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[3,7], 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]School of Chemistry, Faculty of Sciences, Biomedical Sciences Building, University of Bristol, Bristol BS8 1TS, UK.
[3] School of Biological Sciences, University of Bristol, Bristol BS8 1TH, UK.
[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 8238 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 high purity (**Fig. 2A**). Interestingly a number of these proteins MBK5214510, A0A2U3LUU8, A4J5H3, UYP43922 and A0A7C1Q7X0 show a second band on Coomassie gels. This band suggests that COMMD-like proteins form a SDS-resistant dimer, and is reminiscent of the human COMMD proteins, which often show a similar gel staining pattern[20].

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. Out of the eight proteins that produced enough protein for assessment by MALLS five form large complexes that consist of at least eight protomers (**Fig 3A; Fig. S4A**). 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. S4B**). 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. S4**). S4XVP2 and S4Y171 form stable octamers, whereas A0A6A7KI92, A0A7C1Q7X0 and UYP43922 seemed to form relatively stable decameric assemblies. Finally, A0A558G9V7 can only be observed in a dimeric state. (**Fig 3C; Fig. S4**). 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. The *myxococcota* (bacterial) homologues S4XVP2 and S4Y171 form homogenous octameric complexes.

One outstanding question is why we did not observe larger complexes for A0A2U3LUU8, A0A2U3L5I1, A0A558G9V7. 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]. In the case of A0A2U3L5I a low resolution cryo-EM dataset was collected that shows 2D classes consistent with an ability to form a homodecameric ring (**Fig. S5**). 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.

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 eight of the selected 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. S6; 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. S6**). 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. S4**) 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. S7**). 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. S7-S8**). 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.

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. 5A-5B**). 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 5A**), indicating that

242 eukaryotes did not inherit COMMD from either of their immediate prokaryotic ancestors.
243 However, robustly establishing the closest prokaryotic clade – that is, the extant relatives of the
244 prokaryotic donor lineage - proved challenging because of the low level of similarity between
245 sequences and substantial phylogenetic uncertainty in both sequence- and structure-based
246 analyses, which were performed under the best-fit site-heterogeneous model (sequence analysis),
247 and using a recently developed substitution model for 3Di characters (structure analysis[33,34]).
248 Intriguingly, in the maximum likelihood sequence-based tree, eukaryotes branch with a clade
249 predominantly comprising Myxococcota, but also including individual sequences from
250 Desulfobacterota, Gemmatimonadota and Planctomycetota (bootstrap support 98%). This is
251 intriguing because several other genes involved in eukaryotic membrane biology, including genes
252 for sterol biosynthesis, appear to have been acquired from this clade[35]. A close and highly-
253 supported relationship between eukaryotic COMMD and Myxococcota was also observed in the
254 structure-based tree (**Fig. S9)**.
255

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

257 Although the prokaryotic COMMD-like proteins clearly share a conserved structure with their
258 eukaryotic relatives, their functions in either Bacteria or Archaea are unknown. In prokaryotes,
259 genes that function in related pathways often tend to cluster together, for example components of
260 hydrogenase complexes[36] and the ancestral ESCRT and ubiquitination machinery in Archaea[37-39].
261 We therefore investigated the genomic context of the COMMD-like genes with a particular
262 emphasis on the *Myxococcota* bacteria. An example genomic region focused on S4XVP2 from
263 *Sorangium cellulosum* So0157-2 is shown in **Fig. 6A**. Interestingly, by investigating the adjacent
264 genes, we noticed similarities between this region and the VgrG1B region of *Pseudomonas*
265 *aeruginosa* PA01[40] (**Fig. 6A**), although we note *P. aeruginosa* PA01 does not have an identified
266 COMMD-like protein in the genome. The VgrG1B region is known to be critical for *P. aeruginosa*
267 PA01 virulence and contains numerous components related to the Type VI secretions system
268 (T6SS) (**Fig. 6B**).

269 To understand if this was an isolated case or common to other COMMD-like proteins from
270 *Myxcoccotta* we next performed syntenic analyses. Several COMMD-like homologues from
271 *Myxococcota* species are available in the NCBI database that have not yet been catalogued in
272 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. S10**), 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 and E. coli

303 In addition to the Myxococcota bacteria, we also noted an expansion of COMMD-like proteins in
304 *Chlamydia* and *Halobacteria*. The exact reason for this expansion is unclear, and a potential caveat
305 is that due to their respective importance in disease and Archaeal evolution respectively, the
306 *Chlamydia* and *Halobacteria* clades have many representative genome sequences compared with
307 other species we identified. It is also worth noting that while the overall structure of the *Chlamydia*
308 proteins conforms to the classical COMMD-fold, containing a helical HN domain and a COMM
309 domain, they appear to function in a unique manner. This is seen in two previously determined
310 crystal structures (PDB: 3Q9D and 4MLK[47]) in which they form a hexameric assembly
311 coordinated by interactions between the HN domain as opposed to the classical β-sheet
312 augmentation seen in the COMMD and COMMD-like structures presented here and elsewhere.
313 Regardless we investigated the genomic context of these proteins to see if there was continuity
314 with *Myxococcota*. This showed that these COMMD-like proteins clearly exist within species level
315 clusters (**Fig S10-12**). In **Table S5** the clusters are summarised for an exemplar COMMD-like
316 protein: O84588 for *Chlamydia trachomatis serovar D (strain ATCC VR-885 / DSM 19411 / UW-*
317 *3/Cx)* and G0HSR7 for *Haloarcula hispanica (strain ATCC 33960 / DSM 4426 / JCM 8911 /*
318 *NBRC 102182 / NCIMB 2187 / VKM B-1755)*. As before a combination of AlphaFold2 and
319 FoldSeek was used to understand the type of genes these cluster with and if any complexes were
320 forming with nearby encoded proteins. In both cases the COMMD-like protein was not predicted
321 by AlphaFold to make direct interactions with any of the proteins within the respective gene
322 cluster.
323
324

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.

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

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

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

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

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

411 In summary, the combination of genome-wide structure prediction, and structural
412 homology searches and phylogenetics have provided a powerful and comprehensive approach for
413 identifying remote protein homologies that are difficult - if not impossible - using sequence-based
414 methods alone. Eukaryotic COMMD proteins have a unique structural organisation that drives
415 heteromeric assembly into stable ring-shaped complexes, and both archaea and bacteria possess
416 ancestral COMMD-like proteins that form analogous complexes with only a single repeated
417 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.

426 **MATERIALS AND METHODS**

427 ***Structure-based homology search***

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

440

441 ***Protein structure predictions***

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

449

450 ***Molecular cloning***

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

457 *et al*[82]. S4XVP2 and S4Y171 produced more homogeneous higher quality protein when expressed
458 with a N-terminal GST tag compared with the C-terminal His tag.
459
460 ***Protein expression and purification***
461 All the proteins were expressed using *Escherichia coli* BL21(DE3) competent cells grown in the
462 presence of kanamycin at 37°C. Protein expression was induced by 1 mM isopropyl β-D-
463 thiogalactopyranoside (IPTG) when cell density OD600 reached 0.8, the temperature was then
464 reduced to 25°C and cultures were incubated overnight. The cells were harvested the following
465 day by centrifugation (Beckman JLA 8.1 rotor) at 6,000 rpm, 4 °C for 10 minutes and resuspended
466 in lysis buffer (500 mM NaCl, 50 mM Tris-HCl (pH 8.0), 5 mM imidazole, 10 % glycerol, 2mM
467 β-mercaptoethanol, benzamidine, and DNAseI).
468 For protein purification, cells were firstly lysed using a Constant systems cell disruptor at
469 35 Kpsi, followed by centrifugation (Beckman JLA 16.250 rotor) at 14,000 rpm, 4°C for 30
470 minutes, after which the soluble His-tagged proteins were contained in the supernatant.
471 Subsequently, TALON resin (Bio-strategy Pty Limited) was added to the supernatant for binding
472 His-tagged proteins and eluted 500 mM NaCl, 500 mM imidazole, 10% glycerol, 2 mM β-
473 mercaptoethanol. Proteins were further purified using pre-equilibrated size exclusion
474 chromatography (SEC) columns in 100 mM NaCl, 20 Tris-HCl (pH 8.0), either a Superdex 200
475 Increase 10/300 GL column (Cytiva) or HiLoad 16/600 Superdex 200 pg column (Cytiva) attached
476 to an AKTA system (GE Healthcare). The purified protein fractions were checked by sodium
477 dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) using Bolt 12 % Bis-Tris gel in
478 1x MES running buffer at 165 V for 45 minutes.
479
480 ***Size exclusion chromatography-Multiangle laser light scattering (SEC-MALLS)***
481 Each protein sample at 1 mg/ml (30 ul) was injected onto the Superose 6 increase 5/150 GL column
482 equilibrated with the buffer containing 100 mM NaCl, 20 mM Tris-HCl (pH 8.0). MALS was
483 performed using a DAWN HELEOS II 10-angle light-scattering detector paired with an Optilab
484 rEX refractive index detector (Wyatt Technology), attached to a Prominence high-performance
485 liquid chromatography (HPLC) (Shimadzu) at a flow rate of 0.25 ml/min.
486
487 ***Single-Particle Mass Photometry***

488 Molecular mass of all the proteins except S4Y171 was measured by a Refeyn OneMP mass
489 photometer (Refeyn Ltd) in the CMM. Initially bovine serum albumin (BSA) from the Sigma-
490 Aldrich was applied as a mass calibration for MBK5214510, given MBK5214510 ability to for
491 concentration dependent oligomers this was used in subsequent measures for calibration. For
492 actual measurement, 10 µl of 100 nM protein samples (200 nM for A0A2U3L5I1, A0A558G9V7)
493 was added to the coverslip. Movie were collected for 180 s. The data was then analysed using
494 Aquire^{MP} software and generated a gaussian distributions to determine the protein mass
495 composition.
496
497 ***Nano-Differential Scanning Fluorimetry.***
498 NanoDSF was conducted using Prometheus NT.48 equipped with backscattering detectors
499 (NanoTemper Technologies, München, Germany) at the Protein Expression Facility (PEF, The
500 University of Queensland). Each protein sample was prepared in triplicate at a concentration of 1
501 mg/ml and loaded into individual NanoDSF grade high sensitivity capillaries (NanoTemper
502 Technologies, München, Germany). The samples were then exposed at 100% excitation power,
503 with a temperature gradient ranging from 20°C to 95°C with a ramping rate of 1°C/min.
504 Tryptophan fluorescence emission was collected at 330 nm and 350 nm using a dual-UV detection,
505 meanwhile, protein aggregation was monitored by backreflection optics. The PR.ThermControl
506 software was used to calculate thermal stability parameters (Tonset, Tm, and Tagg).
507
508 ***Crystallisation and structure determination***
509 Crystallization trials were performed on a Mosquito liquid handling robot (TTP LabTech) in the
510 commercially available SPARSE matrix 96-well hanging-drop plates (Index, LMB, MIDASplus,
511 PEG/Ion, PEGRx and JCSG+) at 20 °C. Prior to sample loading into wells, protein samples (except
512 S4XVP2, S4Y171) were concentrated to ~10 mg/ml (determined by Bradfold). Crystals hits were
513 observed for all proteins across different plates. To scale up the drop and crystal size, several
514 conditions were optimized in a 24-well hanging drop vapor diffusion (Hampton) plate covered by
515 glass cover slips. On each glass coverslip, three drops were prepared with protein and reservoir
516 solution mixed in ratios of 1:1, 2:1 and 3:1 (protein:reservoir solution). Glycerol (10 %) was added
517 into the mother liquor as cryo-protectant before flash-freezing in liquid nitrogen. The crystals used
518 for MBK5214510 (15mg/ml) structure determination were grown in MIDASplus C9 (10 % v/v

519 DMSO, 0.1M NaCl and 25% PE5/4 which in this case served as the cryo-protectant) a final
520 concentration of 15mg/ml.
521 X-ray diffraction data were collected at the Australian synchrotron on the MX2 beamline.
522 The data was indexed and integrated with XDS[83] and scaled with AIMLESS withing the CCP4
523 suite[84]. The initial structure was resolved using molecular replacement in Phaser[85] with the
524 AlphaFold2 structure as an initial template. Further refinements of the structure were performed
525 with PHENIX[86] and the chimera plugin ISOLDE[87].
526
527 ***Cryo-EM grid preparation and data collection***
528 Freshly glow-discharged Quantifoil 1.2/1.3 Copper 200 mesh grids were used for cryo-EM data
529 collection. For MBK5214510 and A0A6A7KI92 a final concentration of 10 mg/ml in 100 mM
530 NaCl, 20 Tris-HCl (pH 8.0) was loaded onto grids. Grid vitrification was performed in liquid
531 ethane using a Vitrobot Mark IV plunge freezer (Thermo Fisher), using a 3 µL sample application,
532 95% humidity, 4°C, minus 1 blot force and 2.5 s blot time. After freezing, grids were stored
533 immediately under liquid nitrogen until image acquisition.
534 Cryo-EM data for MBK5214510 was acquired using a JOEL CRYOARM 300 (JOEL)
535 microscope operated at an acceleration voltage of 300 kV, equipped with an in column Ω energy
536 filter (20 eV slit width) and a Gatan K3 direct electron detector. A total of 1,905 movies, at 60,000
537 x magnification, with a pixel size of 0.82 Å per pixel and a total exposure dose of 40 e/Å2, within
538 a target defocus range from −2.5 to −0.5 µm. A0A6A7KI92 movies (769 movies) were collected
539 on JOEL CRYOARM200 at 200 kV under 100,000 x magnification (corresponding to calibrated
540 pixel sizes of 0.52 Å/pixel). This microscope included a cold-field emission gun (FEG), an in-
541 column Ω energy filter and a Gatan K2 direct electron detector. As before the total exposure dose
542 was maintained at 40 e/Å2, with the defocus range adjusted between −2.5 and −0.5 µm.
543
544 ***Cryo-EM data processing***
545 Data processing of MBK5214510 and A0A6A7KI92 was performed using the University of
546 Queensland's remote web access desktops (https://github.com/UQ-RCC/hpc-
547 docs/blob/main/guides/OnDemand-Guide.md) in CryoSPARC V4.6.1[88]. For MBK5214510,
548 1,905 movies were imported and motion corrected using Patch Motion Correction, followed by
549 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 S6 and 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.

Sequence-based phylogenetic analysis

Prokaryotic sequences from above were aligned using MAFFT v7.505[89] (--linsi), this alignment was used to infer maximum likelihood trees using IQTREE 2.2.6[90], using 20 profile mixture models (LG+C20)[91] with empirical amino acid frequencies derived from the data (+F), four gamma categories (-G)[92], with 10,000 ultrafast bootstraps[93]. Additional trees were inferred under the same settings, but with alternative constraints built using TreeViewer[94], 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[95] test under the same model, with 100,000 RELL replicates[96]. 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.

581

582 ***Structure-based phylogenetic analysis.***

583 The proteins used in the aforementioned sequence analysis were downloaded from the AlphaFold
584 data base (where available) or predicted used AlphaFold2 and structurally aligned using
585 FOLDMASON[97]. Phylogenetic trees were inferred from the 3Di alignment with IQ-TREE v3.0.1[83]
586 using the substitution rate matrix derived from Foldseek 3Di substitution cost matrices[35] with
587 10,000 ultrafast bootstraps[86]. Phylogenies were then visualised using Figtree
588 (http://tree.bio.ed.ac.uk/software/figtree/).

589

590 ***Genomic context analysis***

591 COMMD-like proteins were identified and enterz was used to download the region 10,000bp
592 upstream and downstream of that reference point. CLINKER[98] was then used to align these
593 genomic regions based on sequence identity. In the case of Myxcoccota sequences they were
594 extracted from the .gbk file and the structure of those sequences was predicted using AlphaFold2
595 as described above.

596

597

598

599

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[99]. 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 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.

DATA AVAILABILITY

All data related to this study is available from the corresponding authors on request. PDB files 9Q6A, 9Q6C, 9Q6D. AlphaFold2 models generated in this work are available at ModelArchive: 10.5452/ma-yfp9l

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

639 **FIGURE LEGENDS**

640 **Figure 1. Discovery of ancestral prokaryotic COMMD-like proteins based on structural**
641 **homology.**

642 (**A**) Workflow to identify and validate prokaryotic ancestors of Commander subunits with
643 FoldSeek[29]. Eukaryotic protein structures of Commander subunits were used as search templates
644 in Foldseek, and searches were then performed iteratively using identified prokaryotic proteins
645 until the search-space was exhausted. Selected prokaryotic proteins were then verified using
646 biophysical and structural approaches, while phylogenetic analyses were performed to uncover
647 evolutionary and potential functional relationships. (**B**) From systematic searches, we did not
648 identify any prokaryotic proteins with structural similarity to the majority of the Commander
649 subunits, but clear structural homologues of the COMMD proteins were found in both archaeal
650 and bacterial species. (**C**) AlphaFold2 prediction of ten selected bacterial and archaeal COMMD-
651 like proteins (coloured by pLDDT score) show a highly conserved architecture when compared to
652 human COMMD9 including HN and COMM domains. Accession numbers, Phyla and species are
653 shown to the right of each image. For the sake of precise identification proteins have been labelled
654 by either a Uniprot ID or Uniparc ID if Uniprot ID was unavailable.

655

656 **Figure 2. Expression and purification of bacterial and archaeal COMMD-like proteins.**

657 **(A)** Coomassie-stained SDS-PAGE of purified COMMD-like proteins. The ten COMMD-like
658 proteins shown in **Fig. 1C** were all expressed in *E. coli* and purified by His-tag affinity
659 chromatography followed by gel filtration. For several of the COMMD-like proteins a second band
660 can be observed reflecting the obligate dimeric nature of this protein fold. (**B)** The gel filtration
661 elution profiles of the purified COMMD-like proteins shows these samples to be mostly
662 monodisperse but typically eluting at an earlier time point indicating that they forming oligomers
663 larger than their expected monomeric molecular weights.

664

665 **Figure 3. COMMD-like proteins form stable concentration dependent oligomers.**

666 **(A)** Multiangle laser light scattering (MALLS) of purified MBK5214510. The protein can form an
667 oligomeric structure, however the plots suggest the sample is not fully monodisperse. MALLS
668 experiments were performed for the majority of the purified COMMD-like proteins and the data
669 is shown in **Fig. S4A**. Proteins were injected at a concentration of ~40 μM based on monomeric

670 mass. (**B**) Table summarising the observed molecular weights by MALLS, uncertainty of the
671 measure as well as the approximate oligomeric state based on the theoretical monomeric molecular
672 weight. (**C**) Mass photometry measurement of purified MBK5214510 shows evidence of dimeric,
673 tetrameric, hexameric, octameric and decameric species. Mass photometry was performed for the
674 other COMMD-like proteins with data shown in **Fig. S4B**. (**D**) A heatmap summarising the
675 percentage of each oligomeric species in the samples measured by mass photometry. Broadly, the
676 bacterial proteins generally exist as a mixture of oligomeric species while the archaeal proteins
677 appear more stable and monodisperse. S4XVP2 and S4Y171 form stable octamers and
678 A0A558G9V7 appears to only form dimers. Proteins were measured at a final 100-200 nM. (**D**)
679 Proposed concentration dependent assembly of bacterial COMMD-like proteins.
680

681 **Figure 4. Structures of COMMD-like proteins reveal ring-shaped homo-oligomers that**
682 **resemble the hetero-decamaric human COMMD complex.**
683 **(A)** Crystal structure of MBK5214510 from *Flavobacteriaceae bacterium*. The asymmetric unit
684 contains a dimer of MBK5214510 similar to the dimers of human COMMD proteins. The crystal
685 lattice reveals the presence of an MBK5214510 octamer along the 4-fold axis. (**B**) and **(C)** Cryo-
686 EM structures of MBK5214510 octamers and decamers. These different oligomeric species are
687 present in the same purified sample, indicating a degree of structural plasticity in the stoichiometry
688 of the ring-shaped assemblies. In both cases a star shaped structure is observed with C-terminal
689 COMM domains forming a central platform and the HN domains lying peripherally on each side.
690 (**D**) The Cryo-EM structure of the human COMMD1-COMMD10 hetero-decameric complex[18]
691 (PDB ID 8F2U) is shown for comparison to the bacterial homo-oligomeric assemblies.
692

693 **Figure 5. Evolutionary origins of the eukaryotic COMMD proteins.**
694 The phylogenetic tree of the prokaryotic COMMD-like proteins and metazoan COMMD proteins
695 based on sequence comparisons revealed the Myxococcota clade of proteins was nested within the
696 metazoan clade with high support (bootstrap 98%), while Asgard Archaea were considerably more
697 distantly related.
698

699 **Figure 6. Genomic context of Myxococcota COMMD like proteins suggests a relationship**
700 **with injection machinery.**

701 (**A**) Gene maps of *Sorangium cellulosum* So0157-2 and *Pseudomonas aeruginosa* PA01 focused
702 on the DUF2169 protein a part of the *P. aeruginosa* Vgr1B virulence cluster. Note *P. aeruginosa*
703 does not have a COMMD-like protein. Coloured genes highlight similar structural folds between
704 the two species. (**B**) Highlighted genes from each species in panel (**A**) were predicted by
705 AlphaFold2 and structurally aligned. Lighter colours are used for *S. cellulosum*, to distinguish
706 them from the *P. aeruginosa* proteins. These proteins represent folds that are commonly associated
707 with bacterial injection machinery. (**C**) To better understand the genomic context of the COMMD-
708 like genes across the Myxococcota clade we used AlphaFold to predict structures of the encoded
709 proteins from genes ±10,000 bp across 55 species. These proteins were clustered using the "easy-
710 cluster" module available in the FoldSeek software. The structures shown are representatives of
711 structures that appear in at least 20% Myxococcota genomes and are coloured based on the
712 percentage of genomes in which they are found. The dotted box indicates proteins with folds
713 related to known injection machinery proteins. (**D**) Iterative AlphaFold predictions were
714 performed to assess potential interactions among the proteins that appeared to be related to
715 bacterial injection machinery complexes. These consistently revealed a complex that was predicted
716 with high confidence that incorporates four genes assembled into a hetero-hexameric complex.
717 The predicted complex shown is an example from *Polyangium fumosum* composed of a
718 homotrimer of TKD03570 and single copies each of TKD03571, TKD03569, TKD03568. The
719 complex shares strong similarities to components of the Type VI secretion system baseplate
720 machinery[46].
721
722
723

724 **SUPPLEMENTARY FIGURES**

725 **Figure S1. AlphaFold2 predicted structures of COMMD-like proteins from Myxococcota.**

726 The structures of S4Y171 and S4XVP2 from *Sorangium cellulosum* were predicted using
727 AlphaFold2 in various homooligomeric forms, with the results shown for dimer, octamer and
728 decamer predictions. Proteins care coloured according to pLDDT, with PAE plots shown below.

729

730 **Figure S2. AlphaFold2 predicted structures of COMMD-like proteins from various bacterial**
731 **species.**

732 The structures of A4J5H3 (*Desulfotomaculum reducens*), A0A2U3L5I1 (*Candidatus*
733 *Sulfopaludibacter*), A0A2U3LUU8 (*Candidatus Sulfopaludibacter*), MBK5214510
734 (*Flavobacteriaceae bacterium*) were predicted using AlphaFold2 in various homooligomeric
735 forms, with the results shown for dimer, octamer and decamer predictions. Proteins care coloured
736 according to pLDDT, with PAE plots shown below.

737

738 **Figure S3. AlphaFold2 predicted structures of COMMD-like proteins from various archaeal**
739 **species.**

740 The structures of UYP43922 (*Candidatus Lokiarchaeum ossiferum*), A0A7C1Q7X0
741 (*Unclassified*), A0A6A7KI92 (*Thermococcus sp. 101 C5*), A0A558G9V7 (*Haloferax volcanii*)
742 were predicted using AlphaFold2 in various homooligomeric forms, with the results shown for
743 dimer, octamer and decamer predictions. Proteins care coloured according to pLDDT, with PAE
744 plots shown below.

745

746 **Figure S4. Biophysical characterisation of archaeal and bacterial COMMD-like proteins**

747 (**A**) MALLS data and (**B**) Mass photomometry data for bacterial and archaeal COMMD-like
748 proteins. (**C**) NanoDSF shows two groupings of COMMD-like proteins with respect to thermal
749 stability with a TM^{50} of ~60°C and 40°C.

750

751 **Figure S5. Cryo-EM of A0A2U3L5I from Candidatus Sulfopaludibacter.**

752 2D class averages indicating a decameric complex of COMMD-like protein A0A2U3L5I from
753 Candidatus Sulfopaludibacter.

754

Figure S6. Cryo-EM workflow for structure determination of MBK5214510.

(A) Cryo-EM data processing workflow using CryoSPARC software. The final 3D reconstruction map was used for model building and refinement. (B) 3D Fourier Shell Correlation of the final reconstructions reporting global resolution, with the FSC at 0.143 denoted. The model-to-map FSC plot from Phenix real space refinement is overlaid. (C) Final reconstruction of octameric structure filtered and colored by local resolution, as estimated in CryoSPARC. (D) Final reconstruction of decameric structure filtered and colored by local resolution, as estimated in CryoSPARC.

Figure S7. Comparison of the experimental MBK5214510 COMMD-like structures with the AlphaFold predictions.

Left panels show AlphaFold2-predicted structures of MBK5214510 COMMD-like protein from *Flavobacteriaceae bacterium* with the lower images showing just the central ring of COMM domains that form the core of the homooligomeric rings. Middle panels show the experimental structures of MBK5214510 octamers and decamers determined by X-ray crystallography and cryo-EM and right-hand panels show the overlay of the predicted and experimental structures.

Figure S8. Comparison of the experimental MBK5214510 COMMD-like homodecameric structure with the human COMMD heterodecameric complex.

Left panels show either the AlphaFold2-predicted or cryo-EM homodecameric structure of MBK5214510 COMMD-like protein from *Flavobacteriaceae bacterium* with the lower images showing just the central ring of COMM domains that form the core of the homooligomeric rings. Middle panels show either the AlphaFold2-predicted or cryo-EM heterodecameric structure the human COMMD1-COMMD10 complex[18] (PDB ID 8F2U) and right-hand panels show the overlay of the bacterial and human structures.

Figure S9. Structure-based phylogenetics.

(A) 3Di phylogeny showing the relationship between archaeal (blue) and bacterial (red) COMMD-like proteins and eukaryotic COMMD1-COMMD10 proteins (green). (B) Example structural alignment of the AlphaFold2 predictions of COMMD10 with two Mycococcota COMMD-like proteins S4Y171 and S4XVP2 from *Sorangium cellulosum*.

Figure S10. Gene clusters surrounding COMMD-like genes in Myxococotta.

Analysis of the genomic surroundings of COMMD-like genes from Myxococotta species using CLINKER.

Figure S11. Gene clusters surrounding COMMD-like genes in Chlamydia.

Analysis of the genomic surroundings of COMMD-like genes from Chlamydial species using CLINKER.

Figure S12. Gene clusters surrounding COMMD-like genes in Haloarchaea.

Analysis of the genomic surroundings of COMMD-like genes from Haloarchaea species using CLINKER.

SUPPLEMENTARY DATA

Supplementary Data S1. Accession numbers and sequences for the prokaryotic COMMD-like proteins identified in this study.

Supplementary Data S2. Summary of the results from purification and crystallisation screens of the ten selected archaeal and bacterial COMMD-like proteins.

Supplementary Table S1. Crystallographic data collection and refinement statistics.

Supplementary Table S2. Cryo-EM data collection and refinement statistics.

Supplementary Table S3. Key Resources Table.

Supplementary Table S4. *Myxococcota* genomic clustering.

Supplementary Table S5. *Chlymidia* and *Haloarchaea* example protein clusters.

815 **Supplementary Table S1.** Crystallographic data collection and refinement statistics.

816

	MBK5214510
Data Collection	
Space group	I222
Unit cell dimensions (a,b,c) (Å)	91.3, 96.0, 115.6
Unit cell angles (α,β,γ) (º)	90 90 90
Wavelength (Å)	0.95373
Resolution range (Å)	45.67 - 2.582 (2.66 - 2.58)
Total reflections	106202
Unique reflections	16103 (1127)
Multiplicity	6.6 (6.4)
Completeness (%)	98.69 (84.74)
Mean I/σ(I)	14.8 (2.8)
R-merge	0.022 (0.146)
R-meas	0.0313 (0.207)
R-pim	0.022 (0.146)
CC(1/2)	0.999 (0.919)
Refinement	
R-work	0.2419 (0.2919)
R-free	0.2895 (0.3782)
Number of non-hydrogen atoms	3573
Number macromolecule atoms	3534
Number water atoms	39
RMS bonds (Å)	0.006
RMS angles (º)	1.17
Ramachandran favored (%)	93.44
Ramachandran allowed (%)	4.92
Ramachandran outliers (%)	1.64
Clashscore	10.97
Overall Molprobity score	
Average B-factor	71.52
PDB ID	9Q6A

817

818 **a.** Statistics for the highest-resolution shell are shown in parentheses.

819

820 **Supplementary Table S2.** Cryo-EM data collection and refinement statistics.

821

	MBK5214510 octamer	MBK5214510 decamer
Data collection and processing		
Magnification	100,000	100,000
Voltage (kV)	300	300
Electron exposure ($e^{-}/Å^2$) Defocus range (μm)	40 $e^{-}/Å^2$ -2.5 μm	40 $e^{-}/Å^2$ -2.5 μm
Pixel size (Å)	0.82	0.82
Symmetry imposed	D4	D5
Programme used	CryoSPARC	CryoSPARC
Initial micrographs collected (no.)	1,905	1,905
Final micrographs processed (no.)	1,331	1,331
Initial particle images (no.)	253,056	253,056
Final particle images (no.)	72,675	48,201
Map resolution (Å)	4.22	4.77
FSC threshold	0.143	0.143
Refinement		
Initial model used	Colabfold	Colabfold
Model resolution range (0/0.143/0.5) (Å)	4.2/4.3/7.7	4.6/4.7/7.6
Model composition		
Non-hydrogen atoms	14,368	17,960
Protein residues	1760	2200
R.m.s. deviations		
Bond lengths (Å)	0.008	0.002
Bond angles (°)	0.861	0.548
Validation		
MolProbity score	2.17	1.86
Clashscore	19.74	12.81
Ramachandran plot		
Favoured (%)	94.5	96.33
Allowed (%)	5.5	3.67
Outliers (%)	0	0
EMDB ID	EMD-72266	EMD-72267
PDB ID	9Q6C	9Q6D

822

823

824 **Supplementary Table S3.** Key Resources Table.

825

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
E. coli DH5α	Invitrogen	Cat# 18265017
BL21 (DE3)	New England Biolabs	Cat# C2527H
DH10EMBacY	ThermoFisher	Cat# 10361012
NEB 5-alpha	New England Biolabs	Cat# C2987H
Chemicals and peptides		
Benzamidine hydrochloride hydrate	Sigma Aldrich	Cat# B6506
Deoxyribonuclease I (DNase I)	Sigma Aldrich	Cat# DN25
Talon® resin	Clontech	Cat# 635503
Glutathione Sepharose 4B	GE Healthcare	Cat# GEHE17-0756-0
Isopropyl β-D-1-thiogalactopyranoside	Bioline	Cat# BIO-37036
Triton-X100	Sigma Aldrich	Cat# X100-500ML
Sf-900 II SFM media	ThermoFisher Scientific	Cat# 10902088
Graphene oxide	Sigma-Aldrich	Cat# 763705
Critical commercial assays		
QIAprep Spin Miniprep Kit	QIAGEN	Cat# 27106
NEBuilder HiFi DNA Assembly System (NEB)	New England Biolabs	Cat# E5520S
TMT seven-plex	ThermoFisher	Cat# 90061
Recombinant DNA		
pET-28(a)	Sigma-Aldrich	Cat# 69864
LM627 vector	Addgene	191661
Deposited Data		
MBK5214510 (crystal structure)	RCSB Protein Data Bank (this study)	PDB ID: 9Q6A
MBK5214510 octamer (cryo-EM structure; RELION map)	RCSB Protein Data Bank (this study)	PDB ID: 9Q6C EMDB ID: EMD-72266
MBK5214510 decamer (cryo-EM structure; CryoSPARC map)	RCSB Protein Data Bank (this study)	PDB ID: 9Q6D EMDB ID: EMD-72267
AlphaFold2 models of Bacterial and Archael COMMD-like proteins	ModelArchive (this study)	10.5452/ma-yfp9l
Software and Algorithms		
XDS	[100]	http://xds.mpimf-heidelberg.mpg.de/
AIMLESS	[101]	http://www.ccp4.ac.uk/html/aimless.html
Phaser	[102]	http://www.phaser.cimr.cam.ac.uk/index.php/Phaser_Crystallographic_Software

Phenix	[103]	https://www.phenix-online.org/
Coot	[104]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/
ChimeraX	[105]	https://www.rbvi.ucsf.edu/chimerax/
iSOLDE	[87]	https://isolde.cimr.cam.ac.uk
RELION 4.0	[106]	https://relion.readthedocs.io/en/release-4.0/
Topaz	[107]	https://cb.csail.mit.edu/cb/topaz/
CryoSPARC V3.3.1	[108]	https://cryosparc.com
AlphaFold2 Multimer	[80,109]	https://github.com/deepmind/alphafold
AlphaFold3		
ColabFold v1.5.5 implementation of AlphaFold2-PTM	[30,79]	https://github.com/sokrypton/ColabFold
ColabFold v1.5.5 implementation of AlphaFold2-Multimer	[79,80]	https://github.com/sokrypton/ColabFold
FoldSeek version c27a629a0a345c288c18d2a0975a3dc7bb143798	[29]	https://github.com/steineggerlab/foldseek
MAFFT v7.505	[89]	https://mafft.cbrc.jp/alignment/server/index.html
IQTREE 2.2.6	[90]	https://iqtree.github.io
FOLDMASON	[97]	https://github.com/steineggerlab/foldmason
Figtree v1.4.4		http://tree.bio.ed.ac.uk/software/figtree/
CLINKER	[98]	https://github.com/gamcil/clinker
Other		
Superose6 Increase10/300 GL column	Cytiva	Cat# 29091596
HiLoad™ Superdex75 PG column	Cytiva	Cat# 28989333
MonoQ 5/50 GL column	Cytiva	Cat# 17516601
C-Flat 1.2/1.3 grids	C-Flat	N/A
Quantifoil 1.2/1.3 Copper 300 mesh grids	Quantifoil	N/A

827 **Supplementary Table S4.** *Myxococcota* genomic clustering

828

CDS	Closest PDB	Description
APR75977.1	8F2U[18]	Human Decameric COMMD ring
MEZ4294706.1		
MEZ4294709.1		
MBM4375919.1	6H3N[110]	VgrG1 of *Pseudomonas aeruginosa PA01*
APR75972.1	6RBK[111]	anti-feeding prophage (AFP) baseplate
APR84297.1		
HEY1692683.1	4JIW[112]/7RPQ[113]	PAAR(N-terminus)/ctpa protease (C-term)
HEY1696975.1	4EQM[42]	Serine-Threonine kinase
NUP07431.1	6KON[43]/2PL2[44]	Component of *Mycobacterium tuberculosis* initial transcription complex/ tetratricopeptide repeat
MBM4375922.1	2M2K[114]	HasB
UCE34441.1	3KRU[45]	Old Yellow Enzyme

829

830 **Supplementary Table S5.** *Chlaymidia* and *Haloarchaea* example protein clusters

831

CDS	Uniprot	Closest PDB	Description
Chlamydia trachomatis serovar D (strain ATCC VR-885 / DSM 19411 / UW-3/Cx)			
AAC68178.1	O84580	2VGX[115]	T3S Chaperon
AAC68179.1	O84581		
AAC68180.1	O84582		
AAC68181.1	O84583		
AAC68182.1	O84584	5I20[116]	DMT transporter
AAC68183.2	O84585	1QF6[117]	THREONYL-TRNA SYNTHETASE
AAC68184.1	O84586	6IUD[118]	DNA-partition
AAC68185.1	O84587		
AAC68186.1	O84588	3Q9D	COMMD-Like
AAC68187.1	O84589	2A4M[119]	Tyrptophan synthase
AAC68188.1	O84590	6O8E[120]	DNA repair/URVB
AAC68189.1	O84591	3ZLF[121]	Enolase
AAC68190.1	O84592	6MAB[122]	Periplasmic environment sensor
AAC68191.1	O84593	3W43[123]/5LTV[124]	Stress signaling/receptor
Haloarcula hispanica (strain ATCC 33960 / DSM 4426 / JCM 8911 / NBRC 102182 / NCIMB 2187 / VKM B-1755)			
AEM58436.1	G0HSR0		
AEM58437.1	G0HSR1	2CY1[125]	Transcription termination factor
AEM58438.1	G0HSR2	7OQ4[126]	RPO1C
AEM58439.1	G0HSR3	7OQ4[126]	RPO1N
AEM58440.1	G0HSR4	7OQ4[126]	RPO2
AEM58441.1	G0HSR5	7OQ4[126]	RPO2
AEM58442.1	G0HSR6	7OQ4[126]	RPO5
AEM58443.1	G0HSR7	3Q9D	COMMD-LIKE
AEM58444.1	G0HSR8	7OYK[127]	Transcription regulator

AEM58445.1	G0HSR9	6XS5[128]	VPS29
AEM58446.1	G0HSS0	7ULE[129]	F420-gamma glutamyl ligase
AEM58447.1	G0HSS1	1Z69[130]	methylenetetrahydromethanopterin reductase
AEM58448.1	G0HSS2		
AEM58449.1	G0HSS3	8SBG[131]	tryptophanases
AEM58450.1	G0HSS4	6F5Z[132]	methyltransferase
AEM58451.1	G0HSS5		
AEM58452.1	G0HSS6	3NBX[133]	Regulatory ATPase variant A - lysine decarboxylase
AEM58453.1	G0HSS7		
AEM58454.1	G0HSS8	6G49[134]	

835 **REFERENCES**

836 1. Schlacht, A., Herman, E.K., Klute, M.J., Field, M.C. & Dacks, J.B. Missing pieces of an ancient
837 puzzle: evolution of the eukaryotic membrane-trafficking system. *Cold Spring Harb Perspect Biol*
838 **6**, a016048 (2014).

839 2. Vosseberg, J. et al. The emerging view on the origin and early evolution of eukaryotic cells. *Nature*
840 **633**, 295-305 (2024).

841 3. Donoghue, P.C.J. et al. Defining eukaryotes to dissect eukaryogenesis. *Curr Biol* **33**, R919-r929
842 (2023).

843 4. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*
844 **521**, 173-179 (2015).

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

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

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

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

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

856 10. Akıl, C. & Robinson, R.C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature*
857 **562**, 439-443 (2018).

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

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

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

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

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

868 16. Laulumaa, S., Kumpula, E.-P., Huiskonen, J.T. & Varjosalo, M. Structure and interactions of the
869 endogenous human Commander complex. *Nature Structural & Molecular Biology* **31**, 925-938
870 (2024).

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

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

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

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

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

881 22. Maine, G.N. & Burstein, E. COMMD proteins: COMMing to the scene. *Cell Mol Life Sci* **64**, 1997-
882 2005 (2007).

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

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

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

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

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

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

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

897 30. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589
898 (2021).

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

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

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

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

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

910 36. Greening, C. et al. Minimal and hybrid hydrogenases are active from archaea. *Cell* **187**, 3357-
911 3372.e19 (2024).

912 37. Lu, Z., Zhang, S., Liu, Y., Xia, R. & Li, M. Origin of eukaryotic-like Vps23 shapes an ancient
913 functional interplay between ESCRT and ubiquitin system in Asgard archaea. *Cell Rep* **43**, 113781
914 (2024).

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

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

919 40. Pissaridou, P. et al. The Pseudomonas aeruginosa T6SS-VgrG1b spike is topped by a PAAR protein
920 eliciting DNA damage to bacterial competitors. *Proc Natl Acad Sci U S A* **115**, 12519-12524
921 (2018).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

965 61. Han, W. et al. Structural basis of plp2-mediated cytoskeletal protein folding by TRiC/CCT. *Sci Adv*
966 **9**, eade1207 (2023).

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

969 63. Montemayor, E.J. et al. Molecular basis for the distinct cellular functions of the Lsm1-7 and Lsm2-
970 8 complexes. *RNA* **26**, 1400-1413 (2020).

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

973 65. Wilusz, C.J. & Wilusz, J. Eukaryotic Lsm proteins: lessons from bacteria. *Nat Struct Mol Biol* **12**,
974 1031-6 (2005).

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

977 67. Yuan, Z. et al. DNA unwinding mechanism of a eukaryotic replicative CMG helicase. *Nat Commun*
978 **11**, 688 (2020).

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

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

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

985 71. Archibald, J.M., Logsdon, J.M., Jr. & Doolittle, W.F. Origin and evolution of eukaryotic
986 chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol Biol Evol* **17**, 1456-
987 66 (2000).

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

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

992 74. Moreira, D. & López-García, P. Symbiosis Between Methanogenic Archaea and δ-Proteobacteria
993 as the Origin of Eukaryotes: The Syntrophic Hypothesis. *Journal of Molecular Evolution* **47**, 517-
994 530 (1998).

995 75. Souza, D.P. et al. Evolutionarily conserved principles of ESCRT-III-mediated membrane
996 remodelling revealed by a two-subunit Asgard archaeal system. *bioRxiv*, 2024.07.01.601590
997 (2024).

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

1000 77. Bernabeu, M., Manzano-Morales, S., Marcet-Houben, M. & Gabaldón, T. Diverse ancestries reveal
1001 complex symbiotic interactions during eukaryogenesis. *bioRxiv*, 2024.10.14.618062 (2024).

1002 78. Stairs, C.W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution.
1003 *Sci Adv* **6**, eabb7258 (2020).

1004 79. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682
1005 (2022).

1006 80. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034
1007 (2022).

1008 81. Meng, E.C. et al. UCSF ChimeraX: Tools for structure building and analysis. *Protein Science* **32**,
1009 e4792 (2023).

1010 82. Wicky, B.I.M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56-61 (2022).

1011 83. Kabsch, W. XDS. *Acta Crystallographica Section D* **66**, 125-132 (2010).

1012 84. Agirre, J. et al. The CCP4 suite: integrative software for macromolecular crystallography. *Acta*
1013 *Crystallographica Section D* **79**, 449-461 (2023).

1014 85. McCoy, A.J. et al. Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658-
1015 674 (2007).

1016 86. Adams, P.D. et al. The Phenix software for automated determination of macromolecular structures.
1017 *Methods* **55**, 94-106 (2011).

1018 87. Croll, T.I. ISOLDE: a physically realistic environment for model building into low-resolution
1019 electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519-530 (2018).

1020 88. Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. cryoSPARC: algorithms for rapid
1021 unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290-296 (2017).

1022 89. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7:
1023 improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).

1024 90. Minh, B.Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
1025 the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).

1026 91. Si Quang, L., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic
1027 reconstruction. *Bioinformatics* **24**, 2317-2323 (2008).

1028 92. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates
1029 over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306-314 (1994).

1030 93. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. UFBoot2: Improving the
1031 Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522 (2018).

1032 94. Bianchini, G. & Sánchez-Baracaldo, P. TreeViewer: Flexible, modular software to visualise and
1033 manipulate phylogenetic trees. *Ecology and Evolution* **14**, e10873 (2024).

1034 95. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492-
1035 508 (2002).

1036 96. Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny and
1037 the origin of chloroplasts. *Journal of Molecular Evolution* **31**, 151-160 (1990).

1038 97. Gilchrist, C.L.M., Mirdita, M. & Steinegger, M. Multiple Protein Structure Alignment at Scale with
1039 FoldMason. *bioRxiv*, 2024.08.01.606130 (2024).

1040 98. Gilchrist, C.L.M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene
1041 cluster comparison figures. *Bioinformatics* **37**, 2473-2475 (2021).

1042 99. The University of Queensland Research Computing Centre. *Bunya supercomputer* (2024).

1043 100. Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-32 (2010).

1044 101. Evans, P.R. & Murshudov, G.N. How good are my data and what is the resolution? *Acta*
1045 *Crystallogr D Biol Crystallogr* **69**, 1204-14 (2013).

1046 102. McCoy, A.J. et al. Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).

1047 103. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure
1048 solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-21 (2010).

1049 104. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D*
1050 *Biol Crystallogr* **60**, 2126-32 (2004).

1051 105. Pettersen, E.F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and
1052 developers. *Protein Sci* **30**, 70-82 (2021).

1053 106. Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S.H.W. New tools for automated cryo-
1054 EM single-particle analysis in RELION-4.0. *Biochem J* **478**, 4169-4185 (2021).

1055 107. Bepler, T. et al. Positive-unlabeled convolutional neural networks for particle picking in cryo-
1056 electron micrographs. *Nat Methods* **16**, 1153-1160 (2019).

1057 108. Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. cryoSPARC: algorithms for rapid
1058 unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296 (2017).

1059 109. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589
1060 (2021).

1061 110. Quentin, D. et al. Mechanism of loading and translocation of type VI secretion system effector
1062 Tse6. *Nature Microbiology* **3**, 1142-1152 (2018).

1063 111. Desfosses, A. et al. Atomic structures of an entire contractile injection system in both the extended
1064 and contracted states. *Nature Microbiology* **4**, 1885-1894 (2019).

1065 112. Shneider, M.M. et al. PAAR-repeat proteins sharpen and diversify the type VI secretion system
1066 spike. *Nature* **500**, 350-353 (2013).

1067 113. Hsu, H.-C., Wang, M., Kovach, A., Darwin Andrew, J. & Li, H. Pseudomonas aeruginosa C-
1068 Terminal Processing Protease CtpA Assembles into a Hexameric Structure That Requires
1069 Activation by a Spiral-Shaped Lipoprotein-Binding Partner. *mBio* **13**, e03680-21 (2022).

1070 114. de Amorim, G.C. et al. The structure of HasB reveals a new class of TonB protein fold. *PLoS One*
1071 **8**, e58964 (2013).

1072 115. Büttner, C.R., Sorg, I., Cornelis, G.R., Heinz, D.W. & Niemann, H.H. Structure of the Yersinia
1073 enterocolitica type III secretion translocator chaperone SycD. *J Mol Biol* **375**, 997-1012 (2008).

1074 116. Tsuchiya, H. et al. Structural basis for amino acid export by DMT superfamily transporter YddG.
1075 *Nature* **534**, 417-420 (2016).

1076 117. Sankaranarayanan, R. et al. The structure of threonyl-tRNA synthetase-tRNA(Thr) complex
1077 enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **97**, 371-81
1078 (1999).

1079 118. Chu, C.H. et al. Crystal structures of HpSoj-DNA complexes and the nucleoid-adaptor complex
1080 formation in chromosome segregation. *Nucleic Acids Res* **47**, 2113-2129 (2019).

1081 119. Buddha, M.R. & Crane, B.R. Structures of tryptophanyl-tRNA synthetase II from Deinococcus
1082 radiodurans bound to ATP and tryptophan. Insight into subunit cooperativity and domain motions
1083 linked to catalysis. *J Biol Chem* **280**, 31965-73 (2005).

1084 120. Lee, S.J., Sung, R.J. & Verdine, G.L. Mechanism of DNA Lesion Homing and Recognition by the
1085 Uvr Nucleotide Excision Repair System. *Research (Wash D C)* **2019**, 5641746 (2019).

1086 121. Cork, A.J. et al. Stability of the Octameric Structure Affects Plasminogen-Binding Capacity of
1087 Streptococcal Enolase. *PLOS ONE* **10**, e0121764 (2015).

1088 122. Soules, K.R. et al. Structural and ligand binding analyses of the periplasmic sensor domain of RsbU
1089 in Chlamydia trachomatis support a role in TCA cycle regulation. *Mol Microbiol* **113**, 68-88 (2020).

1090 123. Teh, A.H. et al. Structure of the RsbX phosphatase involved in the general stress response of
1091 Bacillus subtilis. *Acta Crystallogr D Biol Crystallogr* **71**, 1392-9 (2015).

1092 124. Gavira, J.A. et al. How Bacterial Chemoreceptors Evolve Novel Ligand Specificities. *mBio*
1093 **11**(2020).

1094 125. Tanaka, H. et al. Crystallization of the archaeal transcription termination factor NusA: a significant
1095 decrease in twinning under microgravity conditions. *Acta Crystallogr Sect F Struct Biol Cryst*
1096 *Commun* **63**, 69-73 (2007).

1097 126. Pilotto, S. et al. Structural basis of RNA polymerase inhibition by viral and host factors. *Nat*
1098 *Commun* **12**, 5523 (2021).

1099 127. Soltysova, M. et al. Structural insight into DNA recognition by bacterial transcriptional regulators
1100 of the SorC/DeoR family. *Acta Crystallographica Section D* **77**, 1411-1424 (2021).

1101 128. Chen, K.E. et al. De novo macrocyclic peptides for inhibiting, stabilizing, and probing the function
1102 of the retromer endosomal trafficking complex. *Sci Adv* **7**, eabg4007 (2021).

1103 129. Bashiri, G. et al. Poly-γ-glutamylation of biomolecules. *Nature Communications* **15**, 1310 (2024).

1104 130. Aufhammer, S.W. et al. Crystal structure of methylenetetrahydromethanopterin reductase (Mer) in
1105 complex with coenzyme F420: Architecture of the F420/FMN binding site of enzymes within the
1106 nonprolyl cis-peptide containing bacterial luciferase family. *Protein Sci* **14**, 1840-9 (2005).

1107 131. Graboski, A.L. et al. Mechanism-based inhibition of gut microbial tryptophanases reduces serum
1108 indoxyl sulfate. *Cell Chem Biol* **30**, 1402-1413.e7 (2023).

1109 132. van Tran, N. et al. Evolutionary insights into Trm112-methyltransferase holoenzymes involved in
1110 translation between archaea and eukaryotes. *Nucleic Acids Res* **46**, 8483-8499 (2018).

1111 133. El Bakkouri, M. et al. Structure of RavA MoxR AAA+ protein reveals the design principles of a
1112 molecular cage modulating the inducible lysine decarboxylase activity. *Proc Natl Acad Sci U S A*
1113 **107**, 22499-504 (2010).

1114 134. Uruburu, M. et al. Structural and functional characterization of TgpA, a critical protein for the
1115 viability of Pseudomonas aeruginosa. *Journal of Structural Biology* **205**, 18-25 (2019).
1116

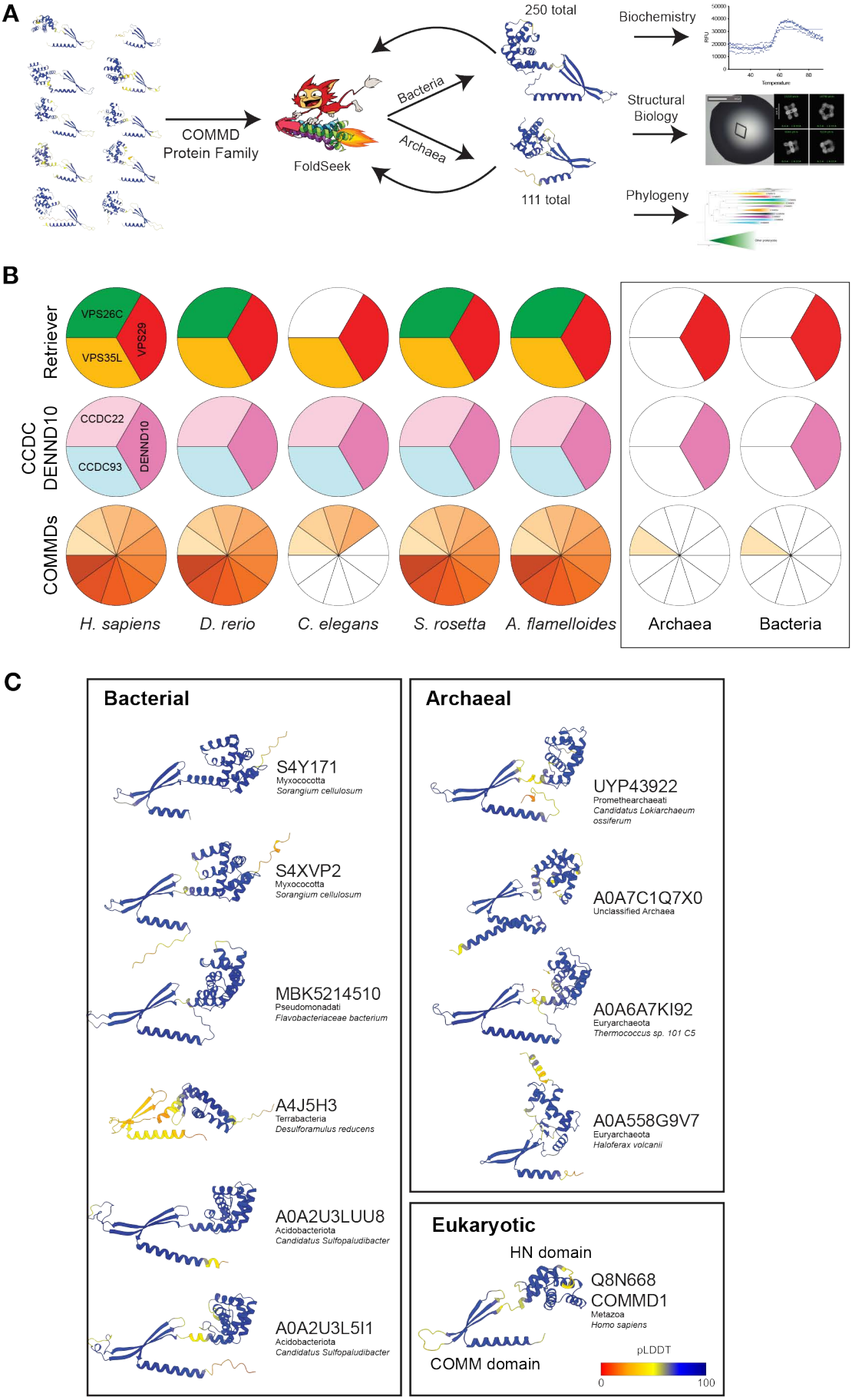


Figure 1. Discovery of ancestral prokaryotic COMMD-like proteins based on structural homology.
(A) Workflow to identify and validate prokaryotic ancestors of Commander subunits with FoldSeek29. Eukaryotic protein structures of Commander subunits were used as search templates in Foldseek, and searches were then performed iteratively using identified prokaryotic proteins until the search-space was exhausted. Selected prokaryotic proteins were then verified using biophysical and structural approaches, while phylogenetic analyses were performed to uncover evolutionary and potential functional relationships. (B) From systematic searches, we did not identify any prokaryotic proteins with structural similarity to the majority of the Commander subunits, but clear structural homologues of the COMMD proteins were found in both archaeal and bacterial species. (C) AlphaFold2 prediction of ten selected bacterial and archaeal COMMD-like proteins (coloured by pLDDT score) show a highly conserved architecture when compared to human COMMD9 including HN and COMM domains. Accession numbers, Phyla and species are shown to the right of each image. For the sake of precise identification proteins have been labelled by either a Uniprot ID or Uniparc ID if Uniprot ID was unavailable.

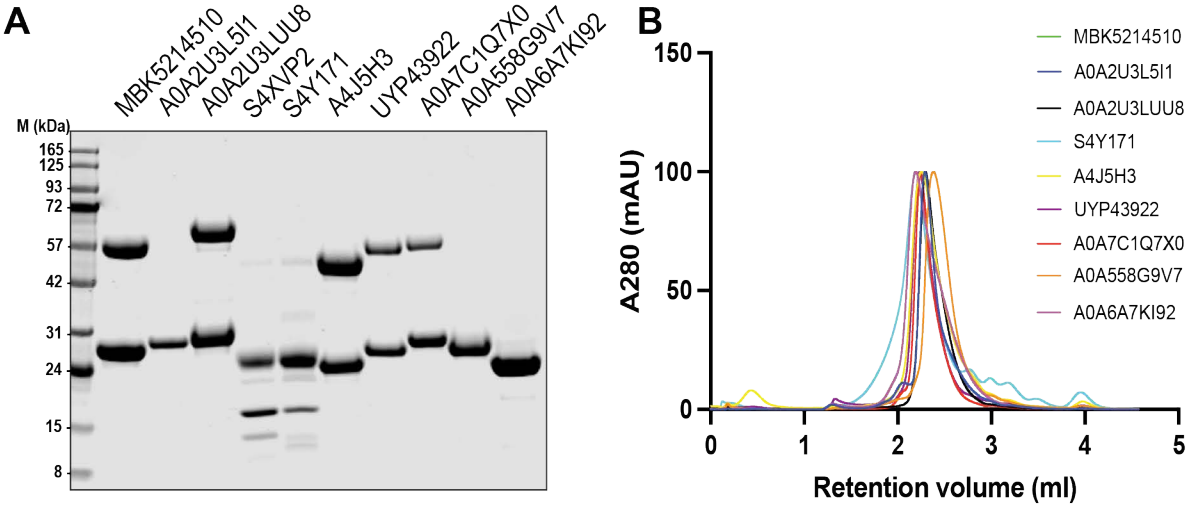
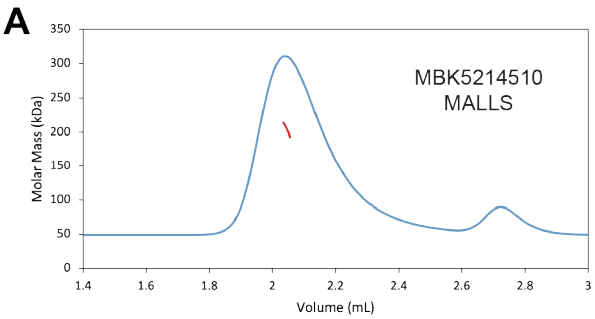
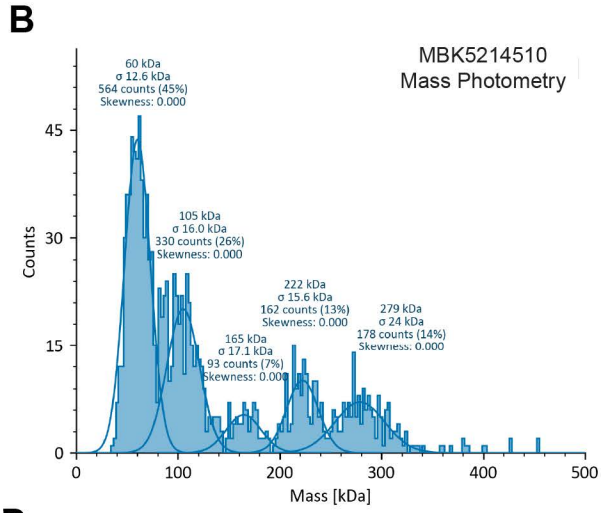
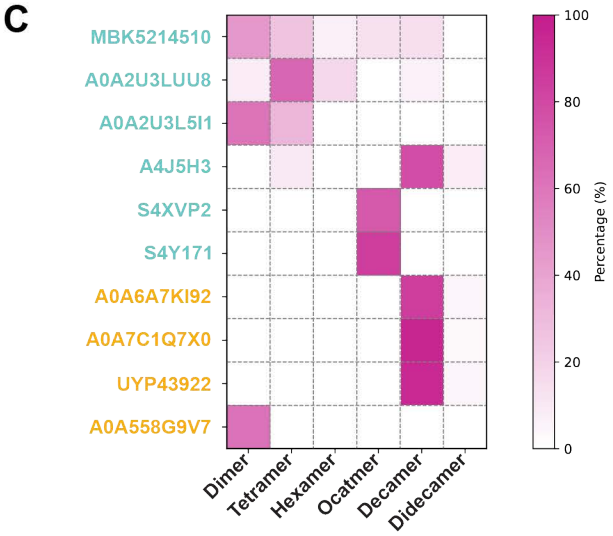


Figure 2. Expression and purification of bacterial and archaeal COMMD-like proteins.
(A) Coomassie-stained SDS-PAGE of purified COMMD-like proteins. The ten COMMD-like proteins shown in Fig. 1C were all expressed in E. coli and purified by His-tag affinity chromatography followed by gel filtration. For several of the COMMD-like proteins a second band can be observed reflecting the obligate dimeric nature of this protein fold. (B) The gel filtration elution profiles of the purified COMMD-like proteins shows these samples to be mostly monodisperse but typically eluting at an earlier time point indicating that they forming oligomers larger than their expected monomeric molecular weights.



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





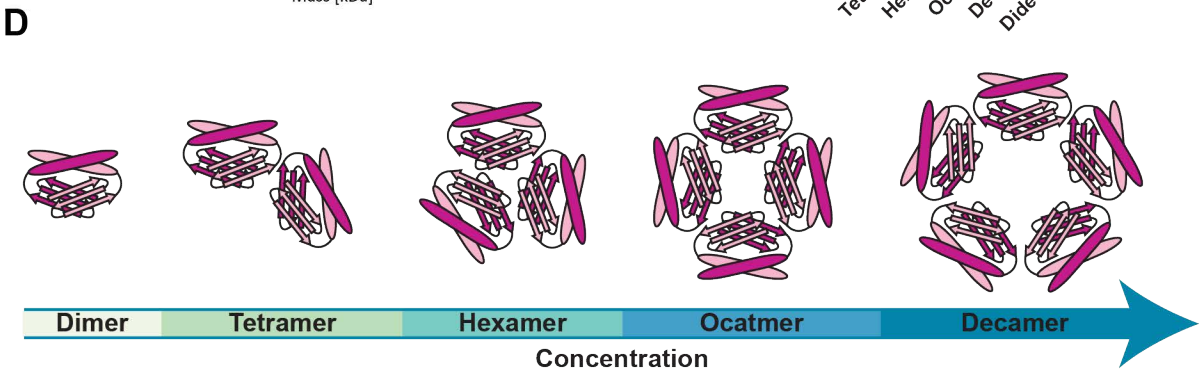


Figure 3. COMMD-like proteins form stable concentration dependent oligomers.
(A) Multiangle laser light scattering (MALLS) of purified MBK5214510. The protein can form an oligomeric structure, however the plots suggest the sample is not fully monodisperse. MALLS experiments were performed for the majority of the purified COMMD-like proteins and the data is shown in Fig. S4A. Proteins were injected at a concentration of ~40 μM based on monomeric mass. (B) Table summarising the observed molecular weights by MALLS, uncertainty of the measure as well as the approximate oligomeric state based on the theoretical monomeric molecular weight. (C) Mass photometry measurement of purified MBK5214510 shows evidence of dimeric, tetrameric, hexameric, octameric and decameric species. Mass photometry was performed for the other COMMD-like proteins with data shown in Fig. S4B. (D) A heatmap summarising the percentage of each oligomeric species in the samples measured by mass photometry. Broadly, the bacterial proteins generally exist as a mixture of oligomeric species while the archaeal proteins appear more stable and monodisperse. S4XVP2 and S4Y171 form stable octamers and A0A558G9V7 appears to only form dimers. Proteins were measured at a final 100-200 nM. (D) Proposed concentration dependent assembly of bacterial COMMD-like proteins.

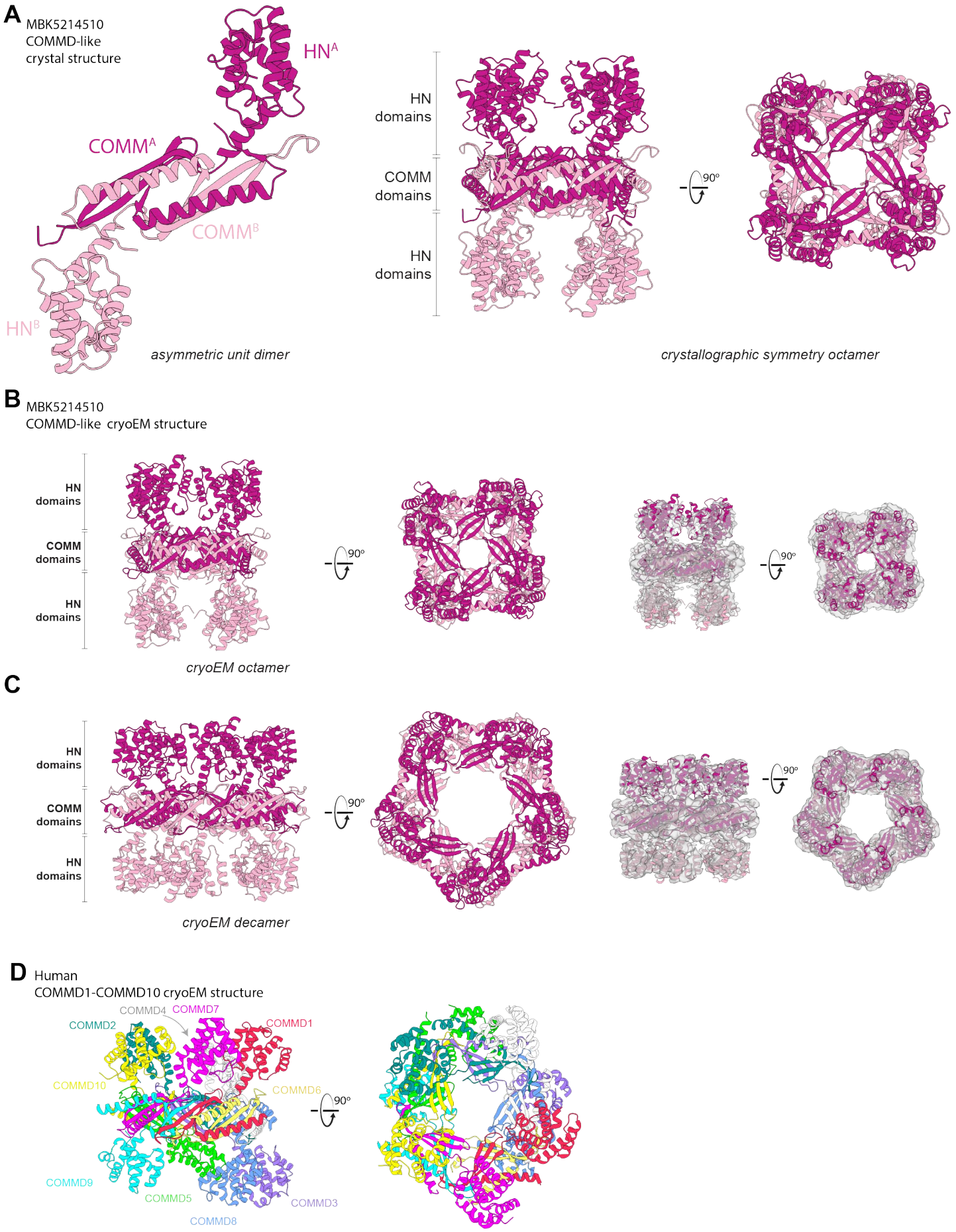


Figure 4. Structures of COMMD-like proteins reveal ring-shaped homo-oligomers that resemble the hetero-decamaric human COMMD complex.
(A) Crystal structure of MBK5214510 from Flavobacteriaceae bacterium. The asymmetric unit contains a dimer of MBK5214510 similar to the dimers of human COMMD proteins. The crystal lattice reveals the presence of an MBK5214510 octamer along the 4-fold axis. (B) and (C) Cryo-EM structures of MBK5214510 octamers and decamers. These different oligomeric species are present in the same purified sample, indicating a degree of structural plasticity in the stoichiometry of the ring-shaped assemblies. In both cases a star shaped structure is observed with C-terminal COMM domains forming a central platform and the HN domains lying peripherally on each side. (D) The Cryo-EM structure of the human COMMD1-COMMD10 hetero-decameric complex18 (PDB ID 8F2U) is shown for comparison to the bacterial homo-oligomeric assemblies.

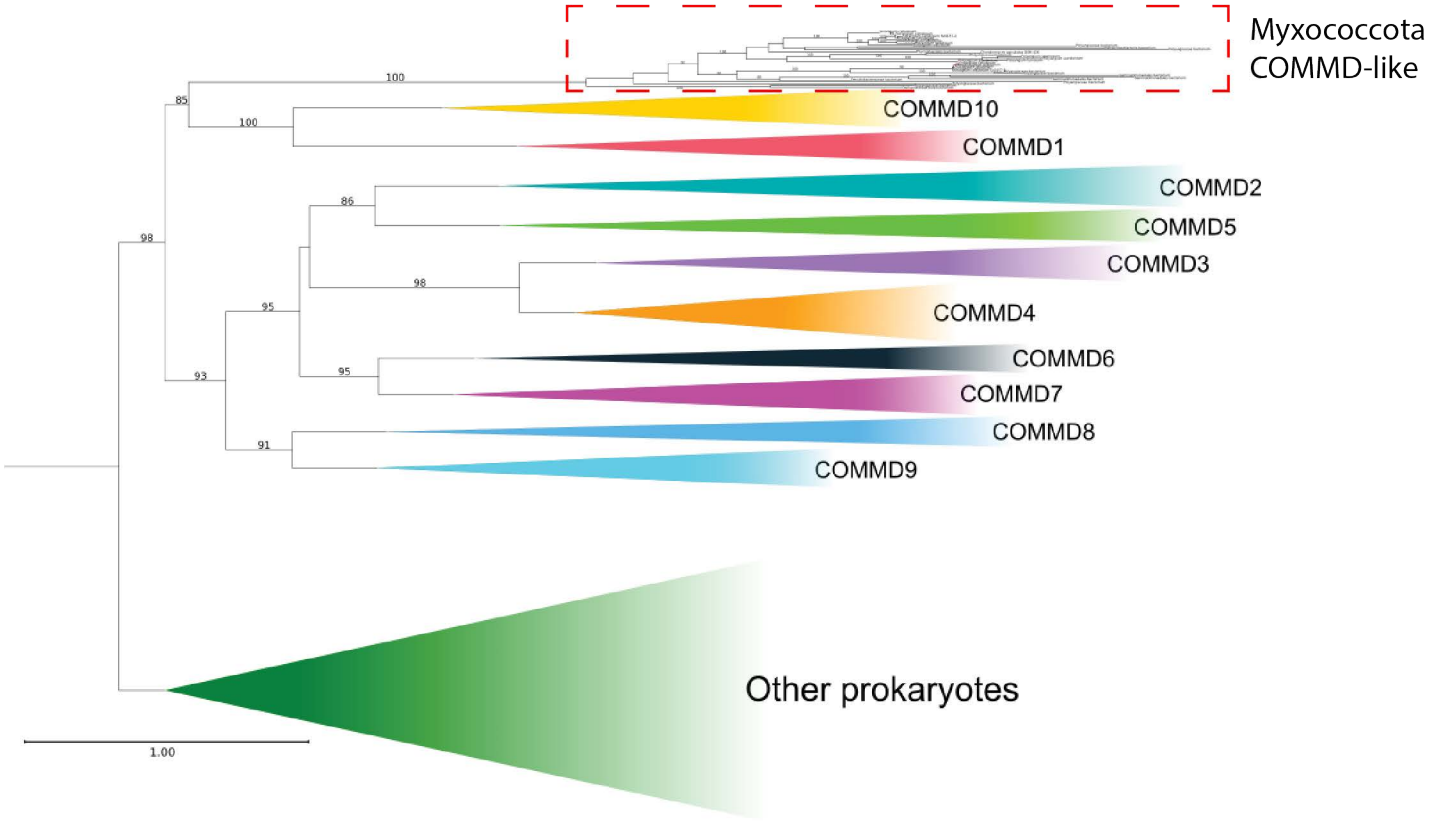


Figure 5. Evolutionary origins of the eukaryotic COMMD proteins.
The phylogenetic tree of the prokaryotic COMMD-like proteins and metazoan COMMD proteins based on sequence comparisons revealed the Myxococcota clade of proteins was nested within the metazoan clade with high support (bootstrap 98%), while Asgard Archaea were considerably more distantly related.

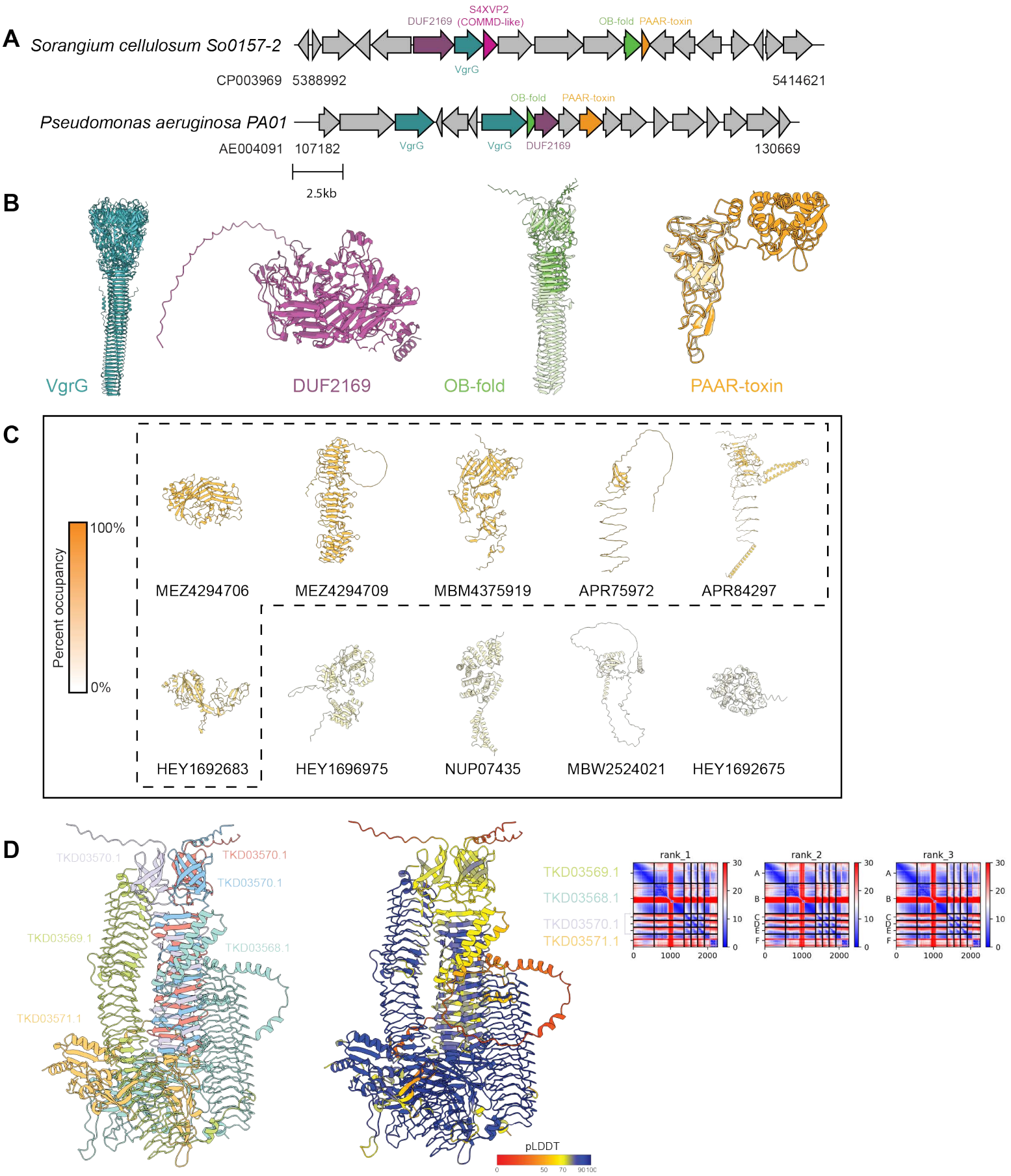


Figure 6. Genomic context of Myxococcota COMMD like proteins suggests a relationship with injection machinery.
(A) Gene maps of Sorangium cellulosum So0157-2 and Pseudomonas aeruginosa PA01 focused on the DUF2169 protein a part of the P. aeruginosa Vgr1B virulence cluster. Note P. aeruginosa does not have a COMMD-like protein. Coloured genes highlight similar structural folds between the two species. (B) Highlighted genes from each species in panel (A) were predicted by AlphaFold2 and structurally aligned. Lighter colours are used for S. cellulosum, to distinguish them from the P. aeruginosa proteins. These proteins represent folds that are commonly associated with bacterial injection machinery. (C) To better understand the genomic context of the COMMD-like genes across the Myxococcota clade we used AlphaFold to predict structures of the encoded proteins from genes ±10,000 bp across 55 species. These proteins were clustered using the "easy-cluster" module available in the FoldSeek software. The structures shown are representatives of structures that appear in at least 20% Myxococcota genomes and are coloured based on the percentage of genomes in which they are found. The dotted box indicates proteins with folds related to known injection machinery proteins. (D) Iterative AlphaFold predictions were performed to assess potential interactions among the proteins that appeared to be related to bacterial injection machinery complexes. These consistently revealed a complex that was predicted with high confidence that incorporates four genes assembled into a hetero-hexameric complex. The predicted complex shown is an example from Polyangium fumosum composed of a homotrimer of TKD03570 and single copies each of TKD03571, TKD03569, TKD03568. The complex shares strong similarities to components of the Type VI secretion system baseplate machinery46.

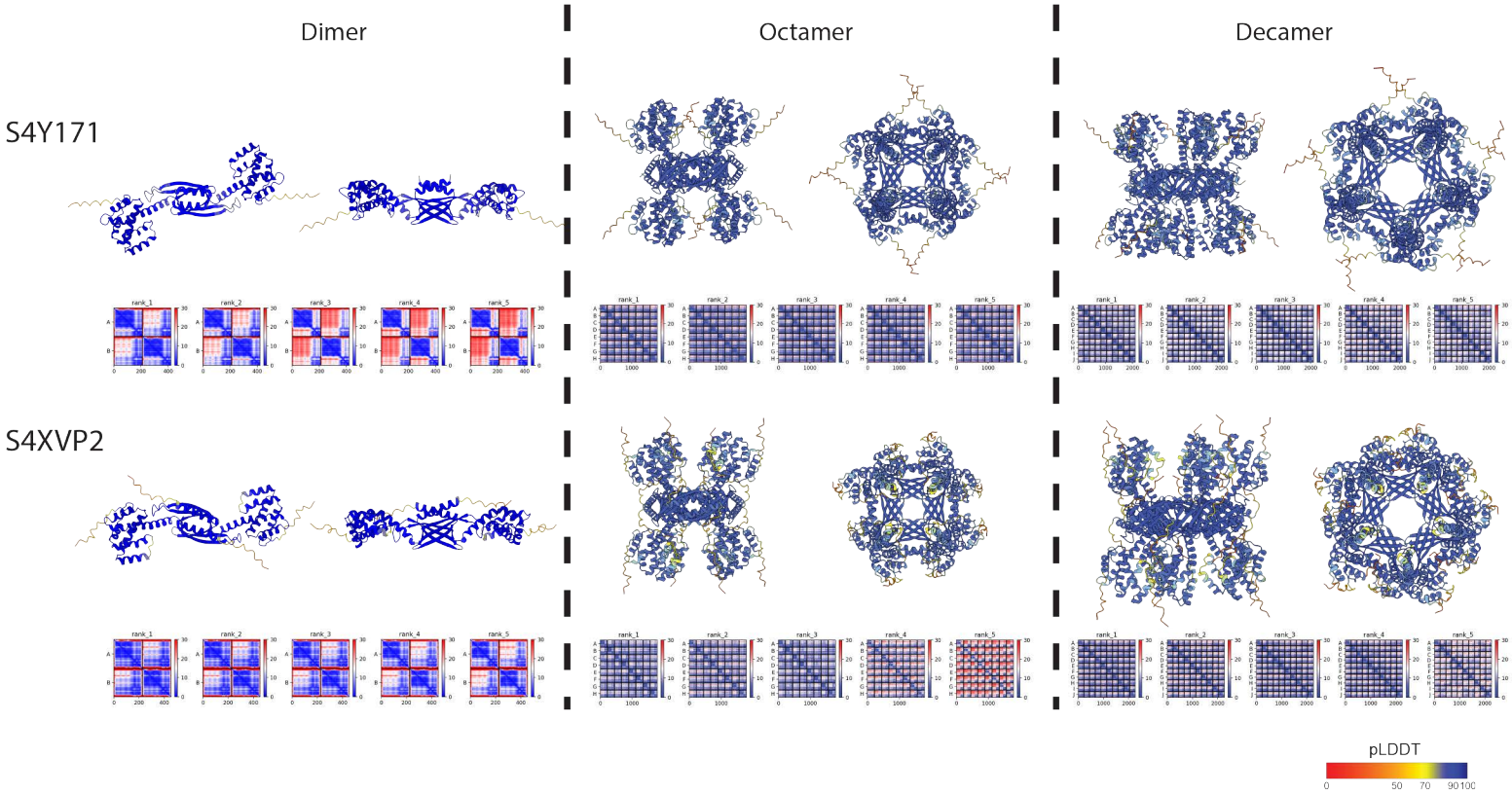


Figure S1. AlphaFold2 predicted structures of COMMD-like proteins from Myxococcota.
The structures of S4Y171 and S4XVP2 from Sorangium cellulosum were predicted using AlphaFold2 in various homooligomeric forms, with the results shown for dimer, octamer and decamer predictions. Proteins care coloured according to pLDDT, with PAE plots shown below.

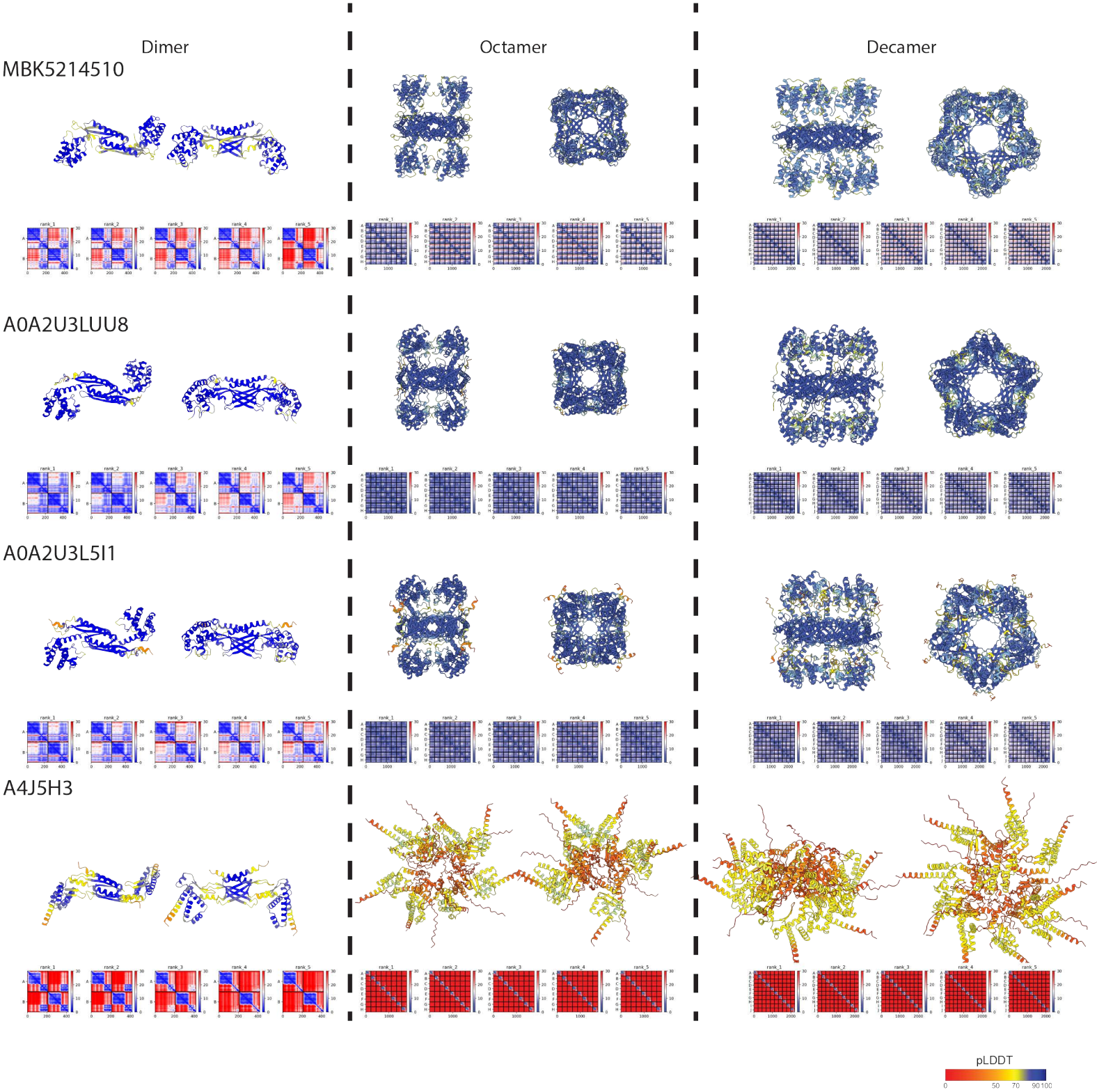


Figure S2. AlphaFold2 predicted structures of COMMD-like proteins from various bacterial species.
The structures of A4J5H3 (Desulfotomaculum reducens), A0A2U3L5I1 (Candidatus Sulfopaludibacter), A0A2U3LUU8 (Candidatus Sulfopaludibacter), MBK5214510 (Flavobacteriaceae bacterium) were predicted using AlphaFold2 in various homooligomeric forms, with the results shown for dimer, octamer and decamer predictions. Proteins care coloured according to pLDDT, with PAE plots shown below.

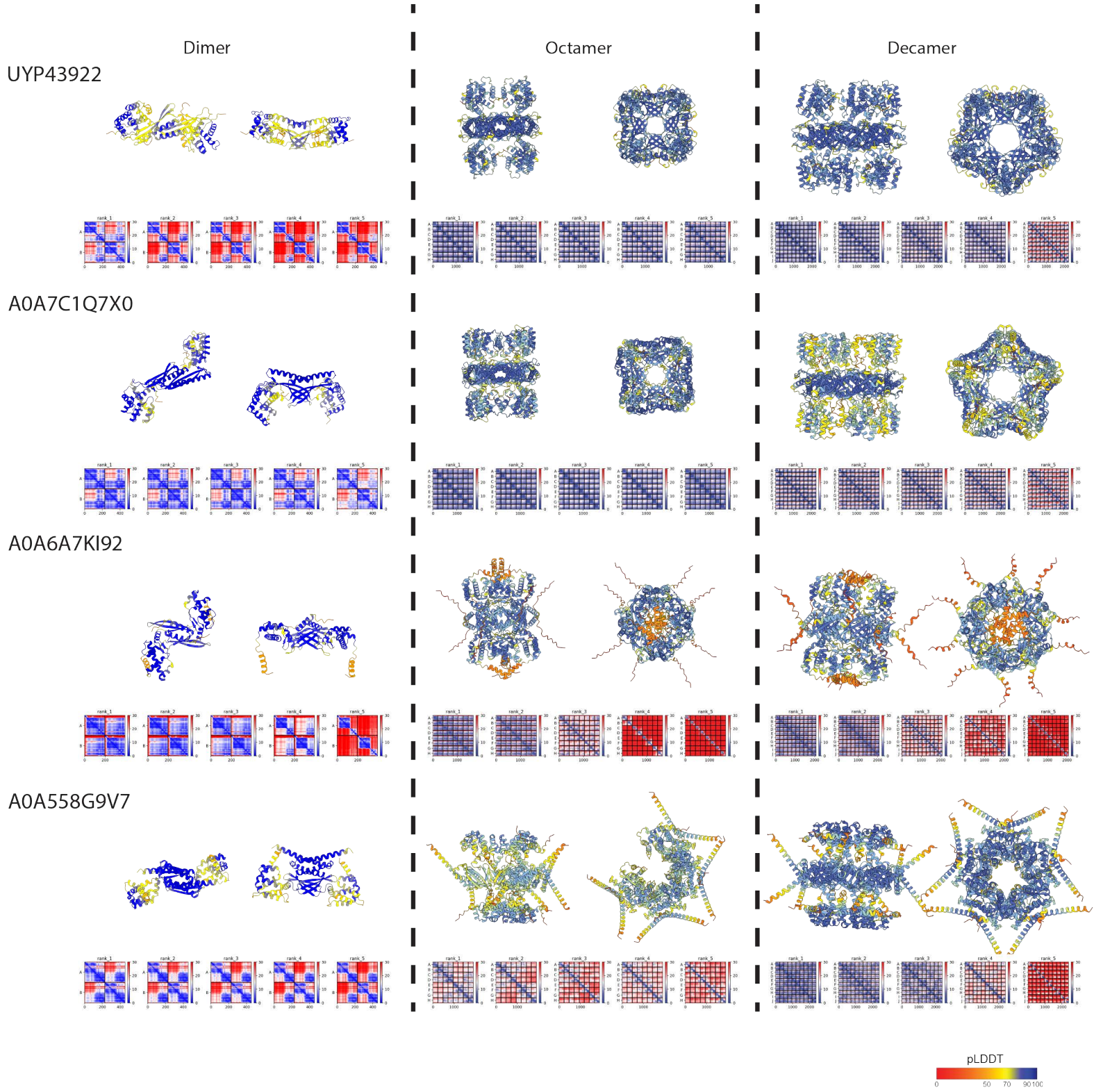


Figure S3. AlphaFold2 predicted structures of COMMD-like proteins from various archaeal species.
The structures of UYP43922 (Candidatus Lokiarchaeum ossiferum), A0A7C1Q7X0 (Unclassified), A0A6A7KI92 (Thermococcus sp. 101 C5), A0A558G9V7 (Haloferax volcanii) were predicted using AlphaFold2 in various homooligomeric forms, with the results shown for dimer, octamer and decamer predictions. Proteins care coloured according to pLDDT, with PAE plots shown below.

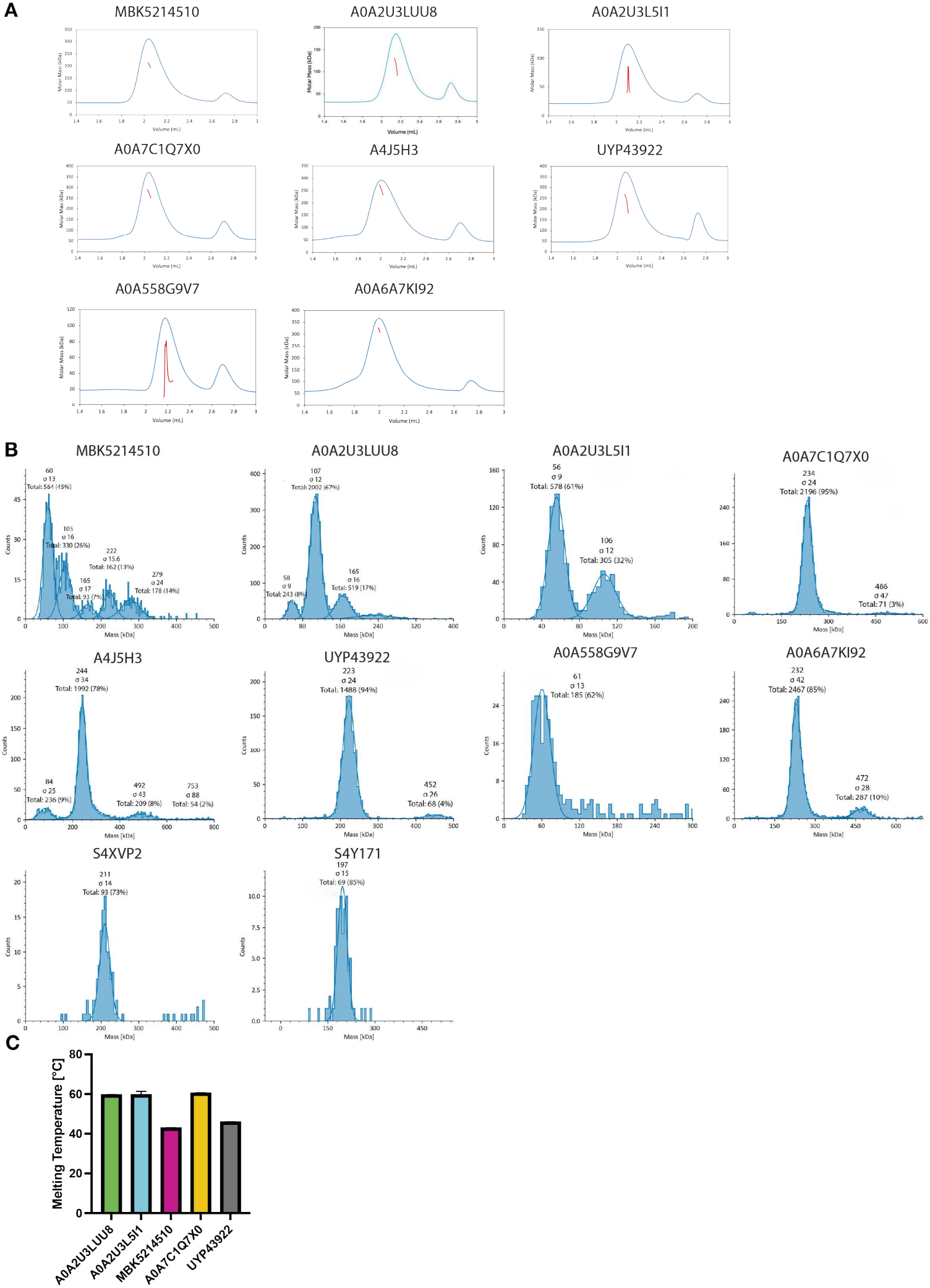


Figure S4. Biophysical characterisation of archaeal and bacterial COMMD-like proteins
(A) MALLS data and (B) Mass photomometry data for bacterial and archaeal COMMD-like proteins. (C) NanoDSF shows two groupings of COMMD-like proteins with respect to thermal stability with a TM50 of ~60oC and 40oC.

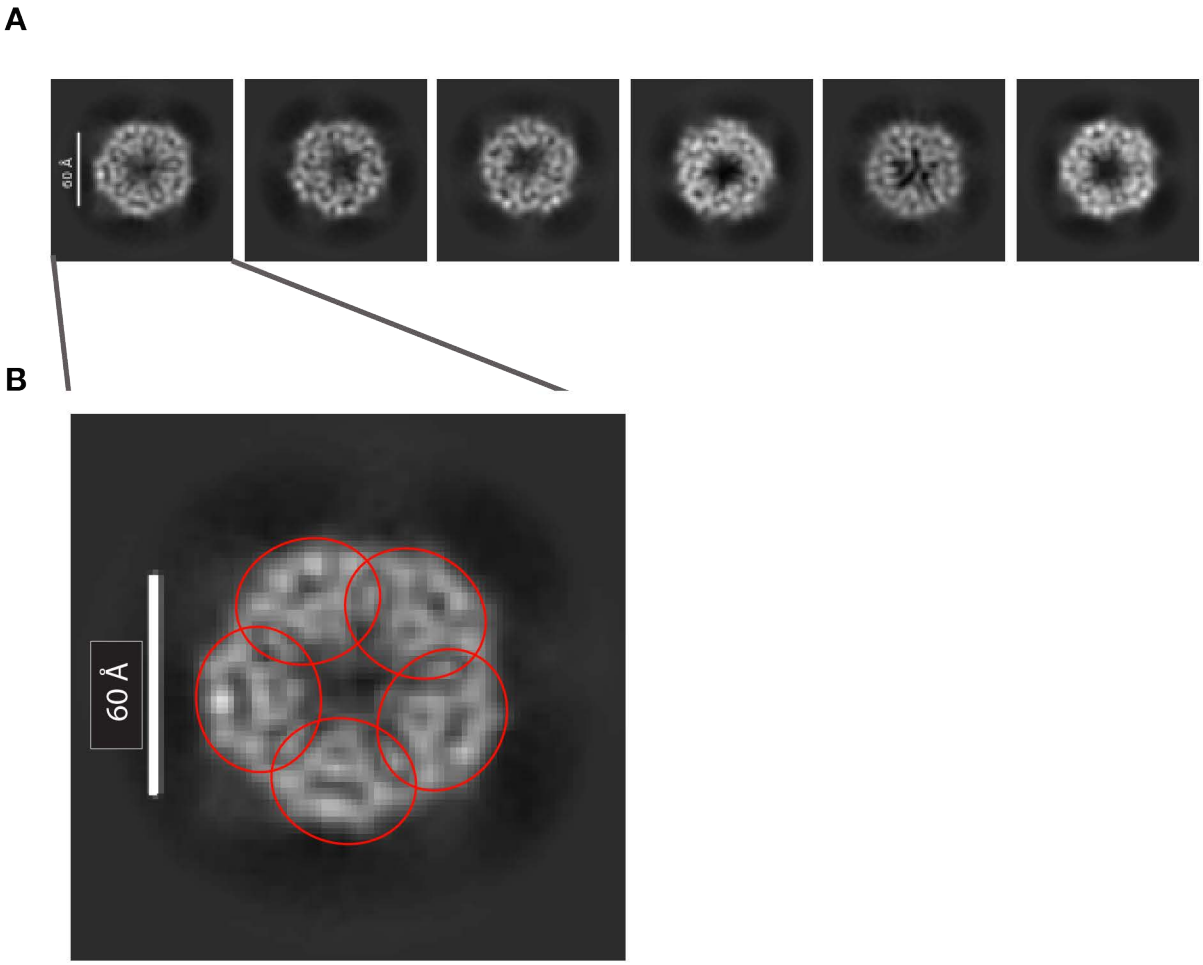
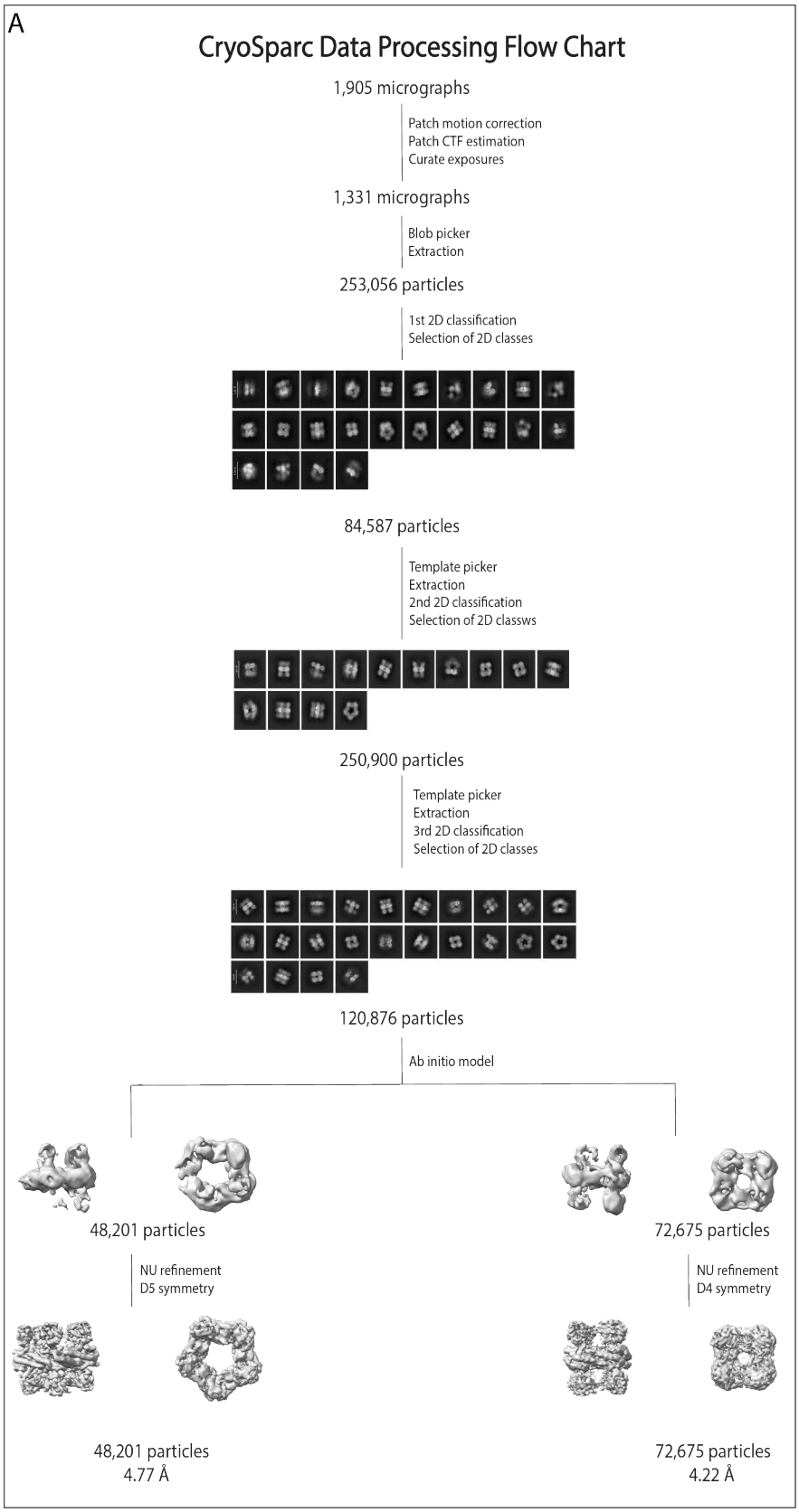
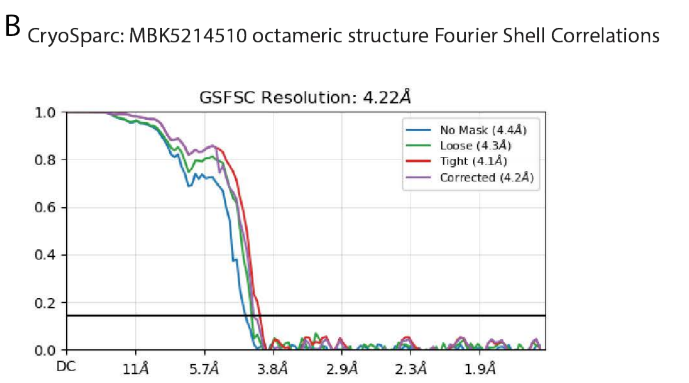
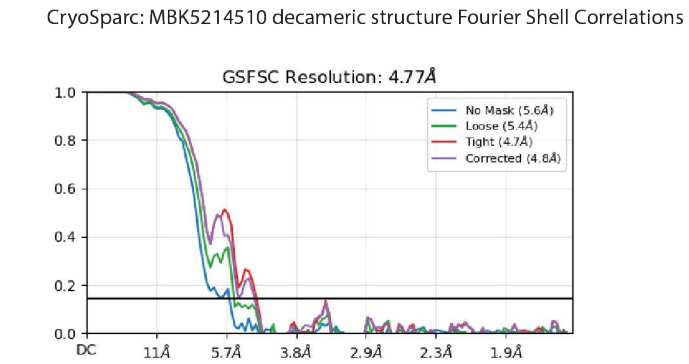
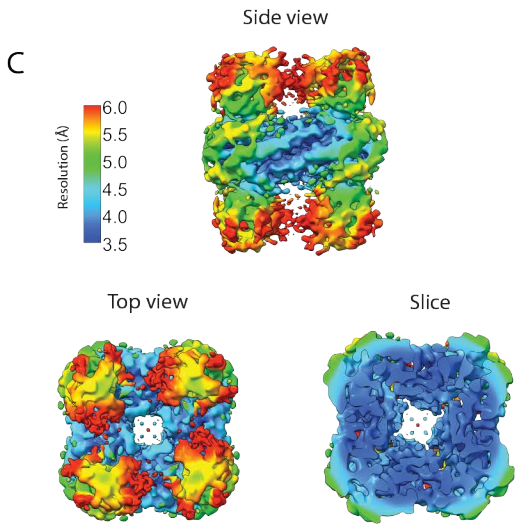


Figure S5. Cryo-EM of A0A2U3L5I from Candidatus Sulfopaludibacter.
2D class averages indicating a decameric complex of COMMD-like protein A0A2U3L5I from Candidatus Sulfopaludibacter.









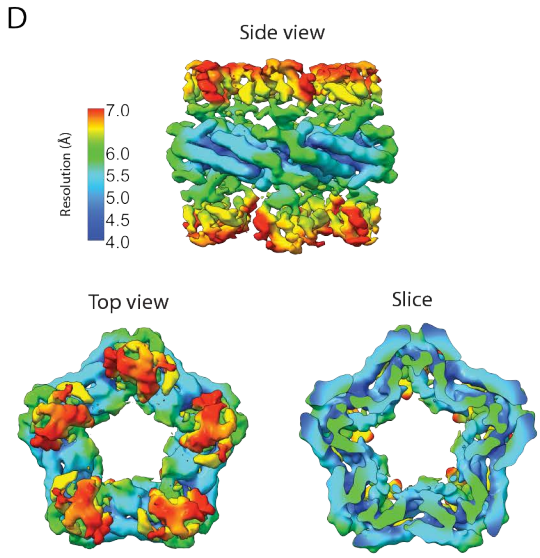


Figure S6. Cryo-EM workflow for structure determination of MBK5214510.
(A) Cryo-EM data processing workflow using CryoSPARC software. The final 3D reconstruction map was used for model building and refinement. (B) 3D Fourier Shell Correlation of the final reconstructions reporting global resolution, with the FSC at 0.143 denoted. The model-to-map FSC plot from Phenix real space refinement is overlaid. (C) Final reconstruction of octameric structure filtered and colored by local resolution, as estimated in CryoSPARC. (D) Final reconstruction of decameric structure filtered and colored by local resolution, as estimated in CryoSPARC.

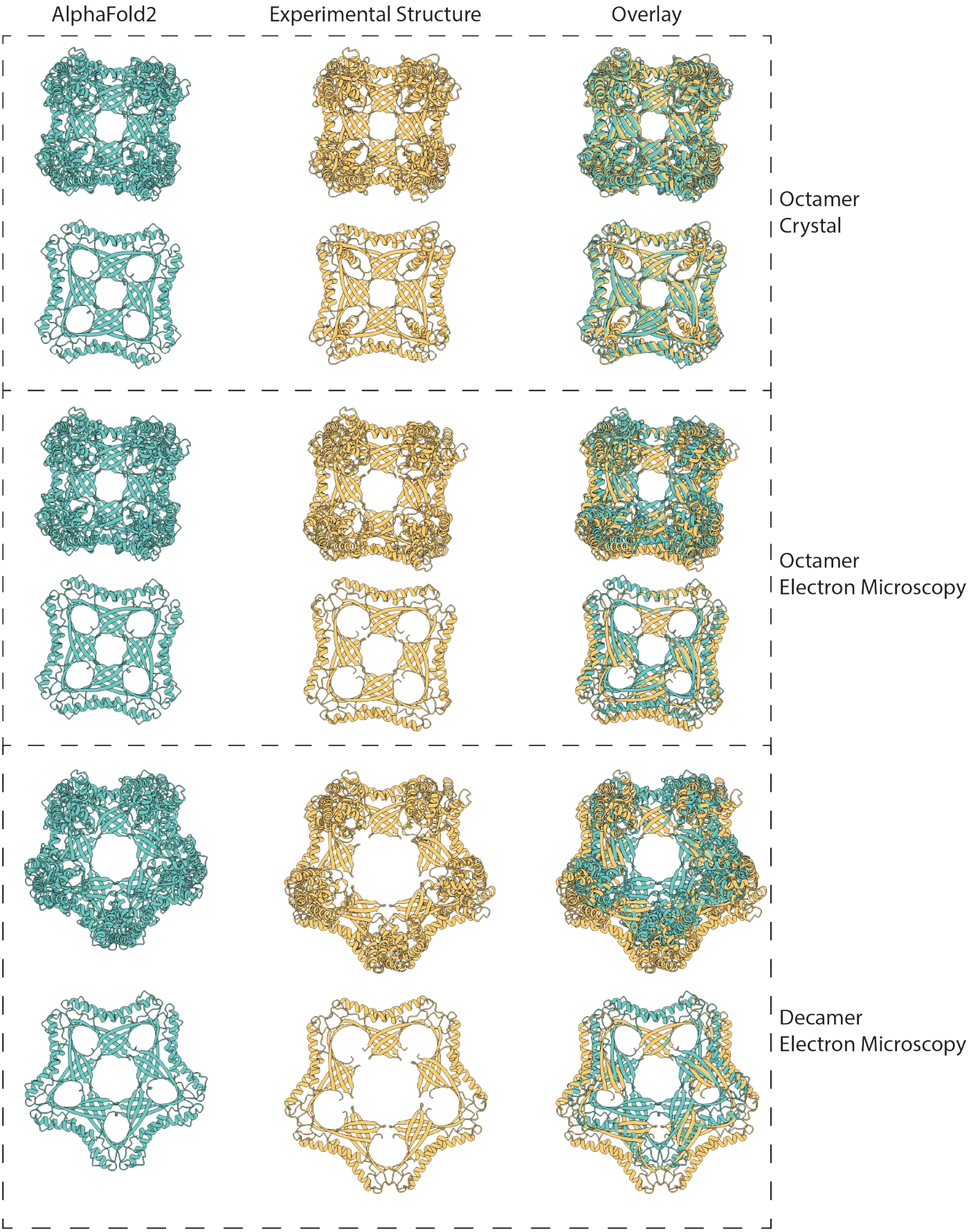


Figure S7. Comparison of the experimental MBK5214510 COMMD-like structures with the AlphaFold predictions.
Left panels show AlphaFold2-predicted structures of MBK5214510 COMMD-like protein from Flavobacteriaceae bacterium with the lower images showing just the central ring of COMM domains that form the core of the homooligomeric rings. Middle panels show the experimental structures of MBK5214510 octamers and decamers determined by X-ray crystallography and cryo-EM and right-hand panels show the overlay of the predicted and experimental structures.

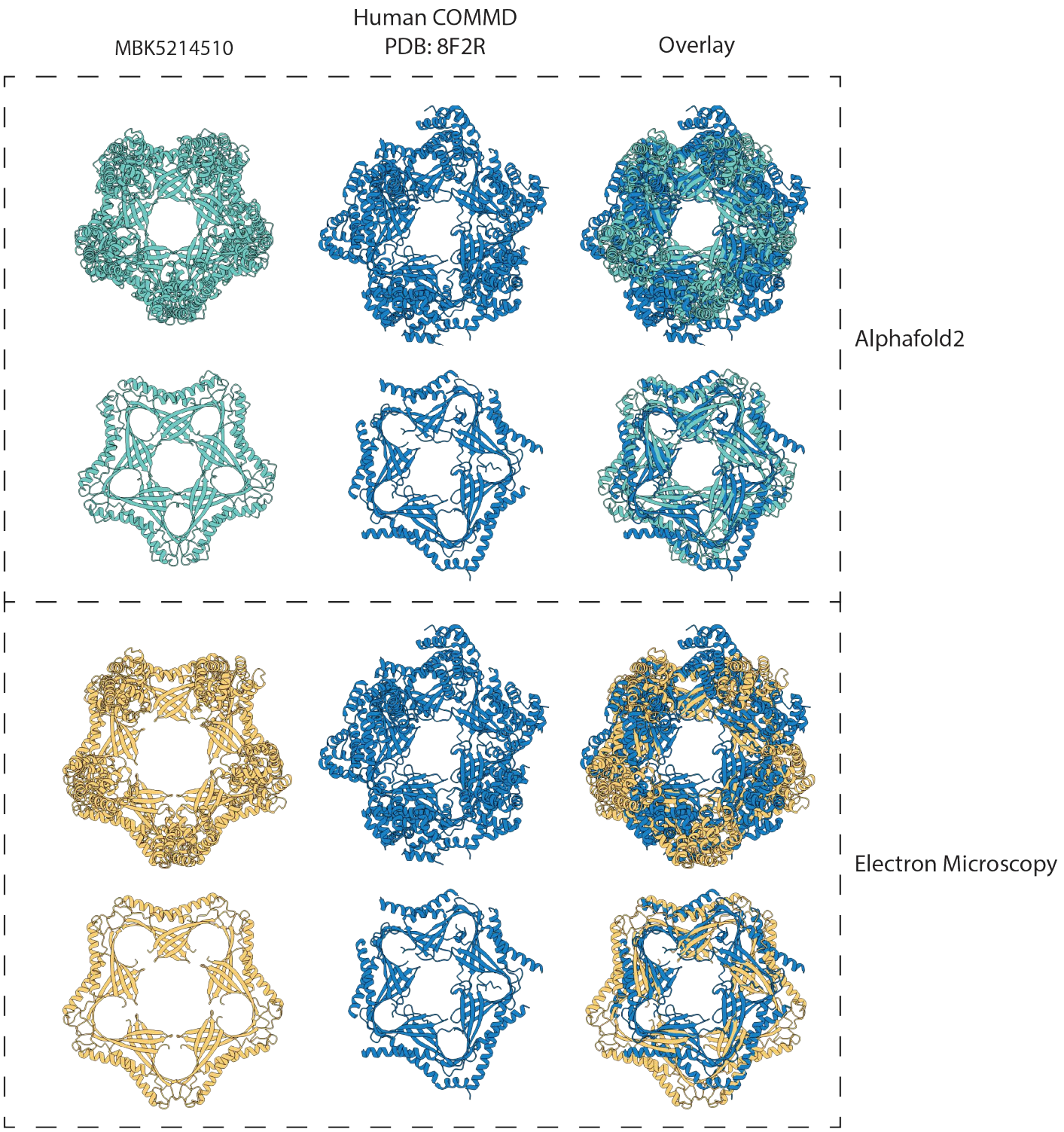


Figure S8. Comparison of the experimental MBK5214510 COMMD-like homodecameric structure with the human COMMD heterodecameric complex. Left panels show either the AlphaFold2-predicted or cryo-EM homodecameric structure of MBK5214510 COMMD-like protein from Flavobacteriaceae bacterium with the lower images showing just the central ring of COMM domains that form the core of the homooligomeric rings. Middle panels show either the AlphaFold2-predicted or cryo-EM heterodecameric structure the human COMMD1-COMMD10 complex18 (PDB ID 8F2U) and right-hand panels show the overlay of the bacterial and human structures.

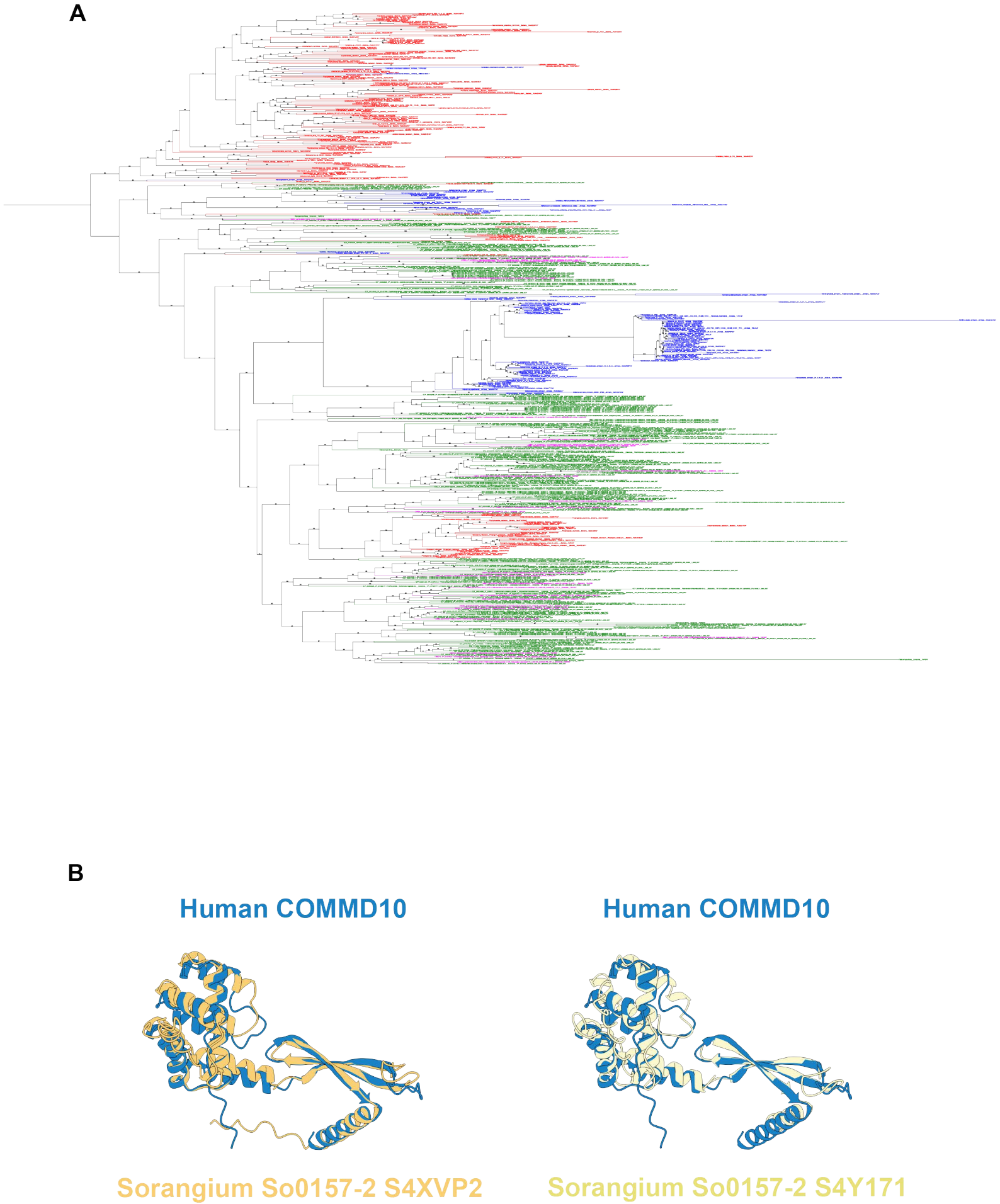


Figure S9. Structure-based phylogenetics.
(A) 3Di phylogeny showing the relationship between archaeal (blue) and bacterial (red) COMMD-like proteins and eukaryotic COMMD1-COMMD10 proteins (green). (B) Example structural alignment of the AlphaFold2 predictions of COMMD10 with two Mycococcota COMMD-like proteins S4Y171 and S4XVP2 from Sorangium cellulosum.

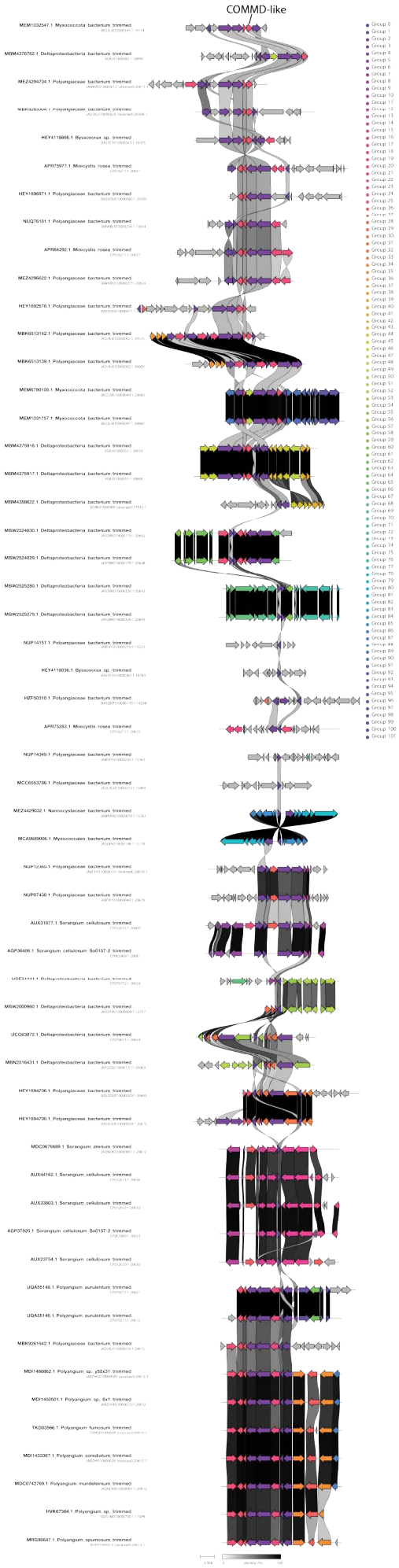


Figure S10. Gene clusters surrounding COMMD-like genes in Myxococotta.
Analysis of the genomic surroundings of COMMD-like genes from Myxococotta species using CLINKER.

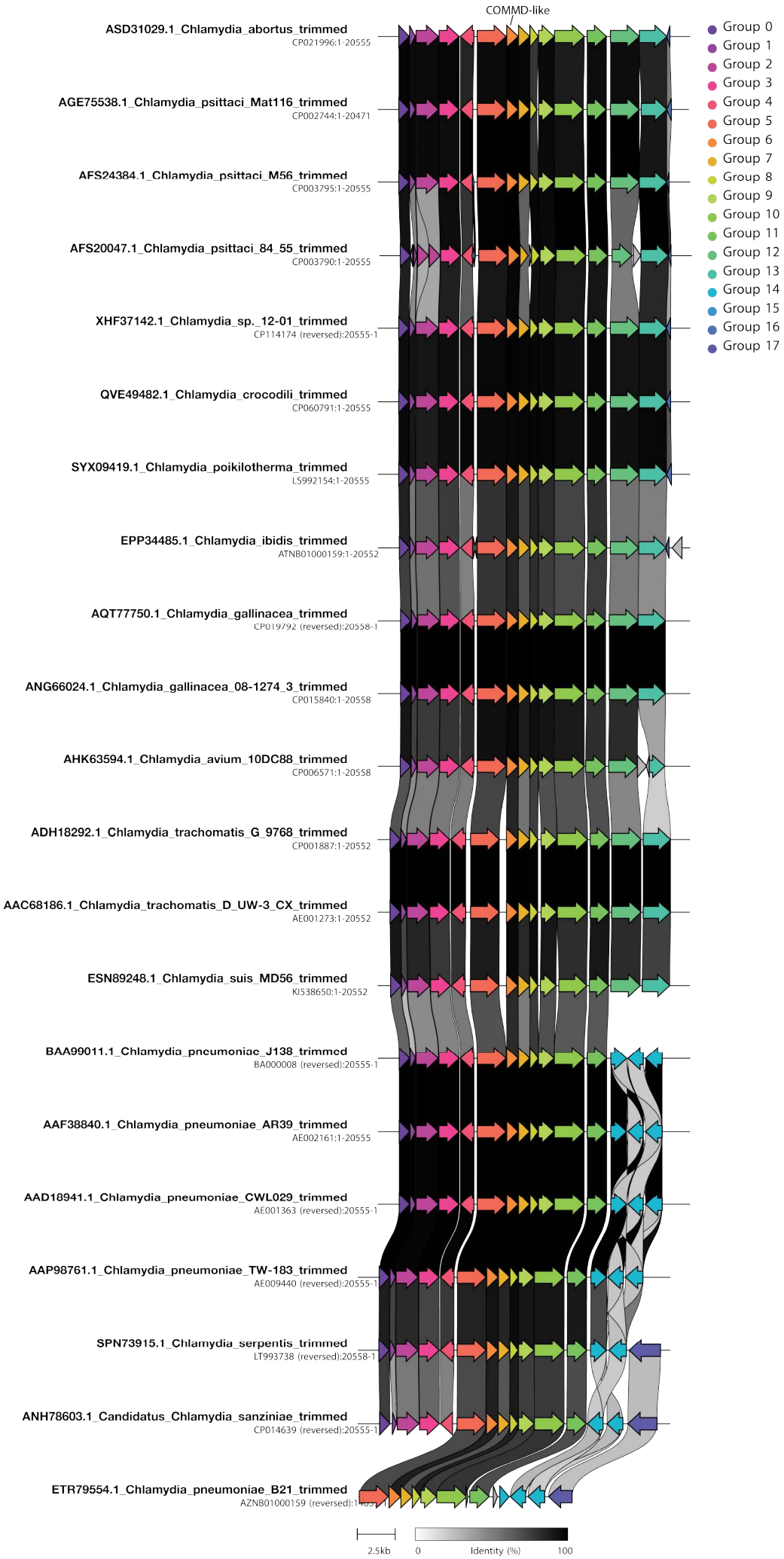


Figure S11. Gene clusters surrounding COMMD-like genes in Chlamydia.
Analysis of the genomic surroundings of COMMD-like genes from Chlamydial species using CLINKER.

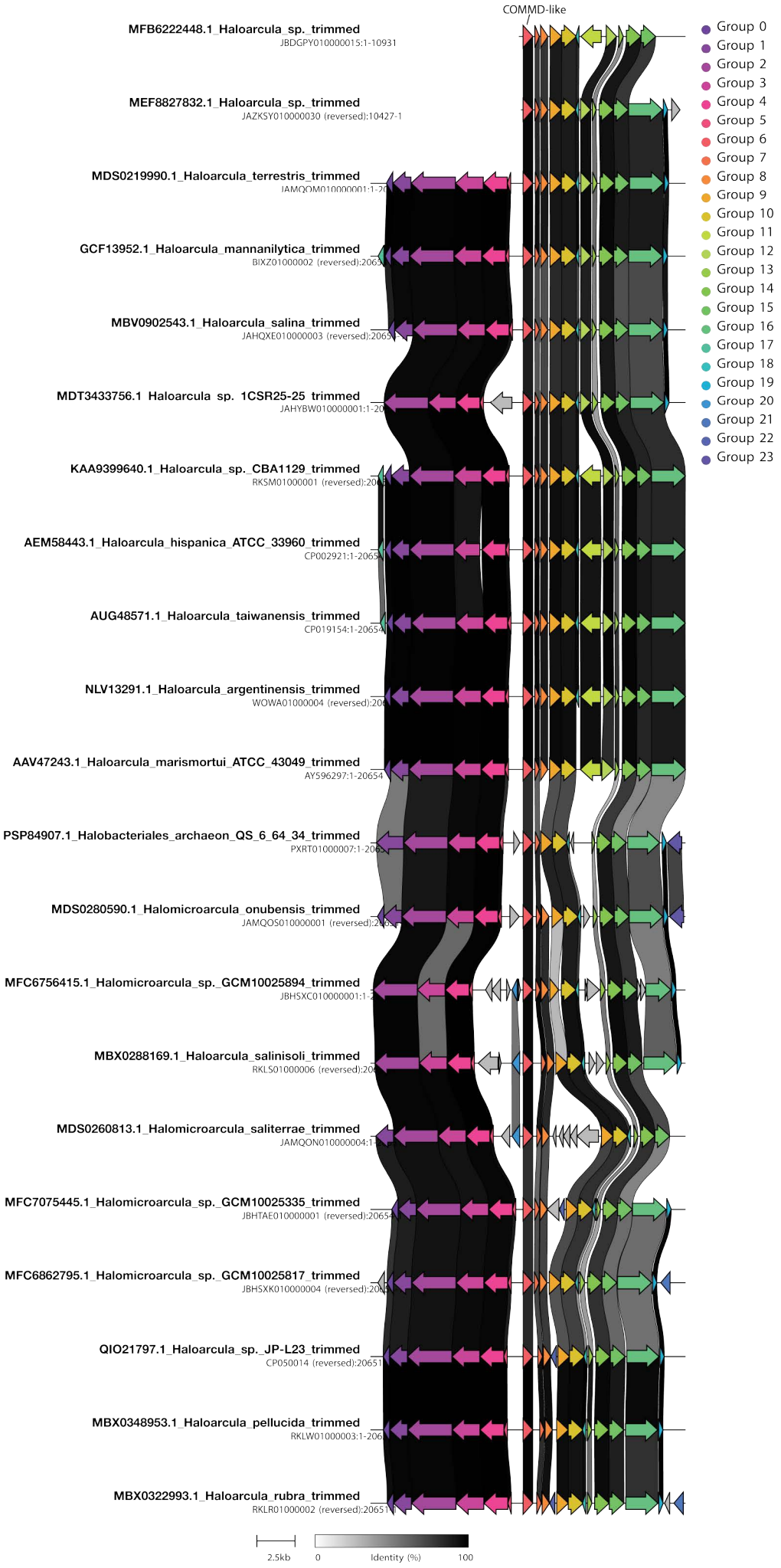


Figure S12. Gene clusters surrounding COMMD-like genes in Haloarchaea.
Analysis of the genomic surroundings of COMMD-like genes from Haloarchaea species using CLINKER.