions of the eukaryotic COMMD assembly.

METHODS

Structure-based homology search

FoldSeek[29], with the full UniProt database, was installed locally. The initial analysis utilized a folder containing the predicted structures of the following human proteins: COMMD1-10 (excluding COMMD6 to focus on full-length COMMD proteins), VPS26C, VPS29, VPS35L, CCDC22, CCDC93, and DENND10. This folder was used to search the UniProt database, restricted to eukaryotic sequences. Default settings were applied, except for coverage, which was set to 90% to exclude PUR-α and its homologues from the results. Proteins identified with a confidence score of 1 were downloaded from the AlphaFold database to create a new starting dataset. This updated folder was then used in a second screening of the UniProt database. The iterative screening process continued until the increase in the total number of identified proteins was less than 2%. The final datasets obtained from the eukaryotic screen were subsequently used as starting points for searches in the archaeal and bacterial databases. The same iterative method was applied to these datasets.

Protein structure predictions

All individual proteins were obtained from the AlphaFold2 database[31] unless no structure was available. In such cases structures were generated using a local instillation of the ColabFold[79] v1.5.5 implementation of AlphaFold2-PTM[30]. A single model was used with 3 recycles to generate additional structures, quality of model was assessed by the ptm score. Protein complex predictions were generated using ColabFold v1.5.5 implementation of AlphaFold2-multimer[79,80], 5 models were generated with 3 recycles. Model quality was determined by extent of overlap and iPTM score. All images were generated using chimeraX[81].

Molecular cloning

The sequences for 10 full-length COMMD-like proteins were synthesized and cloned for bacterial expression into the pET-28a(+) vector by Gene Universal (USA). The two different cloning sites are NcoI & XhoI (A0A558G9V7, A0A6A7KI92, A4J5H3, S4XVP2, S4Y171) and NdeI & XhoI (A0A2U3LUU8, A0A2U3L5I1, A0A7C1Q7X0, MBK5214510, UYP43922). S4XVP2 and S4Y171 were ordered as gene fragments and cloned into a modified version of the LM627 vector (addgene: 191551) containing GST using the Golden Gate cloning method as described by Wicky

et al[82]. S4XVP2 and S4Y171 produced more homogeneous higher quality protein when expressed with a N-terminal GST tag compared with the C-terminal His tag.

Protein expression and purification

All the proteins were expressed using *Escherichia coli* BL21(DE3) competent cells grown in the presence of kanamycin at 37°C. Protein expression was induced by 1 mM isopropyl β-D-thiogalactopyranoside (IPTG) when cell density OD600 reached 0.8, the temperature was then reduced to 25°C and cultures were incubated overnight. The cells were harvested the following day by centrifugation (Beckman JLA 8.1 rotor) at 7,000 xg, 4 °C for 10 minutes and resuspended in lysis buffer (500 mM NaCl, 50 mM Tris-HCl (pH 8.0), 5 mM imidazole, 10 % glycerol, 2mM β-mercaptoethanol, benzamidine, and DNAseI).

For protein purification, cells were firstly lysed using a Constant systems cell disruptor at 35 Kpsi, followed by centrifugation (Beckman JLA 16.250 rotor) at 20,000 xg, 4°C for 30 minutes, after which the soluble His-tagged proteins were contained in the supernatant. Subsequently, TALON resin (Bio-strategy Pty Limited) was added to the supernatant for binding His-tagged proteins and eluted 500 mM NaCl, 500 mM imidazole, 10% glycerol, 2 mM β-mercaptoethanol. Proteins were further purified using pre-equilibrated size exclusion chromatography (SEC) columns in 100 mM NaCl, 20 Tris-HCl (pH 8.0), either a Superdex 200 Increase 10/300 GL column (Cytiva) or HiLoad 16/600 Superdex 200 pg column (Cytiva) attached to an AKTA system (GE Healthcare). The purified protein fractions were checked by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) using Bolt 12 % Bis-Tris gel in 1x MES running buffer at 165 V for 45 minutes.

Size exclusion chromatography-Multiangle laser light scattering (SEC-MALLS)

Each protein sample at 1 mg/ml (30 ul) was injected onto the Superose 6 increase 5/150 GL column equilibrated with the buffer containing 100 mM NaCl, 20 mM Tris-HCl (pH 8.0). MALS was performed using a DAWN HELEOS II 10-angle light-scattering detector paired with an Optilab rEX refractive index detector (Wyatt Technology), attached to a Prominence high-performance liquid chromatography (HPLC) (Shimadzu) at a flow rate of 0.25 ml/min.

Single-Particle Mass Photometry

Molecular mass of all the proteins was measured by a Refeyn OneMP mass photometer (Refeyn Ltd) in the CMM. Initially bovine serum albumin (BSA) from the Sigma-Aldrich was applied as a mass calibration for MBK5214510, given MBK5214510 ability to for concentration dependent oligomers this was used in subsequent measures for calibration. For actual measurement, 10 µl of 100 nM protein samples (200 nM for A0A2U3L5I1, A0A558G9V7) was added to the coverslip. Movie were collected for 180 s. The data was then analysed using Aquire^{MP} software and generated a gaussian distributions to determine the protein mass composition.

Nano-Differential Scanning Fluorimetry.

NanoDSF was conducted using Prometheus NT.48 equipped with backscattering detectors (NanoTemper Technologies, München, Germany) at the Protein Expression Facility (PEF, The University of Queensland). Each protein sample was prepared in triplicate at a concentration of 1 mg/ml and loaded into individual NanoDSF grade high sensitivity capillaries (NanoTemper Technologies, München, Germany). The samples were then exposed at 100% excitation power, with a temperature gradient ranging from 20°C to 95°C with a ramping rate of 1°C/min. Tryptophan fluorescence emission was collected at 330 nm and 350 nm using a dual-UV detection, meanwhile, protein aggregation was monitored by backreflection optics. The PR.ThermControl software was used to calculate thermal stability parameters (Tonset, Tm, and Tagg).

Crystallisation and structure determination

Crystallization trials were performed on a Mosquito liquid handling robot (TTP LabTech) in the commercially available SPARSE matrix 96-well hanging-drop plates (Index, LMB, MIDASplus, PEG/Ion, PEGRx and JCSG+) at 20 °C. Prior to sample loading into wells, protein samples (except S4XVP2, S4Y171) were concentrated to ~10 mg/ml (determined by Bradfold). Crystals hits were observed for all proteins across different plates. To scale up the drop and crystal size, several conditions were optimized in a 24-well hanging drop vapor diffusion (Hampton) plate covered by glass cover slips. On each glass coverslip, three drops were prepared with protein and reservoir solution mixed in ratios of 1:1, 2:1 and 3:1 (protein:reservoir solution). Glycerol (10 %) was added into the mother liquor as cryo-protectant before flash-freezing in liquid nitrogen. The crystals used for MBK5214510 (15mg/ml) structure determination were grown in MIDASplus C9 (10 % v/v

DMSO, 0.1M NaCl and 25% PE5/4 which in this case served as the cryo-protectant) a final concentration of 15mg/ml.

X-ray diffraction data were collected at the Australian synchrotron on the MX2 beamline. The data was indexed and integrated with XDS[83] and scaled with AIMLESS withing the CCP4 suite[84]. The initial structure was resolved using molecular replacement in Phaser[85] with the AlphaFold2 structure as an initial template. Further refinements of the structure were performed with PHENIX[86] , COOT[87] and the chimera plugin ISOLDE[88].

Cryo-EM grid preparation and data collection

Freshly glow-discharged Quantifoil 1.2/1.3 Copper 200 mesh grids were used for cryo-EM data collection. For MBK5214510, A0A6A7KI92, S4XVP2 and S4Y171 a final concentration of 10 mg/ml in 100 mM NaCl, 20 Tris-HCl (pH 8.0) was loaded onto grids. Grid vitrification was performed in liquid ethane using a Vitrobot Mark IV plunge freezer (Thermo Fisher), using a 3 µL sample application, 95% humidity, 4°C, minus 1 blot force and 2.5 s blot time. After freezing, grids were stored immediately under liquid nitrogen until image acquisition.

Cryo-EM data for MBK5214510 was acquired using a JOEL CRYOARM 300 (JOEL) microscope operated at an acceleration voltage of 300 kV, equipped with an in column Ω energy filter (20 eV slit width) and a Gatan K3 direct electron detector. A total of 1,905 movies, at 60,000 x magnification, with a pixel size of 0.82 Å per pixel and a total exposure dose of 40 e/Å2, within a target defocus range from −2.5 to −0.5 µm. A0A6A7KI92 movies (769 movies) were collected on JOEL CRYOARM200 at 200 kV under 100,000 x magnification (corresponding to calibrated pixel sizes of 0.52 Å/pixel). This microscope included a cold-field emission gun (FEG), an in-column Ω energy filter and a Gatan K2 direct electron detector. As before the total exposure dose was maintained at 40 e/Å2, with the defocus range adjusted between −2.5 and −0.5 µm.

Cryo-TEM data for S4XVP2 and S4Y171 were collected on a Glacios 2 cryo-transmission electron microscope (Thermo Fisher Scientific) operated at an acceleration voltage of 200 kV and equipped with an integrated Thermo Scientific Falcon 4i Direct Electron Detector, an extreme high-brightness field emission gun (X-FEG), and Thermo Scientific EPU software. A total of 5,993 and 3,432 movies, respectively, were recorded at 190,000 x magnification, with a pixel size of 0.75 Å per pixel, a total electron exposure of 50 e/Å2, within a target defocus range from -2.5 to -0.5 µm.

Cryo-EM data processing

Data processing of MBK5214510, A0A6A7KI92, S4Y171 and S4XVP2 was performed using the University of Queensland's remote web access desktops (https://github.com/UQ-RCC/hpc-docs/blob/main/guides/OnDemand-Guide.md) in CryoSPARC V4.6.1[89]. For MBK5214510, 1,905 movies were imported and motion corrected using Patch Motion Correction, followed by CTF estimation. During the subsequent exposure curation step, the contrast transfer function (CTF) fit resolution was restricted to 6 Å, and the median relative ice thickness was adjusted to 1.031, resulting in the retention of 1,331 micrographs. From this dataset, 253,056 particles were initially blob-picked with a particle diameter range from 100 Å to 160 Å and template-picked at 160 Å. Following iterative rounds of 2D classification (**Table S2**), only the high-quality 2D classes were selected as templates for further template picking (150 Å particle diameter) and 2D classification (**Fig S7**). Ultimately, 120,876 particles were chosen and used to construct two ab-initio models. At this point, octameric and decameric structure volumes were separated and reconstructed using non-uniform (NU) refinement with D4 and D5 symmetry imposed. The final maps were obtained at resolutions of 4.22 Å for the octamer, and 4.77 Å for the decamer.

The 769 movies collected for A0A6A7KI92 were processed similarly as MBK5214510, with a particle diameter ranging from 120 Å to 200 Å for blob picking and 100 Å for template picking. As a result, 66,476 particles were selected from 2D classification. Due to a strong bias in orientations towards the top view, no further processing was attempted.

The 5,993 and 3,432 movies for S4Y171 and S4XVP2, respectively, were initially processed as MBK5214510. However after particle extraction the protocol varied, in both cases we used more stringent 2D classification and ab-initio model building following a recent protocol from Kim et al[90]. Following multiple rounds of 2D classification we selected a small subset of extremely high quality particles to build a high quality ab-initio model (**Fig S8**). This initial volume was then used in 2 rounds of heterogenous refinement with 4 junk classes generated from a prematurely terminated ab-initio job. These particle stacks were then refined using NU-refinement with an imposed D4 symmetry (**Fig S8**).

Sequence-based phylogenetic analysis

Prokaryotic sequences from above were aligned using MAFFT v7.505[91] (--linsi), this alignment was used to infer maximum likelihood trees using IQTREE 2.2.6[92], using 20 profile mixture

models (LG+C20)[93] with empirical amino acid frequencies derived from the data (+F), four gamma categories (-G)[94], with 10,000 ultrafast bootstraps[95]. Additional trees were inferred under the same settings, but with alternative constraints built using TreeViewer[96], one constrained with the Myxococcota to be monophyletic with respect to other bacteria in the tree, the other constrained the archaeal sequences to be sister to the eukaryote COMMD proteins (to the exclusion of all bacterial sequences). These constrained topologies and the maximum likelihood topology were then compared using the approximately unbiased[97] test under the same model, with 100,000 RELL replicates[98]. This analysis could not significantly reject ($P < 0.05$) the *Myxococcota* clade branching with other bacteria, or the archaeal sequences being more closely related to the eukaryotic COMMD sequences. However, imposing the archaeal constraint resulting in a subtending branch uniting eukaryotes and archaea of length 0.0000030 substitutions per site, suggesting the archaeal sequences are unlikely to be sister to the eukaryotic sequences. In all of these analyses, the Asgard sequences were resolved as only distantly related to the eukaryotic sequences, perhaps because of relatively recent horizontal acquisition from bacteria.

Structure-based phylogenetic analysis.

Protein structures were downloaded from the AlphaFold database (where available) or predicted using AlphaFold2. Structural alignment was performed using Foldmason[99] and visualised with AliView[100]. Phylogenetic trees were inferred from the 3Di alignment with IQ-TREE v3.0.183[92], 10,000 ultrafast bootstraps[95], using the Foldseek 3Di matrix[29]. This is a fixed scoring matrix that was pre-trained with structurally aligned proteins to model the transition probabilities between structural 3Di states. Scores are based on observed versus expected pair frequencies using empirical background frequencies of 3Di states. Phylogenies were visualised and prepared for publication using iTOL v7.5[101].

Genomic context analysis

COMMD-like proteins were identified and enterz was used to download the region 10,000bp upstream and downstream of that reference point. CLINKER[102] was then used to align these genomic regions based on sequence identity. In the case of Myxcoccota sequences they were extracted from the .gbk file and the structure of those sequences was predicted using AlphaFold2 as described above.

DATA AVAILABILITY

The crystallographic and cryo-electron microscopy data generated in this study have been deposited in the PDB database under accession code 9Q6A [https://doi.org/10.2210/pdb9Q6A/pdb], 9Q6C [https://doi.org/10.2210/pdb9Q6C/pdb], 9Q6D [https://doi.org/10.2210/pdb9Q6D/pdb], 9ZA1 [https://doi.org/10.2210/pdb97A1/pdb] and 9ZA2 [https://doi.org/10.2210/pdb97A2/pdb]. In addition, the associated cryo-electron density maps have been deposited in the EMDB uder accession code EMD-72266 (9Q6C) [https://www.ebi.ac.uk/emdb/EMD-72266], EMD-72267 (9Q6D) [https://www.ebi.ac.uk/emdb/EMD-72267], EMD-73963 (9ZA1) [https://www.ebi.ac.uk/emdb/EMD-73963] and EMD-73964 (9ZA2) [https://www.ebi.ac.uk/emdb/EMD-73964]. The AlphaFold2 modelling data used in this study are available in the ModelArchive database under accession code ma-commd-like [https://modelarchive.org/doi/10.5452/ma-commd-like].

REFERENCES

1. Schlacht, A., Herman, E.K., Klute, M.J., Field, M.C. & Dacks, J.B. Missing pieces of an ancient puzzle: evolution of the eukaryotic membrane-trafficking system. *Cold Spring Harb Perspect Biol* **6**, a016048 (2014).

2. Vosseberg, J. et al. The emerging view on the origin and early evolution of eukaryotic cells. *Nature* **633**, 295–305 (2024).

3. Donoghue, P.C.J. et al. Defining eukaryotes to dissect eukaryogenesis. *Curr Biol* **33**, R919–r929 (2023).

4. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

5. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

6. Hatano, T. et al. Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nature Communications* **13**, 3398 (2022).

7. Tran, L.T., Akıl, C., Senju, Y. & Robinson, R.C. The eukaryotic-like characteristics of small GTPase, roadblock and TRAPPC3 proteins from Asgard archaea. *Commun Biol* **7**, 273 (2024).

8. Vargová, R. et al. The Asgard archaeal origins of Arf family GTPases involved in eukaryotic organelle dynamics. *Nat Microbiol* (2025).

9. Neveu, E., Khalifeh, D., Salamin, N. & Fasshauer, D. Prototypic SNARE Proteins Are Encoded in the Genomes of Heimdallarchaeota, Potentially Bridging the Gap between the Prokaryotes and Eukaryotes. *Curr Biol* **30**, 2468–2480.e5 (2020).

10. Akıl, C. & Robinson, R.C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).

11. Akıl, C. et al. Structure and dynamics of Odinarchaeota tubulin and the implications for eukaryotic microtubule evolution. *Sci Adv* **8**, eabm2225 (2022).

12. Rodrigues-Oliveira, T. et al. Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).

13. Chen, K.E., Healy, M.D. & Collins, B.M. Towards a molecular understanding of endosomal trafficking by Retromer and Retriever. *Traffic* **20**, 465–478 (2019).

14. Cullen, P.J. & Steinberg, F. To degrade or not to degrade: mechanisms and significance of endocytic recycling. *Nat Rev Mol Cell Biol* **19**, 679–696 (2018).

15. Gopaldass, N., Chen, K.E., Collins, B. & Mayer, A. Assembly and fission of tubular carriers mediating protein sorting in endosomes. *Nat Rev Mol Cell Biol* **25**, 765–783 (2024).

16. Laulumaa, S., Kumpula, E.-P., Huiskonen, J.T. & Varjosalo, M. Structure and interactions of the endogenous human Commander complex. *Nat Struct Mol Biol* **31**, 925–938 (2024).

17. Boesch, D.J. et al. Structural organization of the retriever-CCC endosomal recycling complex. *Nat Struct Mol Biol* **31**, 910–924 (2024).

18. Healy, M.D. et al. Structure of the endosomal Commander complex linked to Ritscher-Schinzel syndrome. *Cell* **186**, 2219–2237.e29 (2023).

19. Kato, K. et al. Ritscher-Schinzel syndrome can be characterized as an endosomal recyclinopathy. *Sci Transl Med* **17**, eadt2426 (2025).

20. Healy, M.D. et al. Structural insights into the architecture and membrane interactions of the conserved COMMD proteins. *eLife* **7**, e35898 (2018).

21. Healy, M.D. & Collins, B.M. The PDLIM family of actin-associated proteins and their emerging role in membrane trafficking. *Biochemical Society Transactions* **51**, 2005–2016 (2023).

22. Maine, G.N. & Burstein, E. COMMD proteins: COMMing to the scene. *Cell Mol Life Sci* **64**, 1997–2005 (2007).

23. Liebeskind, B.J., Aldrich, R.W. & Marcotte, E.M. Ancestral reconstruction of protein interaction networks. *PLoS Comput Biol* **15**, e1007396 (2019).

24. Mallam, A.L. & Marcotte, E.M. Systems-wide Studies Uncover Commander, a Multiprotein Complex Essential to Human Development. *Cell Syst* **4**, 483–494 (2017).

25. Shinde, A.P., Kučerová, J., Dacks, J.B. & Tachezy, J. The retromer and retriever systems are conserved and differentially expanded in parabasalids. *J Cell Sci* **137**(2024).

26. McNally, K.E. et al. Retriever is a multiprotein complex for retromer-independent endosomal cargo recycling. *Nature Cell Biology* **19**, 1214–1225 (2017).

27. Köstlbacher, S. et al. Structure-based inference of eukaryotic complexity in Asgard archaea. *bioRxiv*, 2024.07.03.601958 (2024).

28. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nature Microbiology* **5**, 655–667 (2020).

29. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2024).

30. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

31. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

32. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2021).

33. Garg, S.G. & Hochberg, G.K.A. A General Substitution Matrix for Structural Phylogenetics. *Molecular Biology and Evolution* **42**, msaf124 (2025).

34. Puente-Lelievre, C. et al. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv*, 2023.12.12.571181 (2024).

35. Hoshino, Y. & Gaucher, E.A. Evolution of bacterial steroid biosynthesis and its impact on eukaryogenesis. *Proceedings of the National Academy of Sciences* **118**, e2101276118 (2021).

36. Greening, C. et al. Minimal and hybrid hydrogenases are active from archaea. *Cell* **187**, 3357–3372.e19 (2024).

37. Lu, Z., Zhang, S., Liu, Y., Xia, R. & Li, M. Origin of eukaryotic-like Vps23 shapes an ancient functional interplay between ESCRT and ubiquitin system in Asgard archaea. *Cell Rep* **43**, 113781 (2024).

38. Hatano, T. et al. Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nat Commun* **13**, 3398 (2022).

39. Hennell James, R. et al. Functional reconstruction of a eukaryotic-like E1/E2/(RING) E3 ubiquitylation cascade from an uncultured archaeon. *Nat Commun* **8**, 1120 (2017).

40. Pissaridou, P. et al. The Pseudomonas aeruginosa T6SS-VgrG1b spike is topped by a PAAR protein eliciting DNA damage to bacterial competitors. *Proc Natl Acad Sci U S A* **115**, 12519–12524 (2018).

41. Gilchrist, C.L.M. & Chooi, Y.-H. clinker and clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).

42. Rakette, S., Donat, S., Ohlsen, K. & Stehle, T. Structural Analysis of Staphylococcus aureus Serine/Threonine Kinase PknB. *PLOS ONE* **7**, e39136 (2012).

43. Li, L., Molodtsov, V., Lin, W., Ebright, R.H. & Zhang, Y. RNA extension drives a stepwise displacement of an initiation-factor structural module in initial transcription. *Proceedings of the National Academy of Sciences* **117**, 5801–5809 (2020).

44. Lim, H., Kim, K., Han, D., Oh, J. & Kim, Y. Crystal Structure of TTC0263, a Thermophilic TPR Protein from Thermus thermophilus HB27. *Molecules and Cells* **24**, 27–36 (2007).

45. Adalbjörnsson, B.V. et al. Biocatalysis with Thermostable Enzymes: Structure and Properties of a Thermophilic 'ene'-Reductase related to Old Yellow Enzyme. *ChemBioChem* **11**, 197–207 (2010).

46. Wang, J., Brodmann, M. & Basler, M. Assembly and Subcellular Localization of Bacterial Type VI Secretion Systems. *Annu Rev Microbiol* **73**, 621–638 (2019).

47. Barta, M.L. et al. Structure of CT584 from Chlamydia trachomatis refined to 3.05 Å resolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **69**, 1196–201 (2013).

48. Stone, C.B. et al. Structural characterization of a novel Chlamydia pneumoniae type III secretion-associated protein, Cpn0803. *PLoS One* **7**, e30220 (2012).

49. Weber, J. et al. Structural basis of nucleic-acid recognition and double-strand unwinding by the essential neuronal protein Pur-alpha. *Elife* **5**(2016).

50. Cappadocia, L., Parent, J.S., Sygusch, J. & Brisson, N. A family portrait: structural comparison of the Whirly proteins from Arabidopsis thaliana and Solanum tuberosum. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **69**, 1207–11 (2013).

51. Cappadocia, L. et al. Crystal structures of DNA-Whirly complexes and their role in Arabidopsis organelle genome repair. *Plant Cell* **22**, 1849–67 (2010).

52. Graebsch, A., Roche, S., Kostrewa, D., Söding, J. & Niessing, D. Of bits and bugs--on the use of bioinformatics and a bacterial crystal structure to solve a eukaryotic repeat-protein structure. *PLoS One* **5**, e13402 (2010).

53. Graebsch, A., Roche, S. & Niessing, D. X-ray structure of Pur-alpha reveals a Whirly-like fold and an unusual nucleic-acid binding surface. *Proc Natl Acad Sci U S A* **106**, 18521–6 (2009).

54. Janowski, R. & Niessing, D. The large family of PC4-like domains - similar folds and functions throughout all kingdoms of life. *RNA Biol* **17**, 1228–1238 (2020).

55. Pieters, B.J., van Eldijk, M.B., Nolte, R.J. & Mecinovic, J. Natural supramolecular protein assemblies. *Chem Soc Rev* **45**, 24–39 (2016).

56. Marsh, J.A. & Teichmann, S.A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem* **84**, 551–75 (2015).

57. Mallik, S. & Tawfik, D.S. Determining the interaction status and evolutionary fate of duplicated homomeric proteins. *PLoS Comput Biol* **16**, e1008145 (2020).

58. Mallik, S., Tawfik, D.S. & Levy, E.D. How gene duplication diversifies the landscape of protein oligomeric state and function. *Curr Opin Genet Dev* **76**, 101966 (2022).

59. Murata, S., Yashiroda, H. & Tanaka, K. Molecular mechanisms of proteasome assembly. *Nat Rev Mol Cell Biol* **10**, 104–15 (2009).

60. Betancourt Moreira, K. et al. A hierarchical assembly pathway directs the unique subunit arrangement of TRiC/CCT. *Mol Cell* **83**, 3123–3139.e8 (2023).

61. Han, W. et al. Structural basis of plp2-mediated cytoskeletal protein folding by TRiC/CCT. *Sci Adv* **9**, eade1207 (2023).

62. Willison, K.R. The structure and evolution of eukaryotic chaperonin-containing TCP-1 and its mechanism that folds actin into a protein spring. *Biochem J* **475**, 3009–3034 (2018).

63. Montemayor, E.J. et al. Molecular basis for the distinct cellular functions of the Lsm1-7 and Lsm2-8 complexes. *RNA* **26**, 1400–1413 (2020).

64. Mura, C., Randolph, P.S., Patterson, J. & Cozen, A.E. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. *RNA Biol* **10**, 636–51 (2013).

65. Wilusz, C.J. & Wilusz, J. Eukaryotic Lsm proteins: lessons from bacteria. *Nat Struct Mol Biol* **12**, 1031–6 (2005).

66. Burnham, D.R., Kose, H.B., Hoyle, R.B. & Yardimci, H. The mechanism of DNA unwinding by the eukaryotic replicative helicase. *Nat Commun* **10**, 2159 (2019).

67. Yuan, Z. et al. DNA unwinding mechanism of a eukaryotic replicative CMG helicase. *Nat Commun* **11**, 688 (2020).

68. Riera, A. et al. From structure to mechanism-understanding initiation of DNA replication. *Genes Dev* **31**, 1073–1088 (2017).

69. Bleichert, F., Botchan, M.R. & Berger, J.M. Crystal structure of the eukaryotic origin recognition complex. *Nature* **519**, 321–6 (2015).

70. Jaremko, M.J., On, K.F., Thomas, D.R., Stillman, B. & Joshua-Tor, L. The dynamic nature of the human origin recognition complex revealed through five cryoEM structures. *Elife* **9**(2020).

71. Archibald, J.M., Logsdon, J.M., Jr. & Doolittle, W.F. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol Biol Evol* **17**, 1456–66 (2000).

72. Veretnik, S., Wills, C., Youkharibache, P., Valas, R.E. & Bourne, P.E. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Comput Biol* **5**, e1000315 (2009).

73. Mura, C., Randolph, P.S., Patterson, J. & Cozen, A.E. Archaeal and eukaryotic homologs of Hfq. *RNA Biology* **10**, 636–651 (2013).

74. Moreira, D. & López-García, P. Symbiosis Between Methanogenic Archaea and δ-Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. *Journal of Molecular Evolution* **47**, 517–530 (1998).

75. Souza, D.P. et al. Evolutionarily conserved principles of ESCRT-III-mediated membrane remodelling revealed by a two-subunit Asgard archaeal system. *bioRxiv*, 2024.07.01.601590 (2024).

76. Krupovic, M., Kuhn, J.H., Fischer, M.G. & Koonin, E.V. Natural history of eukaryotic DNA viruses with double jelly-roll major capsid proteins. *bioRxiv* (2024).

77. Bernabeu, M., Manzano-Morales, S., Marcet-Houben, M. & Gabaldón, T. Diverse ancestries reveal complex symbiotic interactions during eukaryogenesis. *bioRxiv*, 2024.10.14.618062 (2024).

78. Stairs, C.W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci Adv* **6**, eabb7258 (2020).

79. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).

80. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034 (2022).

81. Meng, E.C. et al. UCSF ChimeraX: Tools for structure building and analysis. *Protein Science* **32**, e4792 (2023).

82. Wicky, B.I.M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

83. Kabsch, W. XDS. *Acta Crystallographica Section D* **66**, 125–132 (2010).

84. Agirre, J. et al. The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallographica Section D* **79**, 449–461 (2023).

85. McCoy, A.J. et al. Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658–674 (2007).

86. Adams, P.D. et al. The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94–106 (2011).

87. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

88. Croll, T.I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519–530 (2018).

89. Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290–296 (2017).

90. Kim, K., Li, H. & Clarke, O.B. High-resolution ab initio reconstruction enables cryo-EM structure determination of small particles. *bioRxiv*, 2025.09.08.674935 (2025).

91. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–80 (2013).

92. Minh, B.Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).

93. Si Quang, L., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).

94. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306–314 (1994).

95. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).

96. Bianchini, G. & Sánchez-Baracaldo, P. TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecology and Evolution* **14**, e10873 (2024).

97. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492–508 (2002).

98. Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* **31**, 151–160 (1990).

99. Gilchrist, C.L.M., Mirdita, M. & Steinegger, M. Multiple Protein Structure Alignment at Scale with FoldMason. *bioRxiv*, 2024.08.01.606130 (2024).

100. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

101. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **52**, W78–W82 (2024).

102. Gilchrist, C.L.M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).

103. The University of Queensland Research Computing Centre. *Bunya supercomputer* (2024).

ACKNOWLEDGEMENTS

We acknowledge use of the University of Queensland Remote Operation Crystallization and X-ray (UQ ROCX) Facility. X-ray data were collected on the MX1 and MX2 microfocus beamline at the Australian Synchrotron. The authors acknowledge the support of the Microscopy Australia Research Facility at the Centre for Microscopy and Microanalysis at The University of Queensland. We wish to acknowledge The University of Queensland's Research Computing Centre (RCC) for its support in this research[103]. We would like to thank Prof. Paul Curmi for prompting this work with an insightful suggestion over drinks at the Lorne Protein meeting. To paraphrase, "Remember how during your PhD you thought it might be fun to look for homologues of your protein in archaea? You should try that again!".

FUNDING

B.M.C. is supported by an Investigator Grant from the National Health and Medical Research Council (APP2016410) and an Australian Research Council (ARC) Discovery Project Grant (DP240101315). M.D.H is supported by an Investigator Grant from the National Health and Medical Research Council (GNT2042760) and a project grant from Dementia Australia Research Foundation. C.P-L is supported by the Human Frontier Science Program Grant No. RGY0072/2021 and the University of Auckland Faculty of Science RDF #3732317. T.A.W. is supported by a grant from the Gordon and Betty Moore Foundation (GBMF9741) and the John Templeton Foundation (63451); the opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

AUTHOR CONTRIBUTIONS

Conceptualization, B.M.C, M.D.H;

Methodology, M.L, F.B, K.E.C, E.J.S, R.J.C, B.M.C, M.D.H, E.R.R.M, T.A.W, C.P-L;

Investigation, M.L, F.B, K.E.C E.R.R.M, C.P-L, M.D.H, B.M.C;

Writing – Original Draft, B.M.C, M.D.H;

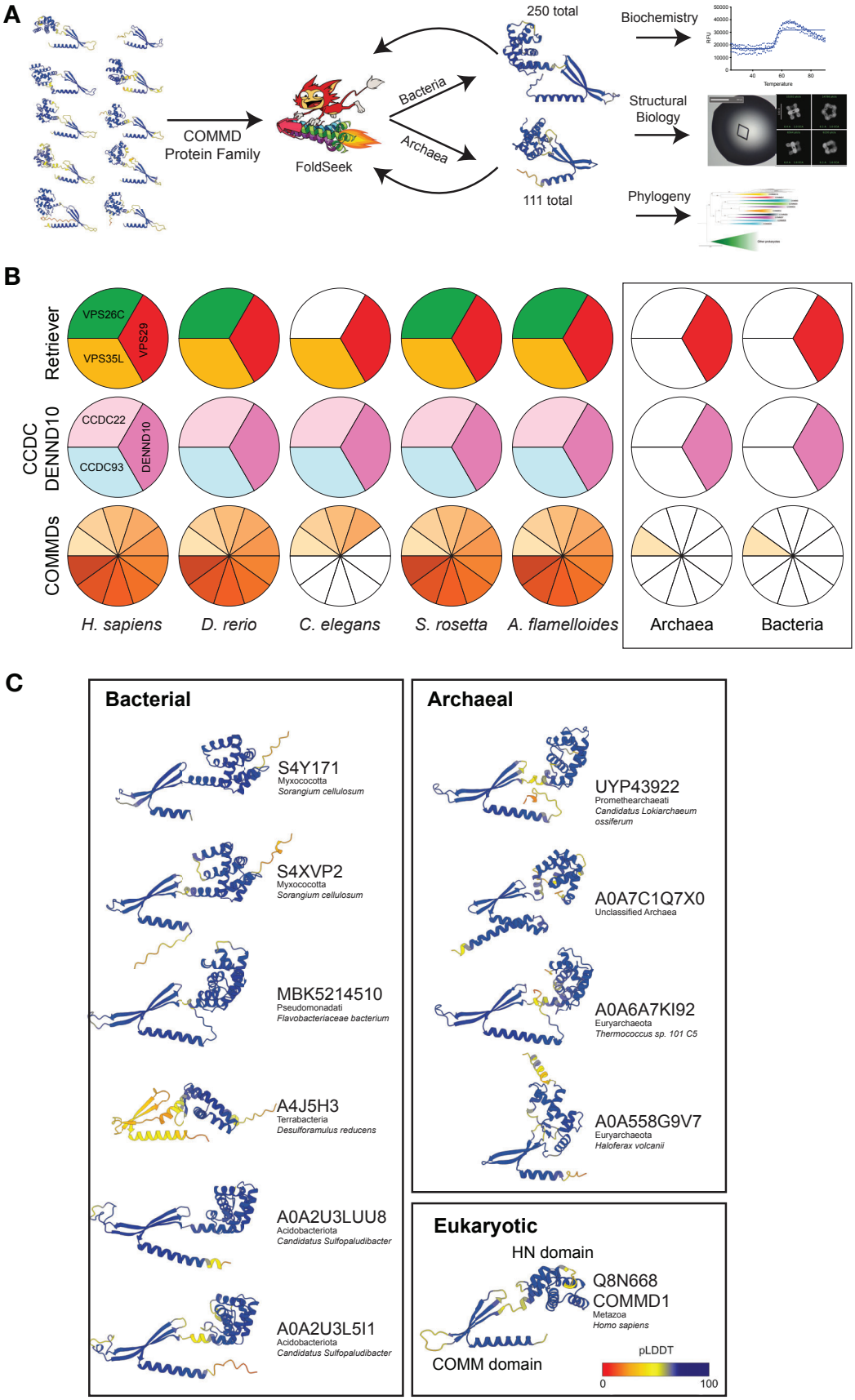
Writing – Review & Editing, M.L, F.B, K.E.C, E.J.S, R.J.C, B.M.C, M.D.H, E.R.R.M, T.A.W, C.P-L, K.A.M;

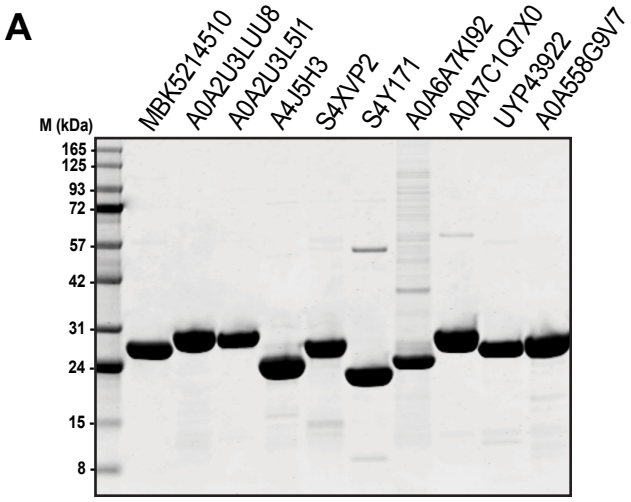
Funding Acquisition, R.J.C, B.M.C, M.D.H;

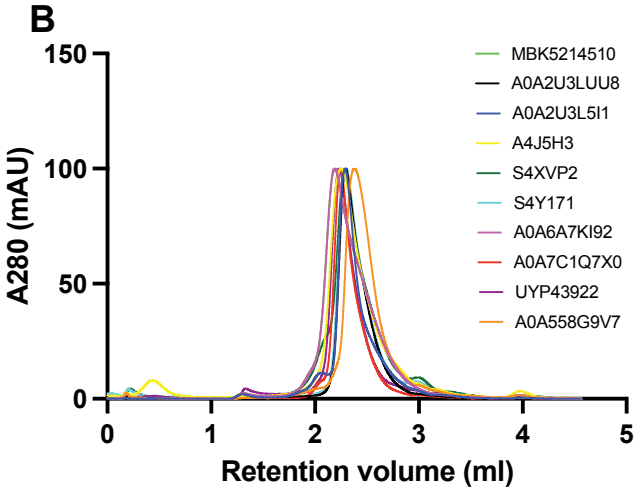
Supervision, B.M.C, M.D.H.

CONFLICTS OF INTEREST

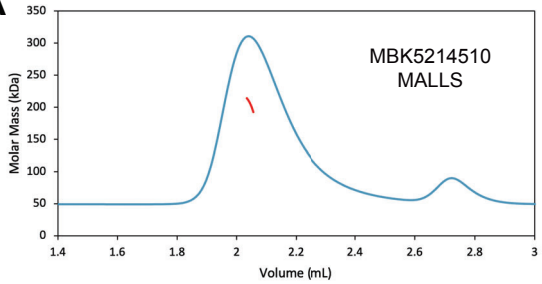
The authors declare no conflict of interest.


A
COMMD
Protein Family
FoldSeek
Bacteria
Archaea
250 total
111 total
Biochemistry
Structural
Biology
Phylogeny
B
Retriever
CCDC
DENND10
COMMDs
VPS26C
VPS29
VPS35L
CCDC22
DENND10
CCDC93
H. sapiens
D. rerio
C. elegans
S. rosetta
A. flamelloides
Archaea
Bacteria
C
Bacterial
S4Y171
Myxococcota
Sorangium cellulosum
S4XVP2
Myxococcota
Sorangium cellulosum
MBK5214510
Pseudomonadati
Flavobacteriaceae bacterium
A4J5H3
Terrabacteria
Desulforamulus reducens
A0A2U3LUU8
Acidobacteriota
Candidatus Sulfopaludibacter
A0A2U3L5I1
Acidobacteriota
Candidatus Sulfopaludibacter
Archaeal
UYP43922
Promethearchaeati
Candidatus Lokiarchaeum ossiferum
A0A7C1Q7X0
Unclassified Archaea
A0A6A7KI92
Euryarchaeota
Thermococcus sp. 101 C5
A0A558G9V7
Euryarchaeota
Haloferax volcanii
Eukaryotic
HN domain
Q8N668
COMMD1
Metazoa
Homo sapiens
COMM domain
pLDDT
0
100


A
M (kDa)
165
125
93
72
57
42
31
24
15
8
MBK5214510
A0A2U3LUU8
A0A2U3L5I1
A4J5H3
S4XVP2
S4Y171
A0A6A7KI92
A0A7C1Q7X0
UYP43922
A0A558G9V7


B
A280 (mAU)
150
100
50
0
0
1
2
3
4
5
Retention volume (ml)
MBK5214510
A0A2U3LUU8
A0A2U3L5I1
A4J5H3
S4XVP2
S4Y171
A0A6A7KI92
A0A7C1Q7X0
UYP43922
A0A558G9V7

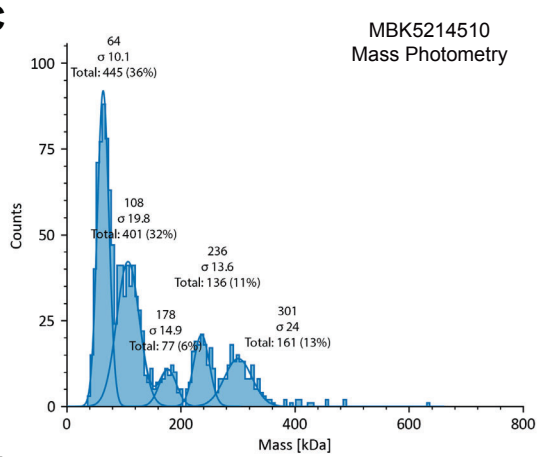
A



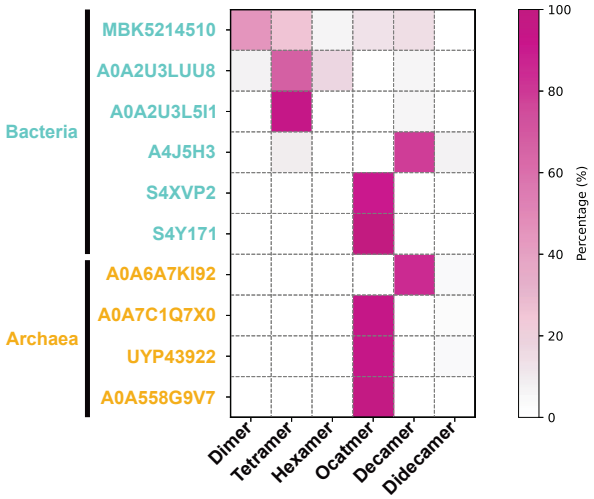
B

Kingdom	Uniprot	MW (kDa)	Uncertainty (%)	Number of Monomers
Bacteria	MBK5214510	205.1	9.6	8
	A0A2U3LUU8	115.8	4.3	4
	A0A2U3L5I1	56.8	9.9	2
	A4J5H3	277.1	1.4	12
	S4XVP2	215.3	2.8	8
	S4Y171	165.3	2.5	7
Archaea	A0A6A7KI92	318.4	10.5	14
	A0A7C1Q7X0	273.9	4.9	10
	UYP43922	235.2	4.8	8
	A0A558G9V7	42.7	27.2	2

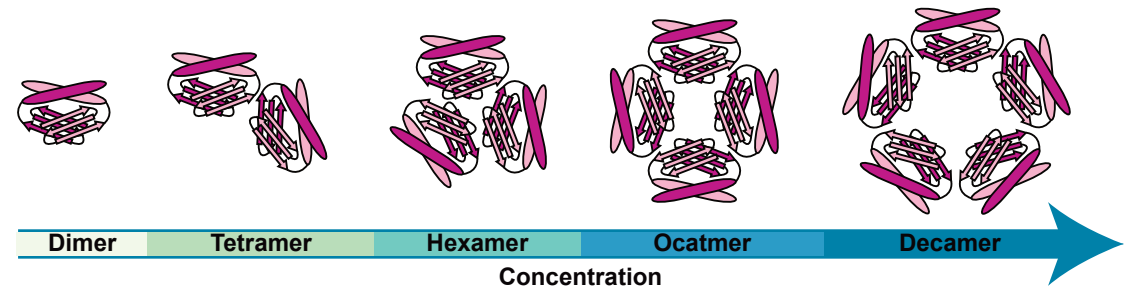
C

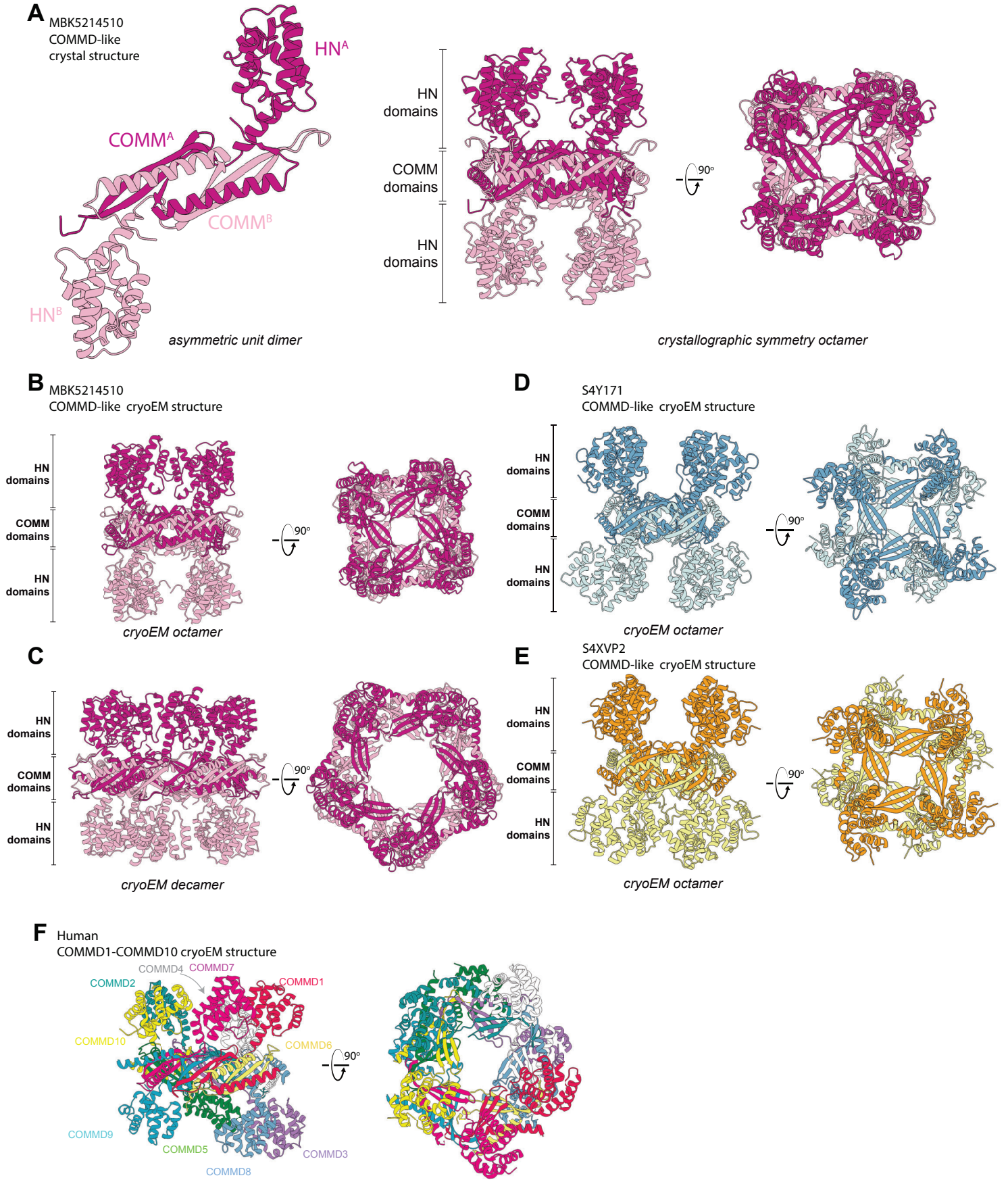


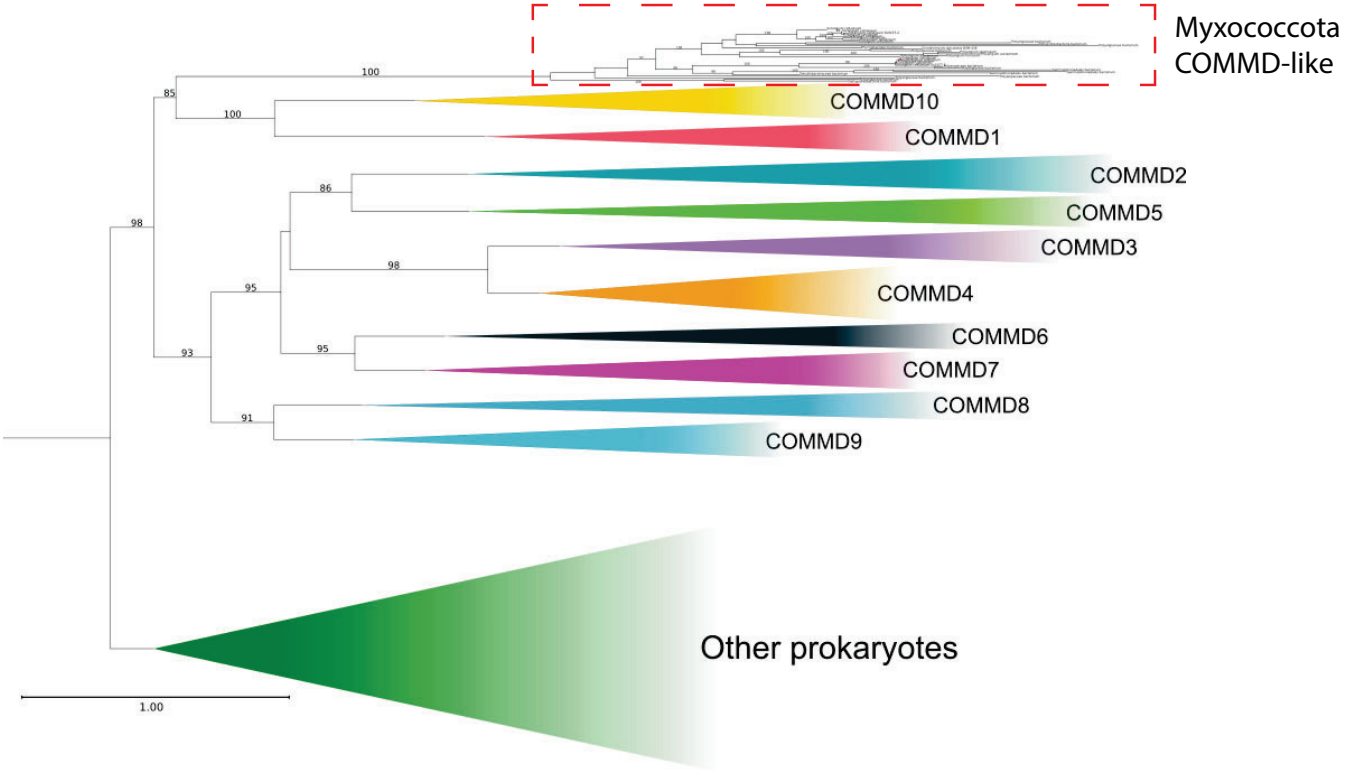
D

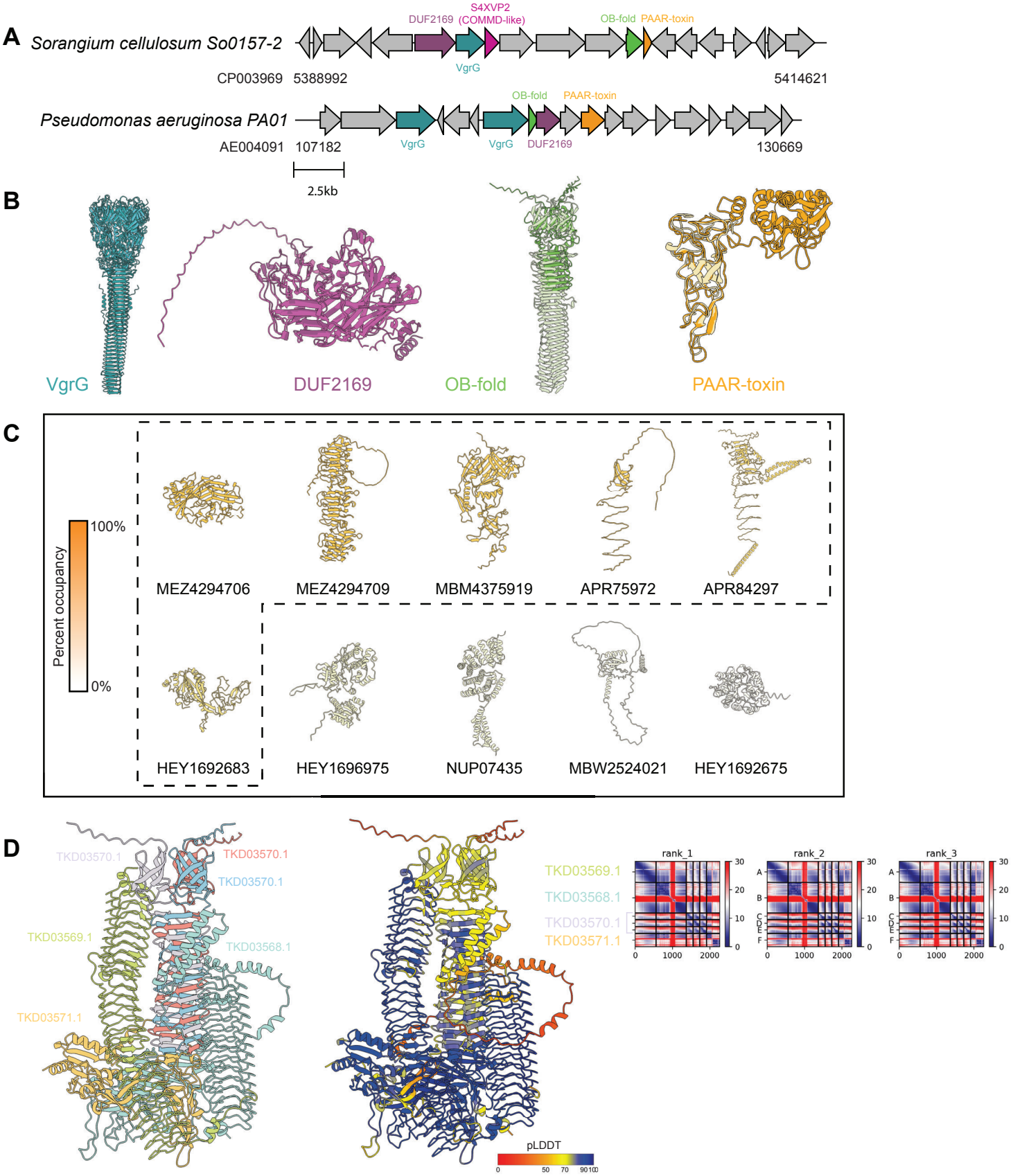


E




A
MBK5214510
COMMD-like
crystal structure
HN^A
COMM^A
COMM^B
HN^B
asymmetric unit dimer
HN domains
COMM domains
HN domains
90°
crystallographic symmetry octamer
B
MBK5214510
COMMD-like cryoEM structure
HN domains
COMM domains
HN domains
90°
cryoEM octamer
C
HN domains
COMM domains
HN domains
90°
cryoEM decamer
D
S4Y171
COMMD-like cryoEM structure
HN domains
COMM domains
HN domains
90°
cryoEM octamer
E
S4XVP2
COMMD-like cryoEM structure
HN domains
COMM domains
HN domains
90°
cryoEM octamer
F
Human
COMMD1-COMMD10 cryoEM structure
COMMD4
COMMD7
COMMD2
COMMD1
COMMD10
COMMD6
90°
COMMD9
COMMD5
COMMD3
COMMD8


Myxococcota
COMMD-like
COMMD10
COMMD1
COMMD2
COMMD5
COMMD3
COMMD4
COMMD6
COMMD7
COMMD8
COMMD9
Other prokaryotes
100
85
100
86
98
98
95
95
93
91
1.00


A
Sorangium cellulosum So0157-2
DUF2169
S4XVP2
(COMMD-like)
OB-fold
PAAR-toxin
VgrG
CP003969
5388992
5414621
Pseudomonas aeruginosa PA01
OB-fold
PAAR-toxin
VgrG
VgrG
DUF2169
AE004091
107182
130669
2.5kb
B
VgrG
DUF2169
OB-fold
PAAR-toxin
C
Percent occupancy
100%
0%
MEZ4294706
MEZ4294709
MBM4375919
APR75972
APR84297
HEY1692683
HEY1696975
NUP07435
MBW2524021
HEY1692675
D
TKD03570.1
TKD03570.1
TKD03570.1
TKD03569.1
TKD03568.1
TKD03571.1
pLDDT
0
50
70
90
100
TKD03569.1
TKD03568.1
TKD03570.1
TKD03571.1
rank_1
rank_2
rank_3
A
B
C
D
E
F
0
1000
2000
30
20
10
0