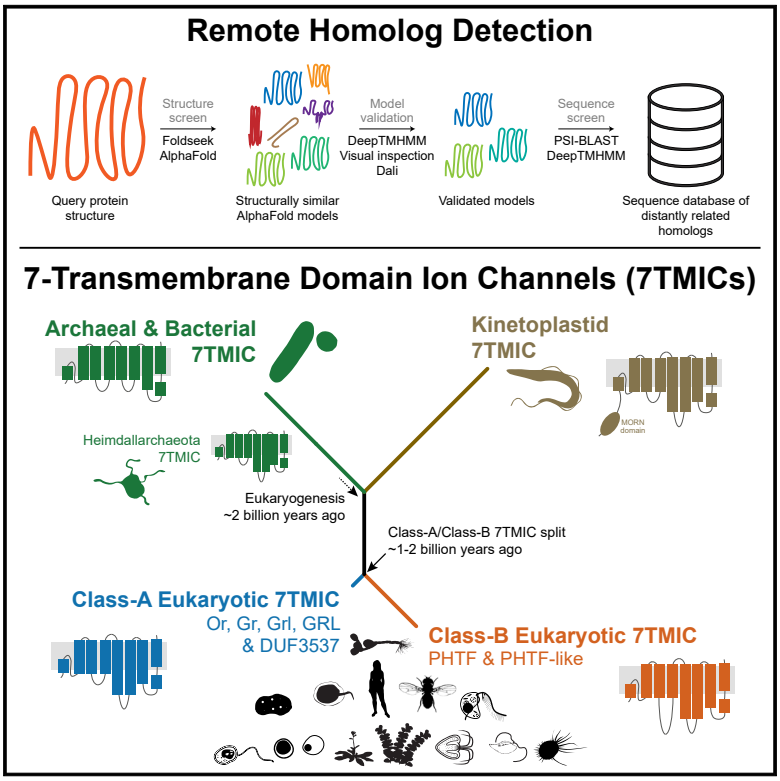
Report

Current Biology

Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life

Graphical abstract



Highlights

- Synergizing protein structure and language models to improve homolog detection
- Relatives of insect Ors are present across the tree of life, including Prokaryota
- These 7-transmembrane domain ion channels (7TMICs) form a homologous superfamily
- 7TMIC evolution involved one early duplication and many taxon-specific radiations

Authors

Nathaniel J. Himmel, David Moi, Richard Benton

Correspondence

nathanieljohn.himmel@unil.ch (N.J.H.), richard.benton@unil.ch (R.B.)

In brief

Homologs of insect odorant receptors (7-transmembrane-domain ion channels [7TMICs]) are difficult to recognize and characterize phylogenetically due to rapid divergence. Himmel et al. describe structure- and sequence-based analyses that identify 7TMICs across the tree of life, demonstrate homology, and trace their evolution over ~3–4 billion years.

Himmel et al., 2023, Current Biology *33*, 5023–5033
November 20, 2023
https://doi.org/10.1016/j.cub.2023.10.008

CellPress

Current Biology

CellPress

Report

Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life

Nathaniel J. Himmel,[1,*] David Moi,[2] and Richard Benton[1,3,4,*]
[1]Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, 1015 Lausanne, Switzerland
[2]Department of Computational Biology, Faculty of Biology and Medicine, University of Lausanne, 1015 Lausanne, Switzerland
[3]X (formerly Twitter): @bentonlab
[4]Lead contact
*Correspondence: nathanieljohn.himmel@unil.ch (N.J.H.), richard.benton@unil.ch (R.B.)
https://doi.org/10.1016/j.cub.2023.10.008

SUMMARY

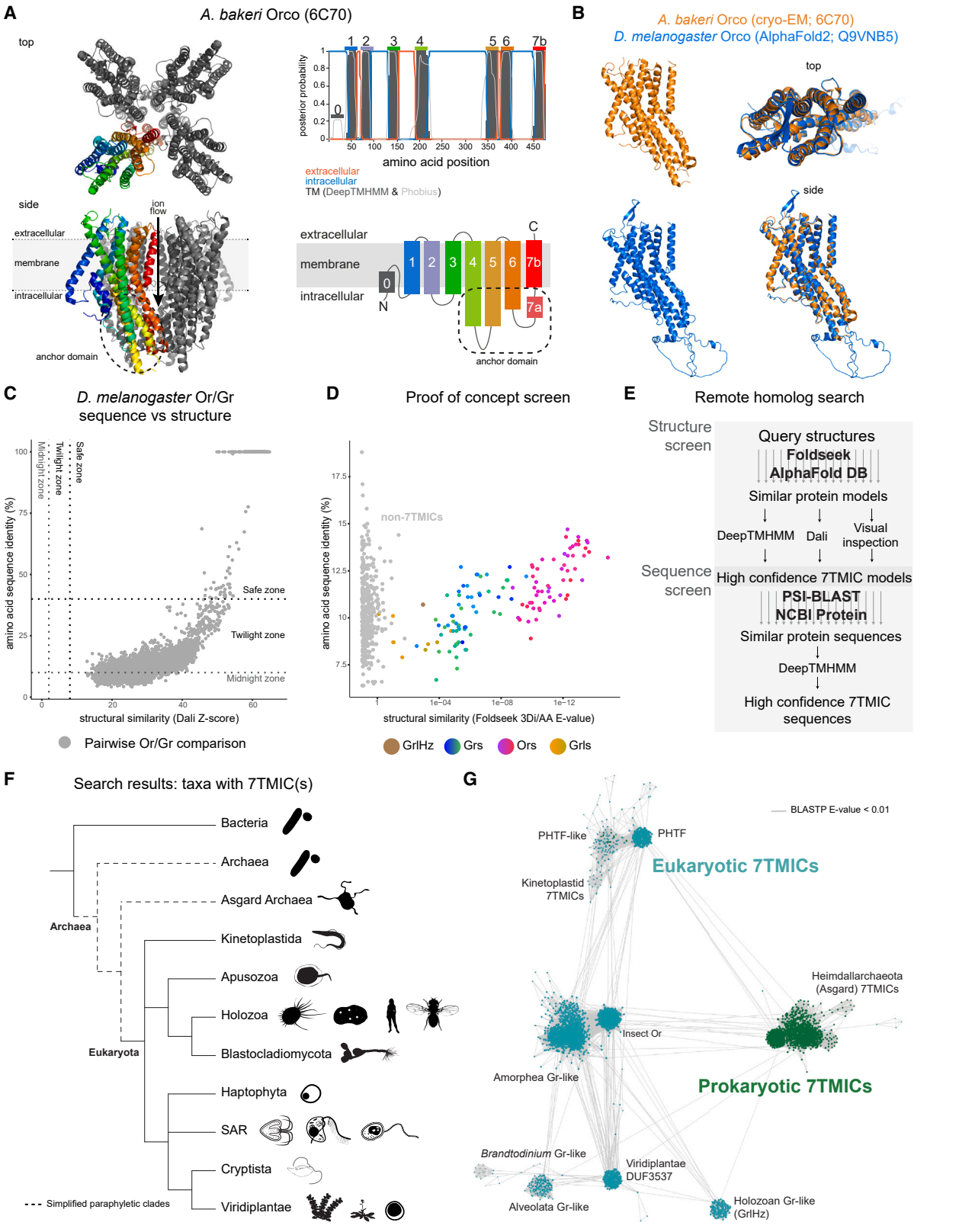
Many proteins exist in the so-called "twilight zone" of sequence alignment, where low pairwise sequence identity makes it difficult to determine homology and phylogeny.[1,2] As protein tertiary structure is often more conserved,[3] recent advances in *ab initio* protein folding have made structure-based identification of putative homologs feasible.[4–6] We present a pipeline for the identification and characterization of distant homologs and apply it to 7-transmembrane-domain ion channels (7TMICs), a protein group founded by insect odorant and gustatory receptors. Previous sequence and limited structure-based searches identified putatively related proteins, mainly in other animals and plants.[7–10] However, very few 7TMICs have been identified in non-animal, non-plant taxa. Moreover, these proteins' remarkable sequence dissimilarity made it uncertain whether disparate 7TMIC types (Gr/Or, Grl, GRL, DUF3537, PHTF, and GrlHz) are homologous or convergent, leaving their evolutionary history unresolved. Our pipeline identified thousands of new 7TMICs in archaea, bacteria, and unicellular eukaryotes. Using graph-based analyses and protein language models to extract family-wide signatures, we demonstrate that 7TMICs have structure and sequence similarity, supporting homology. Through sequence- and structure-based phylogenetics, we classify eukaryotic 7TMICs into two families (Class-A and Class-B), which are the result of a gene duplication predating the split(s) leading to Amorphea (animals, fungi, and allies) and Diaphoretickes (plants and allies). Our work reveals 7TMICs as a cryptic superfamily, with origins close to the evolution of cellular life. More generally, this study serves as a methodological proof of principle for the identification of extremely distant protein homologs.

RESULTS AND DISCUSSION

Insect odorant and gustatory receptors (Ors/Grs) are 7-transmembrane-domain ion channels (7TMICs) critical for the behavior and evolution of insects.[7,11,12] Although originally thought to be an insect-specific protein family,[13–18] the genomic revolution enabled sequence-based searches to identify putative homologs in animals (gustatory-receptor-like proteins [Grls]), plants (DUF3537 proteins), and single-celled eukaryotes (GRLs).[7–9,19] However, 7TMICs remained taxonomically sparse, recognized in only 7 unicellular eukaryotic species, and missing from several holozoan lineages (e.g., chordates, choanoflagellates, and sponges).[7–9,19]

The best-characterized 7TMICs are Ors, which function as odor-gated heterotetrameric (or in some cases, homotetrameric) ion channels.[20–23] A substantial breakthrough came from two Or cryoelectron microscopy (cryo-EM) structures: the fig wasp, *Apocrypta bakeri*, Or co-receptor (Orco)[20] (Figure 1A); and the jumping bristletail, *Machilis hrabei*, Or5.[21] Or monomers have several notable structural features, including: (1) 7 transmembrane α helices; (2) an intracellular N terminus and extracellular C terminus; (3) shorter extracellular than intracellular loops (ILs); (4) long TM4, TM5, and TM6 helices extending into the intracellular space, forming the "anchor domain," where intersubunit interactions occur; (5) a "split" TM7 helix, composed of an intracellular TM7a (part of the anchor domain) and a transmembrane-spanning TM7b (that lines the ion-conducting pore); and (6) an N-terminal re-entrant loop (TM0).[20,21,24,25] These structural features are remarkably conserved despite low sequence conservation; for example, the two experimental structures have virtually indistinguishable folds while sharing only 19% amino acid sequence identity.[21] Importantly, these structures can be accurately predicted *in silico* by several algorithms,[8,25] notably AlphaFold (Figure 1B).[4,10]

Recently, we took advantage of the structural similarity of 7TMICs to perform structure-based screens for putative homologs that had not been identified by sequence-based screening.[10] These screens identified several proteins adopting the 7TMIC fold, including: fly-specific Grls; a highly conserved lineage of eukaryotic proteins (PHTFs, an acronym for the misnomer putative homeodomain transcription factor); a holozoan-specific Grl lineage (GrlHz); and trypanosome 7TMICs. However, these searches



(legend on next page)

were limited by the high computational requirements of the structural alignment tool—Dali[26,27]—and only ~564,000 AlphaFold models from 48 species were screened. Thus, large taxonomic gaps still exist: fewer than 50 proteins have been identified outside of animals and plants, and none have been identified in prokaryotes (despite screening 17 prokaryotes[10]). Beyond the technical limitations leading to sparse taxonomic sampling, the PHTF, GrlHz, Gr/Or, DUF3537, and unicellular eukaryotic 7TMICs share little-to-no recognized sequence similarity. It is therefore unclear how many 7TMICs exist—and whether 7TMIC structure is homologous or convergent—across taxa. We thus sought to build a new pipeline for remote homolog detection and sequence/structure analysis, aiming to resolve the evolutionary history of 7TMICs.

Insect Ors and Grs have high structural similarity despite exceptional sequence dissimilarity

Comparisons of the AlphaFold models of *Drosophila melanogaster* Ors and Grs exemplifies the discordance between sequence and structure similarity: these proteins have on average ~13% pairwise amino acid sequence identity (Figure 1C, y axis), placing them at the border of the "twilight zone" (10%–40% sequence identity)[1] and "midnight zone" (<10% sequence identity)[2] of sequence alignment. By contrast, pairwise comparisons of the corresponding AlphaFold structures—here, by Dali[26,27]—reveal that all pairs fall within the "safe zone" of structural alignments, indicating high statistical confidence in their similarity (Figure 1C, x axis). When visualized as a sequence similarity network (produced by all-to-all BLASTP), *D. melanogaster* 7TMICs segregate into several non-contiguous clusters (Figure S1A). This analysis demonstrates that no single receptor protein can be used to identify all others via simple sequence-based searches. However, structure-based search strategies (e.g., Dali, Figure S1B) are capable of densely networking these proteins. These observations emphasize how structure-based screens are a greatly superior way to search for distant homologs across more phylogenetically diverse species.[3]

A pipeline for identifying extremely distant protein homologs

Foldseek—a new tool for structure-based protein comparisons—operates orders of magnitude faster than Dali and other structural alignment tools, making large protein homolog screens feasible.[5] We first benchmarked Foldseek on *D. melanogaster* 7TMICs. When forced to compare the AlphaFold model of Orco to all other 7TMICs of this species, Foldseek produced structural similarity scores that correlate with Dali *Z* scores (Figure S1C). As proof of concept for screening, we used the AlphaFold model of *D. melanogaster* Orco to survey the AlphaFold structural proteome of *D. melanogaster* (Figure 1D). Foldseek recovered all *D. melanogaster* 7TMICs, except PHTF; thus, the method can result in false negatives. However, with the most permissive settings—which would allow the most sensitive homolog detection—Foldseek also had an extremely high false positive rate (73.3%, gray points; Figure 1D), and some true positives (e.g., Grls) had worse E values and/or lower percent sequence identity than false positives. As we were interested in identifying possibly divergent 7TMIC homologs across much longer evolutionary distances than between Ors/Grs, we recognized that neither E value nor sequence identity could serve as an effective threshold. These benchmarks illustrated the need for additional search and validation steps to minimize both false positive and false negatives.

We therefore implemented Foldseek as part of a screening and validation pipeline, with the goal of determining the presence or absence of 7TMICs across the tree of life (Figures 1E and S1D). This pipeline first uses Foldseek to search for structurally similar models in the AlphaFold Protein Structure Database, which currently consists of ~200,000,000 models from ~1,200,000 species (see STAR Methods for details on exclusions). After structural validation, it employs PSI-BLAST in a sequence-based screen, providing structurally informed access to >400,000,000 sequences—with diverse proteomic, transcriptomic, genomic, and metagenomic origins—that might not have a corresponding protein model. This second step also allows for the identification of proteins with models that were missed in the first structure-based screen, which we expected to occur due to the occurrence of false negatives at hypothetically vast evolutionary distances (e.g., Orco to PHTF).

As false positives can have high scores, and as some public data can be incomplete or of low quality, we implemented several verification steps to extract true hits. For proteins identified by structural model, we: (1) curated proteins based on membrane topology as predicted by the protein language model DeepTMHMM[28]; (2) confirmed structural alignments using Dali; and (3) visually inspected putative hits for the previously

Figure 1. A structure- and sequence-based screen for the identification and validation of extremely distant 7TMIC homologs

(A) Left: top and side views of the cryo-EM structure of the *A. bakeri* Orco homotetramer (PDB: 6C70), with one subunit colored.[20] Right: transmembrane prediction of *A. bakeri* Orco by DeepTMHMM and Phobius illustrating the characteristic membrane topology of 7TMICs. A cartoon representation of 7TMIC membrane topology is shown below.

(B) The cryo-EM structure of *A. bakeri* Orco (PDB: 6C70) and the AlphaFold model of *D. melanogaster* Orco (UniProt/AlphaFold Sequence Database: Q9VNB5). Left: side view of individual monomers. Right: top and side view of aligned monomers.

(C) Sequence identity versus structural similarity for all pairwise comparisons of AlphaFold models of *D. melanogaster* Ors and Grs, using Dali. The cluster of dots at the top right are self-to-self comparisons and isoforms of the same gene.

(D) Proof-of-principle Foldseek screen of the AlphaFold structural proteome of *D. melanogaster*, with results of the screen plotted by Foldseek-derived sequence identity and E values. The color coding is the same as in Figures S1A–S1C and can be matched to exact Hex codes in the supplemental information.

(E) Outline of the screen and validation pipeline.

(F) Cladogram of taxa in which 7TMICs were identified.

(G) All-to-all BLASTP network of 7TMICs (each represented by a dot), which visualizes only pairwise sequence similarity. Several clusters form, suggesting monophyly within clusters (annotated manually, based on CLANS clustering). At presumed longer evolutionary distances, 7TMICs show little-to-no pairwise sequence identity, represented by the weak connectivity (i.e., few edges) between most clusters.

See also Figure S1.

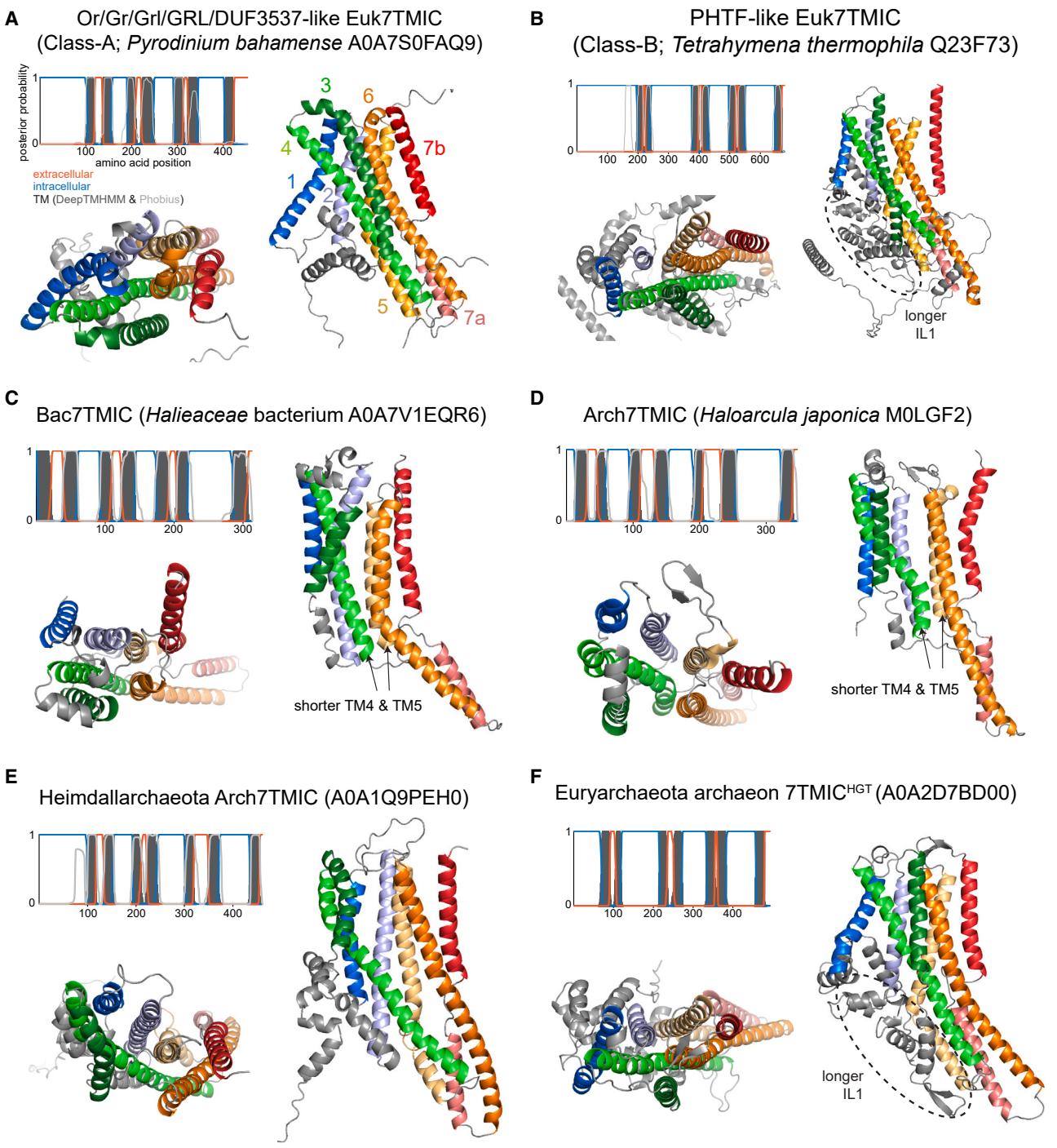


Figure 2. Examples of newly identified 7TMICs

Transmembrane predictions, and top and side views of the AlphaFold structure of newly identified 7TMICs.

(A) Representative example of a Gr- and DUF3537-like Euk7TMIC, subsequently phylogenetically classified as Class-A (Figure 4). These proteins have all of the stereotyped 7TMIC features.

(B) Representative example of a PHTF-like Class-B 7TMIC (Figure 4). These proteins have stereotyped 7TMIC features, with the addition of a long intracellular loop between TM2 and TM3 (IL1).

(legend continued on next page)

described 7TMIC features. Proteins identified through sequence similarity were curated based on predicted membrane topology (DeepTMHMM).

7TMICs are present across the tree of life

Our screen recovered thousands of previously unidentified 7TMICs spanning the tree of life (Figures 1F, 1G, and S1E). The hits not only include new eukaryotic 7TMICs (Euk7TMICs) but also sequences from all major branches of bacteria (Bac7TMICs) and archaea (Arch7TMICs) (see "protein nomenclature" in STAR Methods). These proteins come from several obviously monophyletic clades, apparent as clusters in a network representing all-to-all BLASTP searches (Figure 1G). However, these clades generally exhibit very little pairwise sequence similarity, represented by few edges between clusters (Figure 1G).

Euk7TMICs could be visually sorted into two structure types: Or/Gr/Grl/GRL/DUF3537-like (Figure 2A), having the canonical insect Or-like fold; or PHTF-like (Figure 2B), having the same core structure, but with a long first IL (IL1). Although the various prokaryotic 7TMICs have a striking degree of structural similarity to Euk7TMICs (Figures 2C–2E), we observed that they generally had shorter TM4 and TM5 helices, which constitute a component of the anchor domain in insect Ors (Figure 1A). Heimdallarchaeota 7TMICs (Figure 2E) were an exception: their overall tertiary structure appeared eukaryote-like. This qualitative similarity (supported by subsequent quantitative analyses, described below) is notable, as Heimdallarchaeota are proposed to be the most closely related extant archaea to eukaryotes.[29–33] In addition, a small number of metagenomically identified prokaryotic 7TMICs have Euk7TMIC-like folds (Figure 2F). Importantly, these show high sequence similarity to Euk7TMICs (green nodes in the eukaryotic PHTF-like cluster; Figure 1G), suggesting that these sequences are the result of eukaryote-to-prokaryote horizontal gene transfer (HGT), a hypothesis further supported phylogenetically (see below).

7TMICs have a shared tertiary structure and amino acid sequence profile, supporting homology

Although we observed structural similarities between the proteins our screen identified, it remained unclear whether these sequences were homologous or whether they represent cases of structural convergence. To address this fundamental issue, we adapted established protein comparison tools into a graph-based approach for inferring homology. For protein structures, we calculated all-to-all template modeling (TM) scores, where those >0.5 indicate high statistical confidence of fold similarity.[34] For protein sequences, we performed all-to-all PSI-BLAST searches; PSI-BLAST builds iterative multiple sequence alignments, thereby identifying distant homologs by family-wise sequence profiles, rather than by simple pairwise sequence similarities.[35] In essence, PSI-BLAST networking is equivalent to performing PSI-BLAST searches, starting with every 7TMIC as a query. For both methods, one expects homologous proteins to form bi-directional connections between each other (i.e., pairs will be reciprocal hits) and that homologous families will be highly interconnected, thereby collapsing into visually identifiable clusters in structure and sequence space. We performed these analyses with Type-I and Type-II opsins as control groups, as these large families are 7-transmembrane domain proteins (unrelated to 7TMICs) that adopt highly similar folds to one another, despite no recognized sequence similarity.[36]

In the structural similarity network, 7TMICs formed a densely connected linkage cluster, disconnected from a unified opsin linkage cluster (Figure 3A). 7TMICs also clustered in sequence space—after 3 PSI-BLAST iterations, 7TMICs collapsed into a highly connected community structure (Figures 3B and S2A). In stark contrast, the opsins separated into distinct Type-I and Type-II community structures, demonstrating that structure and sequence are not necessarily linked (Figure 3B). Although there were connections between 7TMICs and the opsins, in the third iteration these constituted only 18 of 1,117,609 connections (0.0016%), almost certainly representing spurious similarity. A small minority of 7TMICs (33/2,421 representative sequences) from diverse eukaryotic taxa showed no connectivity to the core 7TMIC cluster in the second iteration and weak connectivity in the third; these may be extremely rapidly evolving proteins and/or cases of independent structural convergence.

We next sought to determine which, if any, regions of 7TMICs are conserved. It was previously observed that insect Ors display the highest conservation in the anchor domain and pore, with greater divergence in the N-terminal region that forms the odor-binding pocket.[20,21] We calculated sequence embedding-based conservation scores, which identify sites that are evolutionarily constrained.[37] This analysis elucidated a similar conservation pattern for newly identified 7TMICs. Although absolute amino acid sequence identity is low (averaging 15% across sites; Figures 3C and S2B), embedding-based conservation analysis revealed that the most highly conserved regions are in three locations: the hypothetical anchor domain (intracellular sequences spanning TM4-TM5 and TM6-TM7a); the hypothetical pore (TM7b); and TM5-TM6, which form lateral ion permeation conduits in Ors[20,21] (Figures 3D and 3F).

We next used the transformer model PeSTo[38] to predict protein-protein interactions in 7TMICs, revealing two conserved regions (Figures 3E and 3F). The first was N-terminal, corresponding to the re-entrant loop (TM0); this region has an important, albeit poorly understood, function in Orco.[25] The second region was in the hypothetical anchor domain and pore.

These findings are not biased by the inclusion of proteins previously determined to be homologous (insect Ors/Grs and animal Grls); on the contrary, removing these sequences improved average conservation and interaction scores (Figures 3D, 3E, and S2C).

(C–E) Representative examples of a Bac7TMIC and distinct types of Arch7TMICs. When clustered by Foldseek at 90% coverage (data not shown), prokaryotic 7TMICs form three structure clusters (represented here in C–E), although they cannot be easily distinguished visually. These proteins all share the stereotyped 7TMIC features but, with the exception of Heimdallarchaeota 7TMICs (E), have shorter TM4 and TM5.

(F) Representative example of the small number of prokaryotic proteins with high fold and sequence similarity to Euk7TMICs, which are presumed to have arisen through HGT (Figure 4).

See also Figure S1.

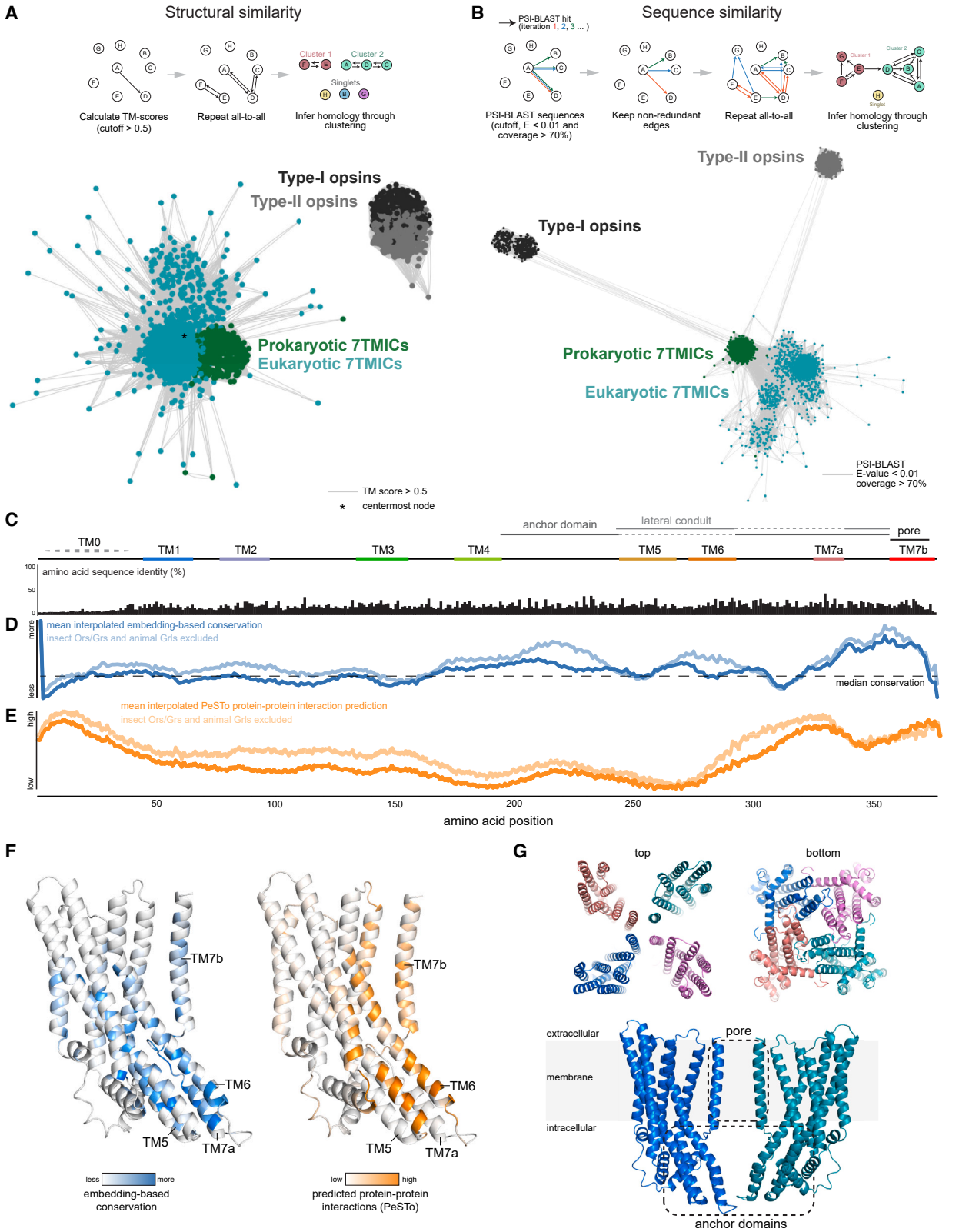


Figure 3. Evidence for 7TMIC homology through structural and sequence similarity

(A) 7TMIC structural similarity network derived from all-to-all TM-scores (schematized at the top). The asterisk marks the most central node, used for visualizations in (C)–(G).

(legend continued on next page)

Although we cannot know *a priori* whether these proteins form tetramers like insect Ors, these patterns of conservation and predicted protein-protein interactions suggests that they may assemble as multimers using the same domains. Consistent with this idea, using AlphaFold-multimer[39–42] to predict complexes of tetramers resulted in quaternary structures with striking similarity to experimentally derived Or structures (Figures 3G and S2D). In these models, the hypothetical anchor domain (particularly TM7a) contains the closest protein-protein interactions and TM7b lines the putative pore.

These results quantitatively demonstrate that 7TMICs have a common predicted structure, a shared sequence profile, and similar patterns of sequence conservation. Thus, the most parsimonious hypothesis is that 7TMICs are a homologous protein superfamily.

The evolutionary history of 7TMICs

Having obtained evidence for the homology of 7TMICs, we next sought to elucidate their evolutionary history. As we expected pairwise sequence dissimilarity would make sequence alignments difficult, we performed sequence-based phylogenetics on an ensemble of alignments, thus resulting in a "forest" of topologically diverse phylogenetic trees (see Figures S3A–S3E for full analysis). We identified two Euk7TMIC families, termed Class-A (frequently monophyletic in tree space, and including animal Ors/Grs/Grls, plant DUF3537, holozoan GrlHz, and various unicellular Euk7TMICs) and Class-B (variably mono- and paraphyletic PHTF-like proteins from diverse taxa, including a small number of bacterial and archaeal proteins; see Figures 1G and 2F). The median sampled tree (Figure 4A) generally represents this diverse tree space: kinetoplastid 7TMICs branch proximally to Arch7TMICs/Bac7TMICs (here deeply, but with low branch support; 0.79 and 0.409 for the two most proximal branches), consistent with the hypothesis that kinetoplastids (and allies, Discoba) split early in eukaryotic evolution[43]; Class-A is monophyletic, with modestly strong branch support (0.91); Class-B is paraphyletic, but with extremely low branch support on the relevant branch (0.22); and, finally, a small number of prokaryotic proteins branch proximally to Class-B Euk7TMICs, suggesting eukaryote-to-prokaryote HGT.

To complement the sequence-based phylogenetic approach, we also employed a recently developed structure-based phylogenetic method, which infers a tree from Foldseek structural alignments.[44] We made three notable observations of the resulting tree (Figure 4B), which shared many similarities to the sequence-based phylogenies (Figure 4A). First, the prokaryotic branch most proximal to the Euk7TMICs included Heimdallarchaeota 7TMICs (Figure 4B, asterisk), consistent with their proposed relation to eukaryotes, and suggesting that the shared Euk7TMIC structure (i.e., longer TM4 and TM5; Figure 2) may have emerged before eukaryogenesis. Second, kinetoplastid 7TMICs were placed as the sister clade to all other Euk7TMICs, consistent with their presumed early branching. Third, Class-B Euk7TMICs were essentially monophyletic.

The most parsimonious interpretation of these data is that the Class-A/Class-B split is the result of a gene duplication that occurred after eukaryogenesis but before the speciation event(s) leading to Amorphea and Diaphoretickes (Figures 4C and 4D).

Concluding remarks

We have described a structure- and sequence-based screening strategy for identifying extremely distant transmembrane protein homologs, revealing that 7TMICs are present across the tree of life, including novel discoveries of representatives in Bacteria and Archaea. We have also shown that, despite substantial pairwise sequence dissimilarity, 7TMICs have extremely high predicted structural similarity and identifiable family-wise sequence similarity. Together, our results provide the first strong evidence that these disparate proteins form a single, homologous superfamily. This finding contrasts with the Type-I and Type-II opsins, whose structural similarity to each other might represent a case of convergent evolution.[36,45] Despite the phylogenetic breadth and conserved predicted structure of 7TMICs, our knowledge of their function is almost entirely restricted to a subset of insect proteins,[46,47] which represent only a single, insect-specific lineage of this family. Our work lays a foundation for the experimental determination of 7TMIC structure and the analysis of the presumably diverse functions of 7TMICs across a wide range of species. Moreover, we suspect that this ancient and cryptic

(B) 7TMIC sequence similarity network produced by all-to-all PSI-BLAST searches (schematized at the top; iteration 3 is shown), providing evidence that 7TMICs have a family-wide sequence profile. This pattern of strong, bi-directional linkages became apparent already in PSI-BLAST iteration 2, while subsequent iterations resembled iteration 3 (Figure S2A; supplemental information).

(C) Amino acid sequence identity derived from a query-centered Foldseek alignment for the centermost node in the structural similarity network (*Symbiodium natans,* UniProt: A0A812K102). Transmembrane predictions are from Phobius. TM7b was annotated manually. TM0 (the re-entrant loop) is indicated with a dashed line, as it is inconsistently predicted by DeepTMHMM and Phobius, and often predicted within a low confidence region in AlphaFold models. Query-centered alignments for all 7TMIC models analyzed here are available in the supplemental information.

(D) Average sequence-embedding-based conservation scores for 7TMICs, with the curves interpolated to match the length of A0A812K102. Column conservation scores are significantly correlated with column sequence identity for A0A812K102 (Figure S2B). The location and strength of conservation likely varies by 7TMIC family and subfamily. Excluding the well-established 7TMICs (insect Ors/Grs and animal Grls) led to overall increased conservation scores (light blue line, also Figure S2C). Embedding-based conservation scores for all 7TMIC models analyzed here are available in the supplemental information.

(E) Average PeSTo protein-protein interaction predictions. The region with the most consistently predicted protein-protein interactions is near the C terminus, correlating with the site of highest sequence conservation; again, exclusion of Ors/Grs/Grls led to higher prediction scores (light orange line). PeSTo predictions for all 7TMIC models analyzed here are available in the supplemental information.

(F) Sequence-embedding-based conservation scores (blue) and PeSTo-derived protein-protein interaction scores (orange) mapped onto the AlphaFold model of A0A812K102.

(G) Top: top (presumed extracellular) and bottom (presumed intracellular) views of a hypothetical tetramer of A0A812K102 (predicted by AlphaFold-multimer), showing that individual subunits have their closest interactions in the pore and anchor domains, similar to Ors.[20,21] Bottom: side view of the A0A812K102 tetramer, with two subunits masked for clarity. In total, we modeled 85 tetramers; 83 of these were Or-like in that they displayed rotational symmetry, with the closest interactions in the hypothetical anchor and pore regions (further examples in Figure S2D and the supplemental information).

See also Figure S2.

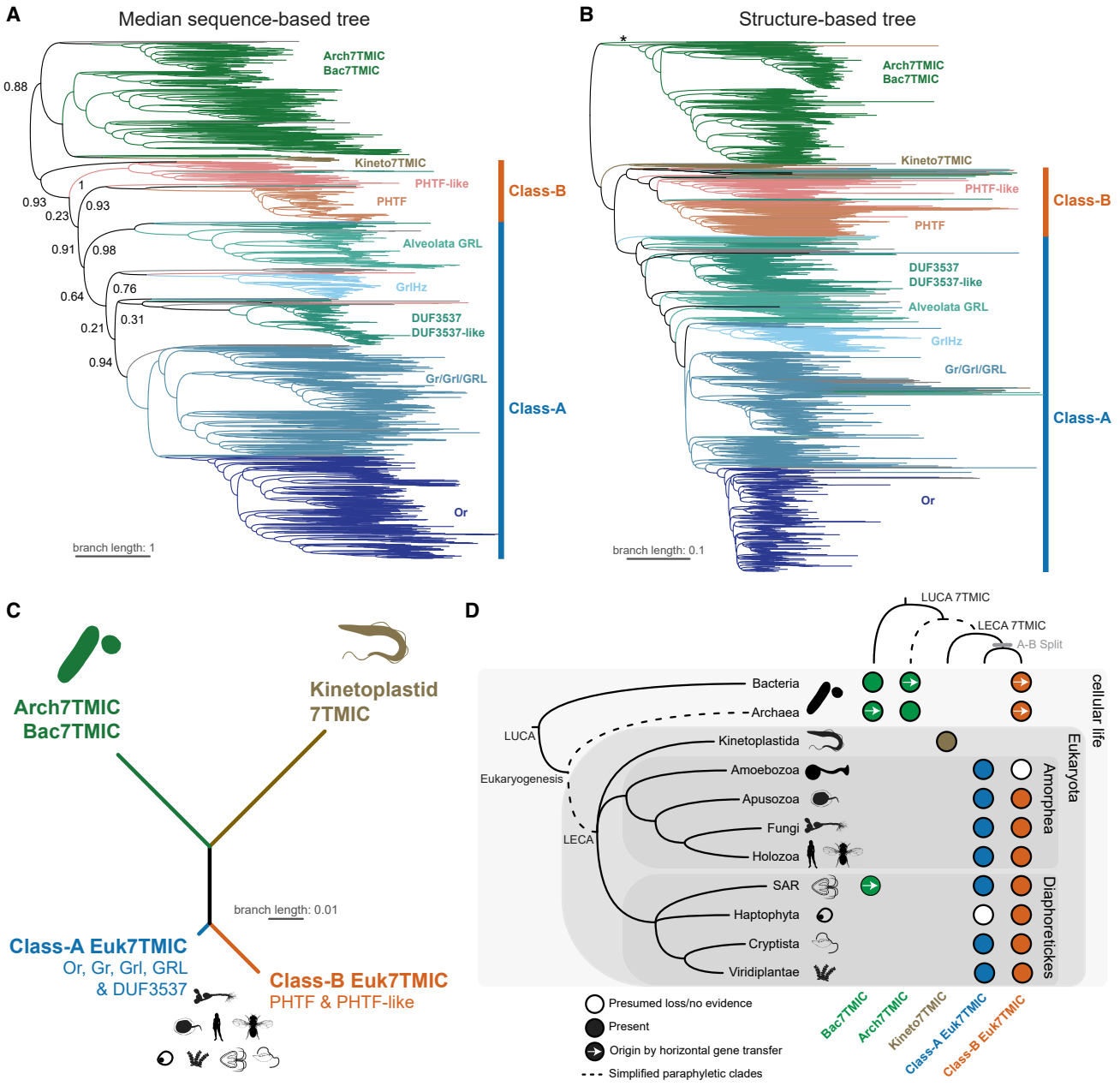


Figure 4. A model for the evolution of the 7TMIC superfamily

(A) The median phylogenetic tree of 7TMICs sampled from the Robinson-Foulds-based tree space of 48 sequence-based phylogenetic trees. For visualization purposes, the tree is arbitrarily rooted in the last common ancestor of all Arch7TMICs/Bac7TMICs (which are highly reticulated); the true root is likely at the unidentified location of the last universal common ancestor (LUCA) within the prokaryotic branch. Branch lengths are derived from the average number of substitutions per site. Tree space is visualized in Figures S3D and S3E, and all trees and alignments are available in the supplemental information.

(B) TM-score-based structural tree of 7TMICs derived from Foldtree, automatically rooted using the minimal ancestor deviation (MAD) method. As in (A), the true root is likely at the unidentified location of LUCA. Branch lengths are derived from the underlying distance matrix of TM-scores. The asterisk marks the branch containing Heimdallarchaeota 7TMICs; a fully annotated tree is available in the supplemental information.

(C) Collapsed version of the tree in (B) highlighting the major branching patterns. Kinetoplastid 7TMICs likely branched early, while the Class-A/Class-B split occurred after the emergence of the last eukaryotic common ancestor (LECA) but before the split(s) leading to Amorphea and Diaphoretickes.

(D) Summary of the results of the screen and evolutionary analyses. The left tree shows assumed relationships between the various taxa in which 7TMICs were identified, while the top tree shows the evolutionary history of 7TMICs themselves. The colored dots represent the presence or absence of 7TMIC families. At the subfamily level, many Class-B PHTF-like proteins may be the result of HGT, as there is broad but sparse taxonomic diversity within this putative subfamily. Note that this screen did not recover previously identified 7TMICs from Amoebozoa or Chytridiomycota, which were inferred to be Class-A, based on previously described sequence similarity to insect Grs/Ors.[8] In addition, "Fungi" only refers to Chytridiomycota and Blastocladiomycota.

See also Figure S3 and Table S1.

superfamily is only one of many that wait to be discovered in the depths of the twilight zone of sequence space.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Structural screen and validation
  - Sequence-based homolog identification
  - *Ab initio* protein folding and structural analyses
  - Network and conservation analyses
  - Phylogenetics
  - Protein nomenclature
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.10.008.

ACKNOWLEDGMENTS

We thank Matteo Dal Peraro, Lucien Krapp, and members of the Benton laboratory for comments on the manuscript and Christophe Dessimoz for support of this work. Silhouettes throughout were sourced from PhyloPic (www.phylopic.org/). N.J.H. is supported by a Human Frontier Science Program Long-Term Postdoctoral Fellowship (LT-0003/2022L). D.M. is supported by a Swiss National Science Foundation grant (216623) to Christophe Dessimoz. Research in R.B.'s laboratory is supported by the University of Lausanne, an ERC Advanced Grant (833548), and the Swiss National Science Foundation (310030B-185377).

AUTHOR CONTRIBUTIONS

Conceptualization, N.J.H. and R.B.; methodology, N.J.H.; software, N.J.H. and D.M.; validation, N.J.H. and D.M.; formal analysis, N.J.H. and D.M.; investigation, N.J.H.; resources, N.J.H.; data curation, N.J.H. and D.M.; writing – original draft, N.J.H.; writing – review & editing, N.J.H., D.M., and R.B.; supervision, R.B.; project administration, N.J.H. and R.B.; funding acquisition, N.J.H. and R.B. D.M. designed and performed the Foldtree analysis. N.J.H. performed all other formal analyses.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. We support inclusive, diverse, and equitable conduct of research.

Received: September 1, 2023
Revised: September 26, 2023
Accepted: October 6, 2023
Published: October 31, 2023

REFERENCES

1. Doolittle, R.F. (1986). Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences (University Science Books).
2. Rost, B. (1997). Protein structures sustain evolutionary drift. Fold. Des. *2*, S19–S24. https://doi.org/10.1016/S1359-0278(97)00059-X.
3. Illergård, K., Ardell, D.H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. Proteins *77*, 499–508. https://doi.org/10.1002/prot.22458.
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
5. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01773-0.
6. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.
7. Benton, R. (2015). Multigene family evolution: perspectives from insect chemoreceptors. Trends Ecol. Evol. *30*, 590–600. https://doi.org/10.1016/j.tree.2015.07.009.
8. Benton, R., Dessimoz, C., and Moi, D. (2020). A putative origin of the insect chemosensory receptor superfamily in the last common eukaryotic ancestor. eLife *9*, e62507. https://doi.org/10.7554/eLife.62507.
9. Robertson, H.M. (2015). The insect chemoreceptor superfamily is ancient in animals. Chem. Senses *40*, 609–614. https://doi.org/10.1093/chemse/bjv046.
10. Benton, R., and Himmel, N.J. (2023). Structural screens identify candidate human homologs of insect chemoreceptors and cryptic *Drosophila* gustatory receptor-like proteins. eLife *12*, e85537. https://doi.org/10.7554/eLife.85537.
11. Joseph, R.M., and Carlson, J.R. (2015). *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain. Trends Genet. *31*, 683–695. https://doi.org/10.1016/j.tig.2015.09.005.
12. Robertson, H.M. (2019). Molecular evolution of the major arthropod chemoreceptor gene families. Annu. Rev. Entomol. *64*, 227–242. https://doi.org/10.1146/annurev-ento-020117-043322.
13. Clyne, P.J., Warr, C.G., Freeman, M.R., Lessing, D., Kim, J., and Carlson, J.R. (1999). A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. Neuron *22*, 327–338. https://doi.org/10.1016/S0896-6273(00)81093-4.
14. Clyne, P.J., Warr, C.G., and Carlson, J.R. (2000). Candidate taste receptors in *Drosophila*. Science *287*, 1830–1834. https://doi.org/10.1126/science.287.5459.1830.
15. Gao, Q., and Chess, A. (1999). Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. Genomics *60*, 31–39. https://doi.org/10.1006/geno.1999.5894.
16. Scott, K., Brady, R., Cravchik, A., Morozov, P., Rzhetsky, A., Zuker, C., and Axel, R. (2001). A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. Cell *104*, 661–673. https://doi.org/10.1016/S0092-8674(01)00263-X.
17. Vosshall, L.B., Amrein, H., Morozov, P.S., Rzhetsky, A., and Axel, R. (1999). A spatial map of olfactory receptor expression in the *Drosophila* antenna. Cell *96*, 725–736. https://doi.org/10.1016/s0092-8674(00)80582-6.
18. Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., et al. (2002). Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science *298*, 149–159. https://doi.org/10.1126/science.1077061.

19. Saina, M., Busengdal, H., Sinigaglia, C., Petrone, L., Oliveri, P., Rentzsch, F., and Benton, R. (2015). A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. Nat. Commun. *6*, 6243. https://doi.org/10.1038/ncomms7243.

20. Butterwick, J.A., del Mármol, J., Kim, K.H., Kahlson, M.A., Rogow, J.A., Walz, T., and Ruta, V. (2018). Cryo-EM structure of the insect olfactory receptor Orco. Nature *560*, 447–452. https://doi.org/10.1038/s41586-018-0420-8.

21. del Mármol, J., Yedlin, M.A., and Ruta, V. (2021). The structural basis of odorant recognition in insect olfactory receptors. Nature *597*, 126–131. https://doi.org/10.1038/s41586-021-03794-8.

22. Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L.B., and Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated ion channels. Nature *452*, 1002–1006. https://doi.org/10.1038/nature06850.

23. Wicher, D., Schäfer, R., Bauernfeind, R., Stensmyr, M.C., Heller, R., Heinemann, S.H., and Hansson, B.S. (2008). *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. Nature *452*, 1007–1011. https://doi.org/10.1038/nature06861.

24. Benton, R., Sachse, S., Michnick, S.W., and Vosshall, L.B. (2006). Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors in vivo. PLoS Biol. *4*, e20. https://doi.org/10.1371/journal.pbio.0040020.

25. Hopf, T.A., Morinaga, S., Ihara, S., Touhara, K., Marks, D.S., and Benton, R. (2015). Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat. Commun. *6*, 6077. https://doi.org/10.1038/ncomms7077.

26. Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics *24*, 2780–2781. https://doi.org/10.1093/bioinformatics/btn507.

27. Holm, L., and Park, J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics *16*, 566–567. https://doi.org/10.1093/bioinformatics/16.6.566.

28. Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Armenteros, J.J.A., Marcatili, P., Nielsen, H., Krogh, A., and Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. Preprint at bioRxiv. https://doi.org/10.1101/2022.04.08.487609.

29. Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., and Embley, T.M. (2008). The archaebacterial origin of eukaryotes. Proc. Natl. Acad. Sci. USA *105*, 20356–20361. https://doi.org/10.1073/pnas.0810647105.

30. Eme, L., Tamarit, D., Caceres, E.F., Stairs, C.W., De Anda, V., Schön, M.E., Seitz, K.W., Dombrowski, N., Lewis, W.H., Homa, F., et al. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature *618*, 992–999. https://doi.org/10.1038/s41586-023-06186-2.

31. Liu, Y., Makarova, K.S., Huang, W.C., Wolf, Y.I., Nikolskaya, A.N., Zhang, X., Cai, M., Zhang, C.J., Xu, W., Luo, Z., et al. (2021). Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature *593*, 553–557. https://doi.org/10.1038/s41586-021-03494-3.

32. Spang, A., Eme, L., Saw, J.H., Caceres, E.F., Zaremba-Niedzwiedzka, K., Lombard, J., Guy, L., and Ettema, T.J.G. (2018). Asgard archaea are the closest prokaryotic relatives of eukaryotes. PLoS Genet. *14*, e1007080. https://doi.org/10.1371/journal.pgen.1007080.

33. Williams, T.A., Foster, P.G., Cox, C.J., and Embley, T.M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. Nature *504*, 231–236. https://doi.org/10.1038/nature12779.

34. Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895. https://doi.org/10.1093/bioinformatics/btq066.

35. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

36. Rozenberg, A., Inoue, K., Kandori, H., and Béjà, O. (2021). Microbial rhodopsins: the last two decades. Annu. Rev. Microbiol. *75*, 427–447. https://doi.org/10.1146/annurev-micro-031721-020452.

37. Yeung, W., Zhou, Z., Li, S., and Kannan, N. (2023). Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. Brief. Bioinform. *24*, bbac599. https://doi.org/10.1093/bib/bbac599.

38. Krapp, L.F., Abriata, L.A., Cortés Rodriguez, F., and Dal Peraro, M. (2023). PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. Nat. Commun. *14*, 2175. https://doi.org/10.1038/s41467-023-37701-8.

39. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., S., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments oding. Nucleic Acids Res. *45*, D170–D176. https://doi.org/10.1093/nar/gkw1081.

40. Mirdita, M., Steinegger, M., and Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics *35*, 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057.

41. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making Protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

42. Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. *48*, D570–D578. https://doi.org/10.1093/nar/gkz1035.

43. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The new tree of eukaryotes. Trends Ecol. Evol. *35*, 43–55. https://doi.org/10.1016/j.tree.2019.08.008.

44. Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. Preprint at bioRxiv. https://doi.org/10.1101/2023.09.19.558401.

45. Larusso, N.D., Ruttenberg, B.E., Singh, A.K., and Oakley, T.H. (2008). Type II opsins: evolutionary origin by internal domain duplication? J. Mol. Evol. *66*, 417–423. https://doi.org/10.1007/s00239-008-9076-6.

46. Benton, R. (2022). *Drosophila* olfaction: past, present and future. Proc. Biol. Sci. *289*, 20222054. https://doi.org/10.1098/rspb.2022.2054.

47. Chen, Y.-C.D., and Dahanukar, A. (2020). Recent advances in the genetic basis of taste detection in *Drosophila*. Cell. Mol. Life Sci. *77*, 1087–1101. https://doi.org/10.1007/s00018-019-03320-0.

48. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

49. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

50. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

51. Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics *20*, 3702–3704. https://doi.org/10.1093/bioinformatics/bth444.

52. Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., and Alva, V. (2020). Protein sequence analysis using the MPI bioinformatics toolkit. Curr. Protoc. Bioinformatics *72*, e108. https://doi.org/10.1002/cpbi.108.

53. Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. *430*, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

54. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501. https://doi.org/10.1107/S0907444910007493.

55. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software

environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

56. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

57. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

58. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. https://doi.org/10.1093/nar/gkab301.

59. Edgar, R.C. (2022). Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat. Commun. *13*, 6968. https://doi.org/10.1038/s41467-022-34630-w.

60. Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2004). A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. *338*, 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016.

61. Käll, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. *35*, W429–W432. https://doi.org/10.1093/nar/gkm256.

62. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

63. Aberer, A.J., Krompass, D., and Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and Webservice. Syst. Biol. *62*, 162–166. https://doi.org/10.1093/sysbio/sys078.

64. Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour. *17*, 1385–1392. https://doi.org/10.1111/1755-0998.12676.

65. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

66. Holm, L. (2020). Using Dali for protein structure comparison. In Structural Bioinformatics Methods in Molecular Biology, Z. Gáspári, ed. (Springer), pp. 29–42. https://doi.org/10.1007/978-1-0716-0270-6_3.

67. Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One *4*, e4345. https://doi.org/10.1371/journal.pone.0004345.

68. Brand, P., Robertson, H.M., Lin, W., Pothula, R., Klingeman, W.E., Jurat-Fuentes, J.L., and Johnson, B.R. (2018). The origin of the odorant receptor gene family in insects. eLife *7*, e38340. https://doi.org/10.7554/eLife.38340.

69. Robertson, H.M., Warr, C.G., and Carlson, J.R. (2003). Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *100*, 14537–14542. https://doi.org/10.1073/pnas.2335847100.

70. Bulzu, P.-A., Andrei, A.-Ş., Salcher, M.M., Mehrshad, M., Inoue, K., Kandori, H., Beja, O., Ghai, R., and Banciu, H.L. (2019). Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. Nat. Microbiol. *4*, 1129–1137. https://doi.org/10.1038/s41564-019-0404-y.

71. Ramirez, M.D., Pairett, A.N., Pankey, M.S., Serb, J.M., Speiser, D.I., Swafford, A.J., and Oakley, T.H. (2016). The last common ancestor of most bilaterian animals possessed at least nine opsins. Genome Biol. Evol. *8*, 3640–3652. https://doi.org/10.1093/gbe/evw248.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
All data	This paper	https://doi.org/10.5061/dryad.fqz612jz9
Software and algorithms		
AlphaFold Protein Structure Database	Jumper et al.[4]; Varadi et al.[6]	https://alphafold.ebi.ac.uk
ape	Paradis and Schliep[48]	http://ape-package.ird.fr
BLAST+	Altschul et al.[35]	https://www.ncbi.nlm.nih.gov
CD-HIT	Fu et al.[49]; Li and Godzik[50]	http://cd-hit.org
CLANS	Frickey and Lupas[51]; Gabler et al.[52]; Zimmermann et al.[53]	https://toolkit.tuebingen.mpg.de/tools/clans
ColabFold	Mirdita et al.[39–41]; Mitchell et al.[42]	https://github.com/sokrypton/ColabFold
Coot	Emsley et al.[54]	https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot
Cytoscape	Shannon et al.[55]	https://cytoscape.org
DeepTMHMM	Hallgren et al.[28]	https://dtu.biolib.com/DeepTMHMM
esm2_t33_650M_UR50D	Lin et al.[56]	https://huggingface.co/facebook/esm2_t33_650M_UR50D
FastTree 2.1	Price et al.[57]	http://www.microbesonline.org/fasttree
Foldtree	Moi et al.[44]	https://github.com/DessimozLab/fold_tree
Foldseek	van Kempen et al.[5]	https://github.com/steineggerlab/foldseek
Illustrator	Adobe Inc.	https://www.adobe.com/products/illustrator.html
iTol	Letunic and Bork[58]	https://itol.embl.de
JASP	JASP Team	https://jasp-stats.org
Muscle5	Edgar[59]	https://www.drive5.com/muscle
PeSTo	Krapp et al.[38]	https://github.com/LBM-EPFL/PeSTo
Phobius	Käll et al.[60,61]	https://phobius.sbc.su.se
phytools	Revell[62]	https://github.com/liamrevell/phytools
PyMol	Schrödinger, LLC.	https://pymol.org
RogueNaRok	Aberer et al.[63]	https://github.com/aberer/RogueNaRok
R studio	R Studio Team	http://www.rstudio.com
Sequence conservation scripts	Yeung et al.[37]	https://github.com/esbgkannan/kibby
treespace	Jombart et al.[64]	https://github.com/thibautjombart/treespace

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
trimal	Capella-Gutiérrez et al.[65]	http://trimal.cgenomics.org/trimal

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Richard Benton (richard.benton@unil.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data reported in this paper have been deposited in Dryad (https://doi.org/10.5061/dryad.fqz612jz9) and are publicly available as of the date of the publication.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Structural screen and validation

Proof-of-concept screens were carried out using local implementations of Foldseek[5] and DaliLite.[26,27] Subsequent structure-based screens of the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/)[4,6] were performed on the Foldseek server (https://search.foldseek.com/), using the following query structures/models: *A. bakeri* Orco (PDB: 6C70); *M. hrabei* Or5 (PDB: 7LIC); *D. melanogaster* GrlHz (Uniprot: Q9W1W8); *B. belcheri* GrlHz (UniProt: A0A6P5ACQ6); *T. adhaerens* GrlHz (UniProt: B3RTY0); *Z. mays* DUF3537 proteins (UniProt: A0A1D6LEW8, B4FJ88 and B6SUZ0); *P. patens* DUF3537 (UniProt: A0A2K1ICX7, A0A2K1JKU0 and A0A2K1L324); *D. melanogaster* Phtf (UniProt: Q9V9A8); *H. sapiens* PHTF1 (UniProt: Q9UMS5) and PHTF2 (UniProt: Q8N3S3); *P. halstedii* PHTF (UniProt: A0A0P1B782); *L. infantum* GRL1 (UniProt: A4HWQ9); and *T. brucei brucei* GRL1 (UniProt: Q57U78) (Figure S1D). For the initial screen, we masked the MORN repeats in trypanosome GRL1 and the long intracellular loop 1 in PHTFs, thus restricting the search to the core 7TMIC domain. We did not set a statistical threshold (E-value) for putative homolog identification. For eukaryotic hits, we initially considered all hits from the screen. For archaeal and bacterial hits, we took the more stringent approach of only further analyzing those that were hits for all the query groups (annotated on the dendrogram in Figure S3A). We did not formally screen animal or vascular land plant species because we considered that these taxa have been sufficiently screened,[7–10,19] and we were most interested in the very early evolution of 7TMICs. Indeed, preliminary Foldseek screens did not elucidate any obvious new plant- and/or animal-specific 7TMICs (data not shown).

Subsequent validation was performed in several steps. First, transmembrane topology was predicted using DeepTMHMM (the BioLib implementation at https://dtu.biolib.com/DeepTMHMM/ and a local implementation).[28] For eukaryotic hits, we further assessed all models with >6 DeepTMHMM predicted transmembrane segments. For archaea and bacteria, we took a more conservative approach and only further assessed hits with exactly 7 predicted transmembrane segments. We also used Phobius (https://phobius.sbc.su.se/)[60,61] and the transformer model PeSTo (https://pesto.epfl.ch/)[38] to predict transmembrane topology and protein-lipid interactions, respectively. Both of these tools were used for visualization, but neither was employed to quantitatively curate sequences. Second, we assessed these predictions alongside the corresponding AlphaFold structural models (visualized in PyMol), looking for: (i) 7 predicted transmembrane alpha helices; (ii) shorter extracellular than intracellular loops; (iii) an intracellular N-terminus and extracellular C-terminus; (iv) longer TM4, TM5 and TM6 helices; and (v) the exceptional "split" TM7 helix.[20,21,24,25] We did not consider the re-entrant loop (TM0) as a criterion, as it is inconsistently predicted by transmembrane prediction methods.[8,10] We caution that this visual inspection is inherently subjective in nature, and in no instance was it used as the sole piece of evidence for a true hit. Third, we used a local implementation of DaliLite to compare all remaining hits with the original query structures and three negative controls. For the negative controls we selected an Adiponectin receptor (*Homo sapiens* ADPR1; PDB: 5LXG) and a channelrhodopsin (*Chlamydomonas reinhardtii* Channelrhodopsin-2; PDB: 6EID) in advance of the screen, as both have 7 transmembrane domains but are unrelated to 7TMICs; we added the ABC transporter permease (*Escherichia coli,* UniProt: A0A061Y968) *post hoc*, as many of the screen hits were errantly annotated as ABC transporters. Only hits with Dali Z-scores >8 as compared to 7TMIC queries were further analyzed. This threshold is based on Holm's criteria,[66] where Z-scores >20 indicate definite homology and 8–20 probable homology (here, collectively the "safe zone"); 2-8 a "grey area" (here, "twilight zone"); and <2 non-significant (here, "midnight zone"). We conceptualized these scores as "protein fold similarity" in place of "homology," as we infer homology based on a holistic view of sequence, structure and taxonomic features.

It is important to note that this approach assumes high quality AlphaFold predictions, which require rich multiple sequence alignments and are improved by experimentally-derived structures in the training dataset (e.g., *A. bakeri* Orco, PDB: 6C70). Furthermore, the validation steps used here are adapted specifically for transmembrane proteins, and should be customized for the particular protein family of interest. The application of this methodology to other protein families will require mindfulness of the limitations of protein structure modelling, particularly if no experimentally-derived structures exist.

Sequence-based homolog identification

For putative eukaryotic homologs, the results of the Foldseek screen were used to select query sequences. CLANS was used to generate an all-to-all BLASTP network (E-value cutoff 0.01), which was subsequently clustered by the global network clustering option.[51–53] PSI-BLAST homolog searches were carried out using all singlets and a representative sequence from each cluster (the node with the highest neighborhood connectivity). Searches were run on the NCBI server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the clustered non-redundant (clustered_nr) sequence database, until convergence. PSI-BLAST searches were performed with an E-value cutoff of 0.05, but final candidates were selected only if they had a minimum coverage of 50% (with coverage of the transmembrane region) and a final E-value at or below 10^{-10}. For searches recovering canonical animal Grs/Ors/Grls, the PSI-BLAST searches were stopped when the top 1000 hits were recovered, as these searches quickly converged on tens of thousands of predominately insect sequences, which was computationally time-consuming and methodologically unnecessary for this study.

For potential Arch7TMIC homologs, sequence databases were likewise assembled using PSI-BLAST, using each of the structural screen hits as a query sequence. Compared to the eukaryote-based searches, we took a more stringent approach, setting an E-value cutoff of 10^{-10} for both the PSI-BLAST search and final hit selection. Query sequences that were orphans, or which had very few sequence-based homologs (<10), were excluded from further analyses. These searches recovered the Bac7TMICs, so efforts were not repeated using Bac7TMIC queries.

After preliminary homolog identification, DeepTMHMM was used to predict transmembrane topology. For all sequences identified via PSI-BLAST, we kept sequences with >6 (rather than 7) TM segments, as DeepTMHMM had previously failed to predict TM7 despite the presence of TM7 helices in the associated structural models.[10] Finally, to reduce redundancy, and thus simplify computation and presentation, CD-HIT (https://cd-hit.org)[49,50] was used to cluster sequences—first by 70% for the initial BLASTP sequence similarity network (Figure 1G), then by 50% for all subsequent analyses—keeping the longest sequence as the cluster representative.

A notable limitation of this approach is the use of metagenomic data for the identification of some prokaryotic 7TMICs. As these datasets are assembled from environmental samples, the possibility of sequence misidentification cannot be completely discounted. However, we believe it is not a serious problem, as most of the metagenomically identified sequences described here correspond to tens-to-hundreds of homologous proteins in closed prokaryotic genomes. The only obvious exceptions are the small number of archaeal and bacterial sequences most closely resembling Class-B Euk7TMICs.

Ab initio protein folding and structural analyses

All monomer models were downloaded from the AlphaFold Protein Structure Database. Protein multimers (5 models each) were generated for *A. bakeri* Orco, the 6 example 7TMICs in Figure 2, *Symbiodium natans* A0A812K102 (the most central node in the structural network, Figure 3A, asterisk), and 7 additional structures derived from Foldseek clustering of 7TMICs by 50% alignment coverage (thus representing nearly the entire 7TMIC fold space; Figure S2D; supplemental information). Predictions were performed in Google Colaboratory (https://research.google.com/colaboratory) using AlphaFold2+MMSeqs2 as implemented by Colabfold (https://github.com/sokrypton/ColabFold).[39–42] These models were not interpreted as accurate predictions of protein stoichiometry, but rather as hypothetical tetramers and as indirect predictions of protein-protein interactions. We also generated hypothetical dimers, trimers and pentamers for *A. bakeri* Orco (available in the supplemental information) and observed that the protein subunits assembled in a globally similar way – i.e., closest contact at the anchor domain(s). Transmembrane prediction was performed using DeepTMHMM, Phobius and PeSTo webservers, as described above. Protein-protein interactions were predicted using a local implementation of PeSTo (https://github.com/LBM-EPFL/PeSTo). All proteins were visualized in PyMol. Visualized structural alignments were generated using Coot.[54]

Network and conservation analyses

We used graph-based strategies for visualizing relatedness among proteins.[67] Structure-based networks were generated from the results of all-to-all DaliLite or Foldseek searches, where connections are derived from Z-scores >8 or TM scores >0.5, respectively. BLASTP sequence-based networking was performed using the CLANS webserver (https://toolkit.tuebingen.mpg.de/tools/clans) and a local implementation of CLANS,[51–53] using attraction values derived from E-values <0.01; clusters were identified using the built-in network clustering algorithm with the global averages option.

PSI-BLAST networking was performed via all-to-all PSI-BLAST searches using a local implementation of BLAST+.[35] First, BLAST databases were prepared from the sequence databases described above. Insect Ors were excluded, as they are an insect-specific radiation[68,69]; their removal thus reduced the likelihood of spurious connectivity between distantly related 7TMICs, as demonstrated by their relatively high connectivity in the BLASTP network (see Figure 1G). In other words, the removal of Ors hypothetically weakened network connectivity overall, but increased our confidence in homology between linked sequences. BLAST+ was then used to perform all-to-all PSI-BLAST searches, stopping at either convergence or 10 iterations. Position-specific scoring matrices (PSSMs) were generated with an E-value cutoff of 0.01 and the final network was assembled from hits where the PSSM query coverage was

>70%. For any query-to-subject relationship, only the first significant PSI-BLAST hit was kept, corresponding to the weakest significant connection (as connections tend to strengthen in subsequent PSI-BLAST iterations), thus providing the most conservative interpretation of the network. The opsin control/outgroup databases were from previous studies.[70,71]

All networks were visualized, annotated and quantitatively analyzed in CLANS, CytoScape[55] and Adobe Illustrator.

For conservation analyses, query-centered sequence alignments were first produced by Foldseek; in the figures, we visualized the alignment from the model with the highest closeness centrality (i.e., the centermost model; UniProt: A0A812K102) in the structural similarity network. Amino acid sequence identity scores were calculated in Jalview. Embedding-based conservation scores were calculated using the esm2_t33_650M_UR50D protein language model,[56] via methods and scripts described previously[37] (https://github.com/esbgkannan/kibby). The mean conservation scores were calculated by spline interpolating each individual data series (corresponding to each protein) to match the length of A0A812K102, then averaging those values; as such, family- and subfamily-specific conservation patterns are likely not represented in the average curve. The embedding-based conservation scores, PeSTo predictions, and query-centered multiple sequence alignments for all representative models are available in the supplemental information.

Phylogenetics

7TMIC GenBank accession numbers from our 50% clustered sequence database were matched to UniProt and 1947 AlphaFold-derived protein models were downloaded from the AlphaFold Protein Structure Database. All subsequent phylogenetic analyses were carried out on these 1947 representative proteins.

Muscle5 was used to generate the ensemble of multiple sequence alignments (MSAs).[59] Because the alignments were extremely long and gap rich (Table S1), MSAs were trimmed using trimal with the -gappyout option.[65] Each trimmed MSA was then used to generate phylogenetic trees using FastTree2,[57] using 3 different amino acid substitution models (JTT, WAG and LG), and with branch lengths rescaled to optimize the Gamma20 likelihood. The initial MSAs had extremely high dispersion and extreme lack of consensus (Figure S3B) indicating widespread alignment errors.[59] These errors led to non-trivial topological differences in the phylogenetic trees, resulting in extreme non-consensus (even for obviously monophyletic clades, such as the insect Ors). This suggested a high degree of phylogenetic instability, likely due to both alignment errors (from low pairwise sequence identity) and phylogenetic errors (e.g., long branch attraction).

To minimize alignment and phylogenetic errors, we repeated MSA and tree inference after identifying and removing rogue taxa (i.e., the most unstable leaves in the previous ensemble analysis) via RogueNaRok.[63] Although the resulting ensemble of MSAs still had high dispersion, the resulting phylogenetic trees were more consistent in the assignment of the various subfamilies as monophyletic clades (Figure S3C). These trees were used for subsequent analysis.

The structural phylogeny was generated using Foldtree.[44] Here, we emphasize the tree derived from all-against-all TM-scores, thereby sampling structural space based on pairwise global rigid structural comparisons, mirroring our network-based analysis, as described above. Structural trees derived from pairwise distances based on the Foldseek structural alphabet (Figure S3F) or pairwise lDDT scores (Figure S3G) produced radically different topologies; neither has obviously high congruence with the sequence-based phylogenetics nor with the presumed taxonomy of the species included in this analysis. All trees were analyzed using the ape,[48] phytools[62] and treespace[64] R packages. Tree topology space was explored by principal coordinate analysis of the Robison-Folds distances between the unrooted phylogenies. Trees were visualized and annotated using R, iTol (https://itol.embl.de/)[58] and Adobe Illustrator.

Protein nomenclature

Most previous naming conventions have not been evolutionarily informed. Terms such as Gustatory receptor-like (Grl and GRL) do not refer to monophyletic clades, but instead correspond to many taxon-specific 7TMIC branches. For animal Grls and unicellular eukaryotic GRLs, the terms were chosen because they resembled insect Grs in either amino acid sequence and/or tertiary structure[8,9,19]; by contrast, insect Grls were named based on the *absence* of sequence similarity to Grs despite the presence of structural similarity.[10] While these terms are useful in situational contexts, they are uninformative at long evolutionary scales. We propose that the 7TMIC superfamily be split into domain-specific families; for eukaryotes, these are Class-A and Class-B. We suggest that the more complex nomenclature of previous work (e.g., Or, Gr, GrlHz) should be reserved for taxon-specific contexts. Relatedly, the evolution of Arch7TMICs and Bac7TMICs is highly reticulated. Although we saw proximity between Heimdallarchaeota 7TMICs and Euk7TMICs in our structure-based phylogeny, there were no other clear recapitulations of Asgard/Eukaryota monophyly or the Archaea-Bacteria split. Therefore, "Arch7TMIC" and "Bac7TMIC" serve only as terms of convenience, and we strongly caution that they do not refer to monophyletic clades.

QUANTIFICATION AND STATISTICAL ANALYSIS

Pearson's correlation analysis and the Bayesian equivalent were performed in JASP (https://jasp-stats.org/); the correlation coefficient (r), p-value (α=0.05), and Bayes factor (BF_{10}; <1 evidence consistent with the null hypothesis and >1 evidence consistent with the alternative hypothesis) are reported in the corresponding figure panels (Figures S1C and S2B).

Current Biology, Volume *33*

Supplemental Information

Remote homolog detection places insect chemoreceptors in a cryptic protein superfamily spanning the tree of life

Nathaniel J. Himmel, David Moi, and Richard Benton

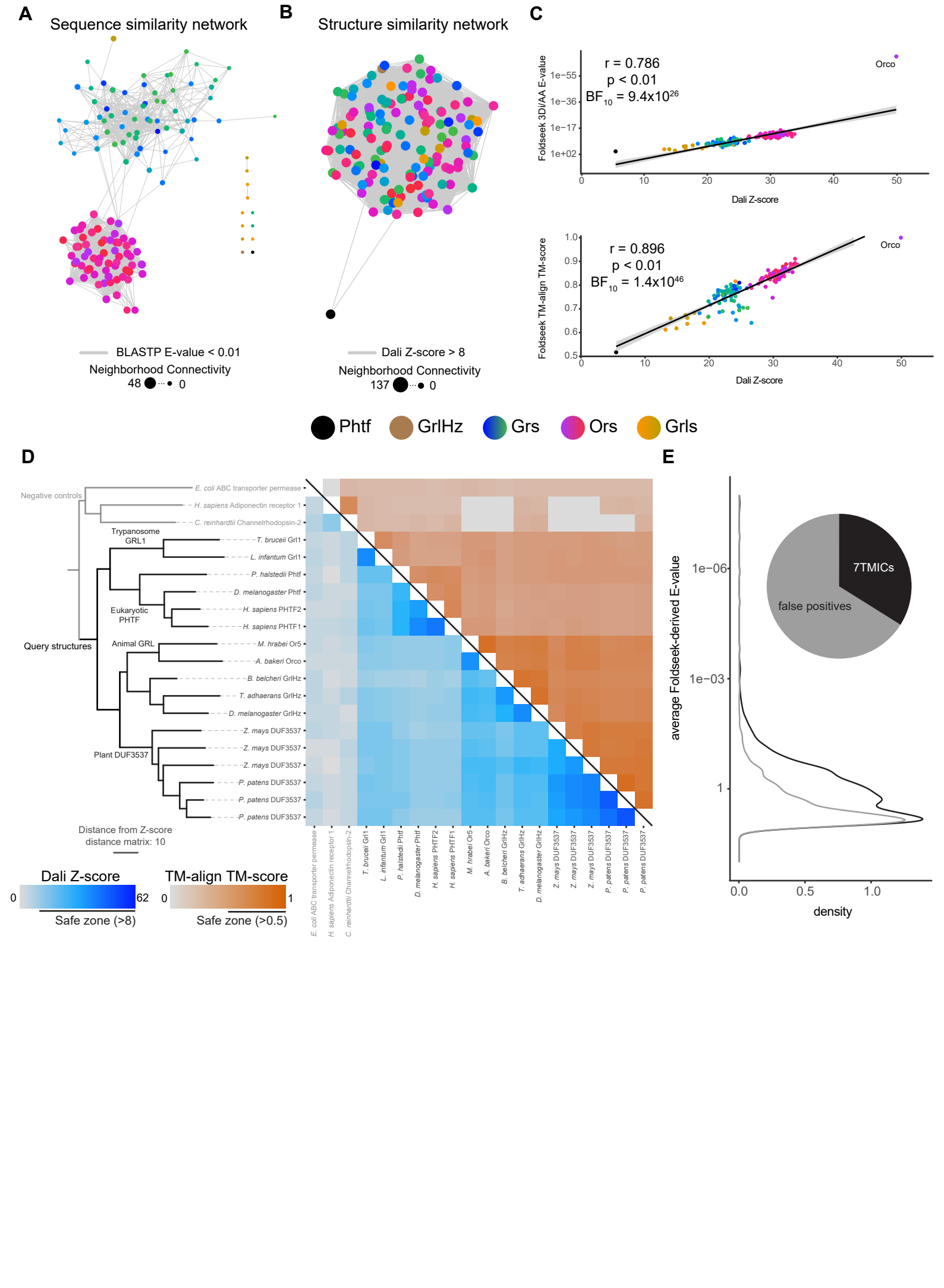

A
Sequence similarity network
BLASTP E-value < 0.01
Neighborhood Connectivity
48 0
B
Structure similarity network
Dali Z-score > 8
Neighborhood Connectivity
137 0
C
r = 0.786
p < 0.01
BF10 = 9.4x10^26
Foldseek 3Di/AA E-value
Dali Z-score
Orco
r = 0.896
p < 0.01
BF10 = 1.4x10^46
Foldseek TM-align TM-score
Dali Z-score
Orco
Phtf
GrlHz
Grs
Ors
Grls
D
Negative controls
E. coli ABC transporter permease
H. sapiens Adiponectin receptor 1
C. reinhardtii Channelrhodopsin-2
Trypanosome GRL1
T. bruceii Grl1
L. infantum Grl1
P. halstedii Phtf
D. melanogaster Phtf
Eukaryotic PHTF
H. sapiens PHTF2
H. sapiens PHTF1
Query structures
Animal GRL
M. hrabei Or5
A. bakeri Orco
B. belcheri GrlHz
T. adhaerans GrlHz
D. melanogaster GrlHz
Plant DUF3537
Z. mays DUF3537
Z. mays DUF3537
Z. mays DUF3537
P. patens DUF3537
P. patens DUF3537
P. patens DUF3537
Distance from Z-score distance matrix: 10
Dali Z-score
0
62
Safe zone (>8)
TM-align TM-score
0
1
Safe zone (>0.5)
E
average Foldseek-derived E-value
1e−06
1e−03
1
7TMICs
false positives
0.0
0.5
1.0
density

Figure S1. All-to-all pairwise protein similarity networks of *D. melanogaster* 7TMICs, Foldseek benchmarking, and summary of the Foldseek screen, related to Figure 1 and Figure 2.

(**A**) All-to-all BLASTP network of *D. melanogaster* 7TMICs; consistent with their low pairwise sequence similarity, this analysis fails to link every 7TMIC to all others. Rather, the major *D. melanogaster* classes (Ors and Grs) are separated into two identifiable community structures, with sparse connectivity among the Grs, and between the Grs and Ors. Other 7TMICs—including Grls, GrlHz, Phtf and two Grs—form singlets, indicating an inability to identify hypothetical homologs using BLASTP. The color key is shown below panels (A-C), and matches that of **Figure 1E**.

(**B**) All-to-all Dali network of *D. melanogaster* 7TMICs. In contrast to (A), structural comparisons result in a "hairball" network, where nearly all proteins are linked to all others, excepting Phtf, which is presumed to be the most distantly related.

(**C**) Plots of structural similarity scores between Orco and other *D. melanogaster* 7TMICs, comparing Dali to Foldseek-derived scores. Foldseek generates Orco-to-all E-values that tightly correlate with the rapidly generated 3Di+AA-derived E-values (top) and the slowly generated TM-align-derived TM-scores (bottom).

(**D**) Protein models used in the Foldseek screen, and negative controls used for subsequent Dali-based validation, with a clustering dendrogram based on all-to-all Dali comparisons between the queries and negative controls. The dendrogram is derived from the Dali Z-score distance matrix. The annotation on the heatmap corresponds to the "groups" described in the methods. The heatmap shows all-to-all Dali Z-scores and TM-scores. All 7TMIC-to-control comparisons are well below thresholds of confidence in fold similarity: the Dali Z-score maximum is 7.1 and averages 4.3; and the TM-align TM-score maximum is 0.31 and averages 0.2. By contrast, for 7TMIC-to-7TMIC comparisons: the Z-score minimum is 8.5 and averages 22.0, and the TM-score minimum is 0.4 and averages 0.6.

(**E**) Stacked density plot showing the frequency distribution of the hits of the Foldseek screen, by E-value, with the inset pie-chart showing the proportion of true positives to false positives. Many true positives had relatively poor E-values, with similar or worse scores than many false positives, demonstrating the need for structural validation in a Foldseek screen.

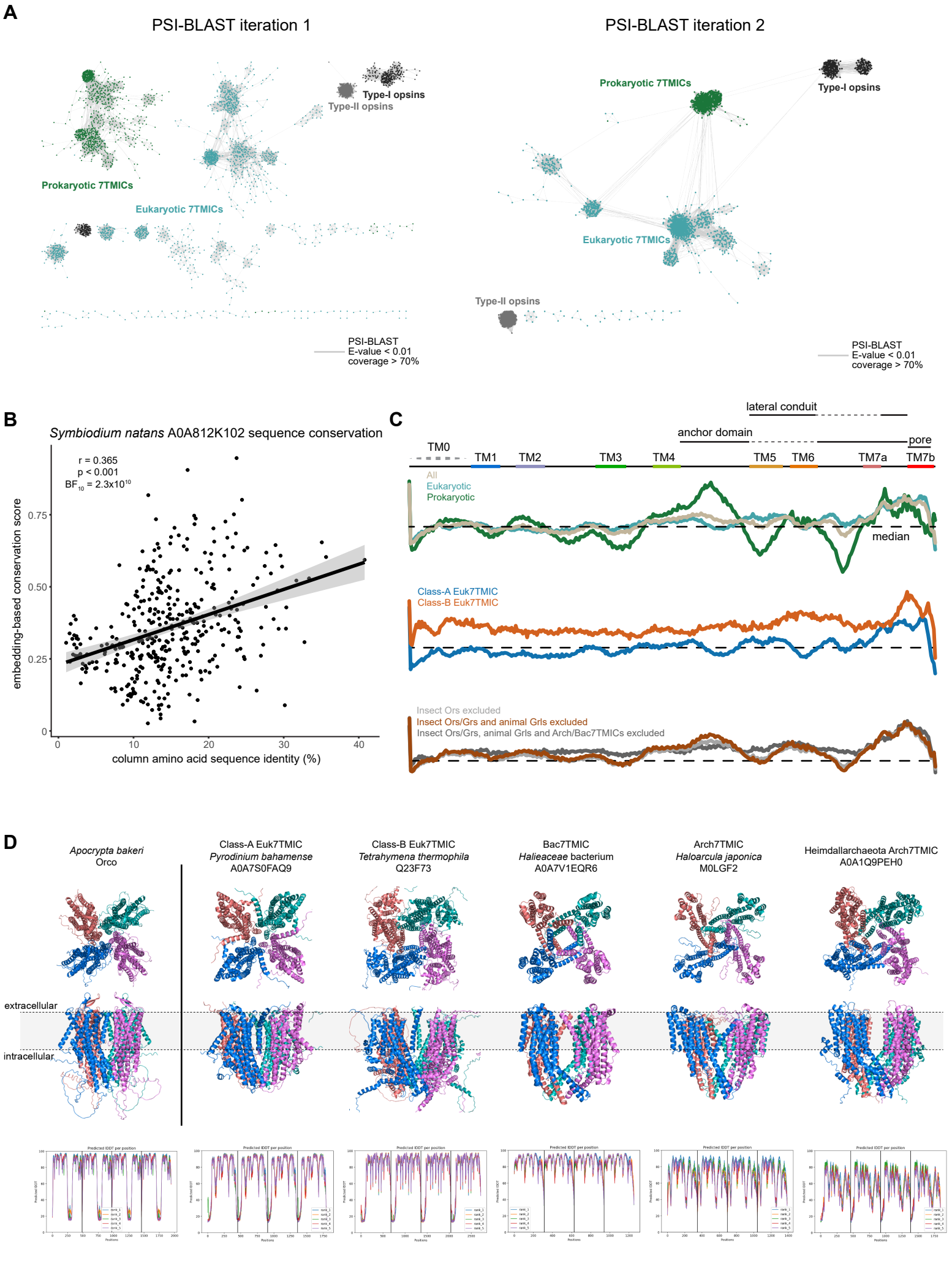

A
PSI-BLAST iteration 1
Type-I opsins
Type-II opsins
Prokaryotic 7TMICs
Eukaryotic 7TMICs
PSI-BLAST
E-value < 0.01
coverage > 70%
PSI-BLAST iteration 2
Prokaryotic 7TMICs
Type-I opsins
Eukaryotic 7TMICs
Type-II opsins
PSI-BLAST
E-value < 0.01
coverage > 70%
B
Symbiodium natans A0A812K102 sequence conservation
r = 0.365
p < 0.001
BF10 = 2.3x10^10
embedding-based conservation score
column amino acid sequence identity (%)
C
lateral conduit
anchor domain
pore
TM0
TM1
TM2
TM3
TM4
TM5
TM6
TM7a
TM7b
All
Eukaryotic
Prokaryotic
median
Class-A Euk7TMIC
Class-B Euk7TMIC
Insect Ors excluded
Insect Ors/Grs and animal Grls excluded
Insect Ors/Grs, animal Grls and Arch/Bac7TMICs excluded
D
Apocrypta bakeri Orco
Class-A Euk7TMIC Pyrodinium bahamense A0A7S0FAQ9
Class-B Euk7TMIC Tetrahymena thermophila Q23F73
Bac7TMIC Halieaceae bacterium A0A7V1EQR6
Arch7TMIC Haloarcula japonica M0LGF2
Heimdallarchaeota Arch7TMIC A0A1Q9PEH0
extracellular
intracellular

Figure S2. Initial iterations of the PSI-BLAST sequence similarity networks, 7TMIC sequence conservation analysis, and predicted quaternary structures of select, newly-identified 7TMICs, related to Figure 3.

(**A**) Sequence similarity networks were generated by all-to-all PSI-BLAST searches of a 50% clustered sequence database of 7TMICs, alongside databases of Type-I and Type-II opsins. Iterations 1 and 2 are visualized here. Subsequent iterations resemble the clustering pattern of iteration 3, as visualized in **Figure 3B**, albeit with strengthening community structures. Left: PSI-BLAST iteration 1. In this network, sequences formed several non-contiguous clusters, and failed to cluster together 7TMICs and Type-I opsins, which is expected given the substantial sequence dissimilarity of 7TMICs. Right: PSI-BLAST iteration 2. Surprisingly, PSI-BLAST networking produced bidirectional linking of the majority of 7TMICs, although presumed spurious linkages to outgroups began to form (which did not greatly multiply in subsequent iterations), and a small number of 7TMICs do not form links to the core 7TMIC cluster(s) (although all join a 7TMIC community structure by iteration 3 [**Figure 3B**]).

(**B**) Embedding-based conservation scores weakly but significantly correlate with column sequence identity from the A0A812K102-centered sequence alignment.

(**C**) Average embedding-based conservation scores for different subsets of 7TMICs, demonstrating that, while family-specific patterns exist, the conservation of anchor domain and pore regions is consistent. The TM and domain labels are derived from A0A812K102, as visualized in **Figure 3**.

(**D**) Predicted tetramers for select 7TMICs. Top: top (presumed extracellular) and side views of the tetrameric arrangement of 7TMICs predicted by AlphaFold-Multimer, showing the formation of a hypothetical pore along TM7b, similar to *A. bakeri* Orco (far-left). Bottom: local Distance Difference Test (lDDT) scores (used to assess model confidence), plotted for each of the 5 replicate models generated. Each color represents a different replicate; vertical black lines separate each of the modelled subunits. Generally, the transmembrane-spanning alpha helices are the most confidently predicted, leading to the similar patten of lDDT peaks and troughs across models.

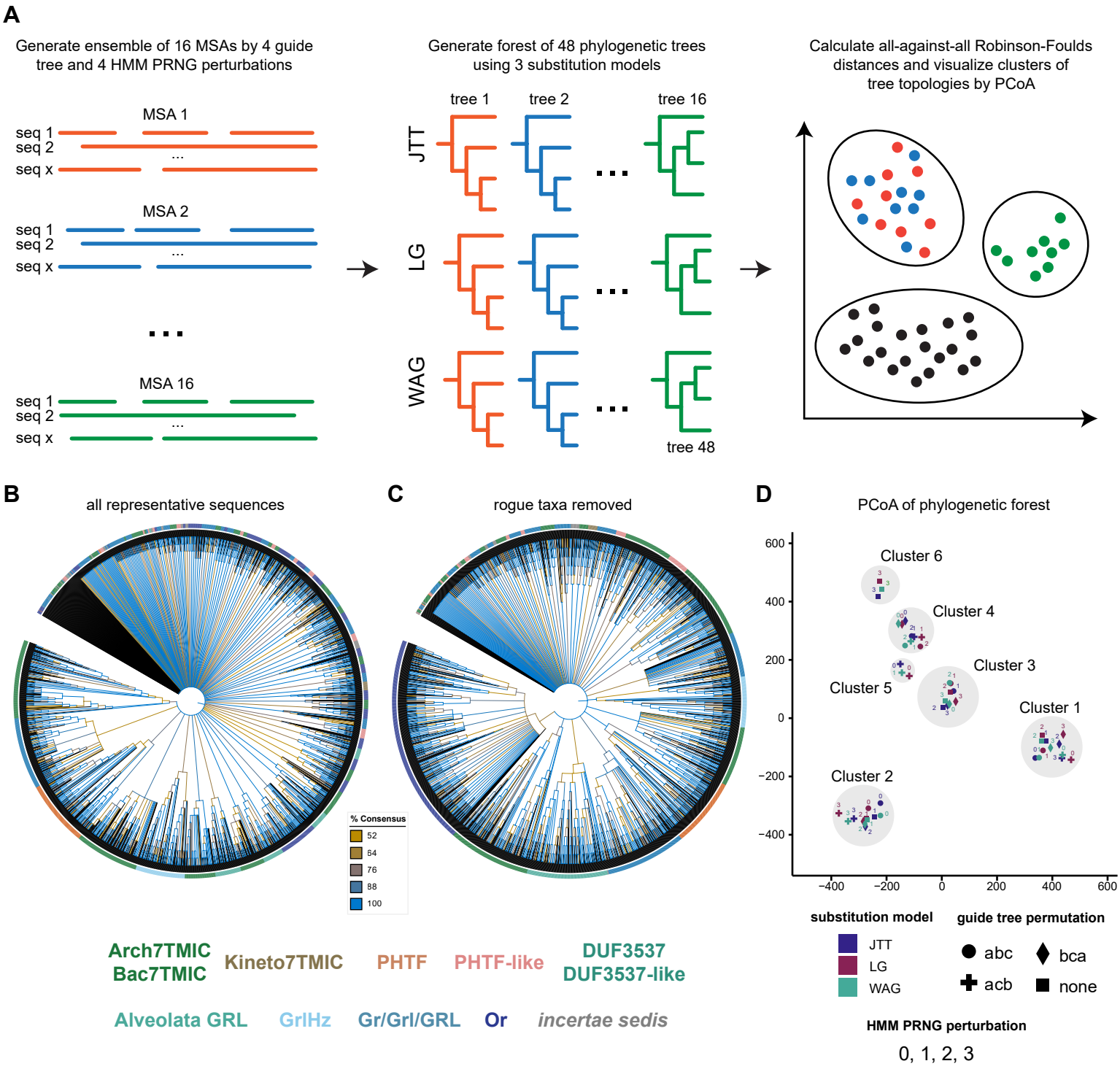

A
Generate ensemble of 16 MSAs by 4 guide tree and 4 HMM PRNG perturbations
MSA 1
seq 1
seq 2
...
seq x
MSA 2
seq 1
seq 2
...
seq x
MSA 16
seq 1
seq 2
...
seq x
Generate forest of 48 phylogenetic trees using 3 substitution models
tree 1
tree 2
tree 16
JTT
LG
WAG
tree 48
Calculate all-against-all Robinson-Foulds distances and visualize clusters of tree topologies by PCoA
B
all representative sequences
C
rogue taxa removed
% Consensus
52
64
76
88
100
D
PCoA of phylogenetic forest
Cluster 6
Cluster 4
Cluster 3
Cluster 5
Cluster 1
Cluster 2
substitution model
JTT
LG
WAG
guide tree permutation
abc
bca
acb
none
HMM PRNG perturbation
0, 1, 2, 3
Arch7TMIC
Bac7TMIC
Kineto7TMIC
PHTF
PHTF-like
DUF3537
DUF3537-like
Alveolata GRL
GrlHz
Gr/Grl/GRL
Or
incertae sedis
E
1
2
3
4
5
6
F
branch length: 0.1
G
branch length: 0.1

Figure S3. Phylogenetic and tree space analysis, related to Figure 4.
(**A**) Pipeline for sequence-based phylogenetic analysis. First, an ensemble of 16 multiple sequence alignments (MSAs) are made by perturbating the guide tree and the Hidden Markov model's pseudorandom number generator (HMM PRNG). Second, phylogenetic trees are generated for each of the MSAs, using 3 different amino acid substitution models, resulting in 48 trees. Finally, differences in the topology of the 48 trees are calculated by pairwise Robinson-Foulds distances; the resulting distance matrix is subsequently visualized in two dimensions by principal coordinate analysis (PCoA).
(**B**) Majority consensus tree for the 48 phylogenetic trees based on alignments of the representative 7TMIC sequences. 7TMIC clades/colors were assigned manually based on visual inspection of a CLANS-based clustering analysis; the color key for 7TMIC subfamilies is shown below panels (B-C). Branch colors indicate the percent consensus. There is essentially no clear consensus among these 48 initial trees; obviously monophyletic clades—such as insect Ors—are not reliably predicted, suggesting substantial alignment/phylogenetic errors (as expected for this highly divergent superfamily).
(**C**) Majority consensus tree for the 48 phylogenetic trees based on alignments of a 7TMIC dataset where rogue taxa (i.e. the most phylogenetically unstable leaves) have been removed. While there is still no greatly informative majority consensus topology, this analysis better recapitulates more obvious monophyletic clades, with higher branch consensus, indicating that errors have been minimized (but not eliminated, which we did not expect to occur at these levels of sequence dissimilarity).
(**D**) PCoA of Robinson-Foulds tree space for trees from (C). Trees form 6 topology clusters.
(**E**) Majority consensus trees for each of the 6 clusters, with colors matching (B) and (C). Five of these clusters agree that Kineto7TMICs branch proximally to prokaryotic 7TMICs, consistent with the hypothesis that kinetoplastids (and allies: Discoba) split early in eukaryotic evolution. Clusters 1 and 4 do not have majority consensus on deep 7TMIC branching. The remaining clusters suggest there are at least two Euk7TMIC families, termed Class-A and Class-B Euk7TMICs, but do not agree on the monophyly of Class-B Euk7TMICs. Clusters 4-6 suggest Class-B monophyly, while clusters 1-3 suggest that many proteins are basally branching (and thus, paraphyly). Given that structure-based phylogenetics suggest a monophyletic Class-B, this discordance may be the result of lingering long branch attraction or other errors resulting from the inclusion of rapidly evolved, horizontally-transferred and/or structurally-convergent proteins.
(**F**) Structural phylogeny derived from pairwise distances used the Foldseek 3Di structural alphabet, with colors matching the panels above. This tree is presented as rooted, but as in Figure 4, the true root is likely within the prokaryotic 7TMICs, at the location of the Last Universal Common Ancestor.
(**G**) Structural phylogeny derived from pairwise lDDT scores, with colors matching the panels above. As in (F), the true root is likely at location of the Last Universal Common Ancestor.

	MSA Perturbations		All Sequences		No Rogue Taxa	
MSA Replicate	**Guide Tree**	**PRNG**	**Columns**	**Column Confidence**	**Columns**	**Column Confidence**
1	abc	0	38154	0.315	36865	0.311
2	abc	1	38357	0.297	34833	0.295
3	abc	2	36970	0.314	35376	0.325
4	abc	3	37064	0.314	31247	0.304
5	acb	0	38375	0.31	35627	0.304
6	acb	1	40342	0.298	34754	0.302
7	acb	2	37080	0.317	34994	0.321
8	acb	3	35735	0.313	31345	0.305
9	bca	0	38391	0.315	35914	0.314
10	bca	1	39831	0.3	35813	0.294
11	bca	2	37647	0.316	34847	0.322
12	bca	3	36515	0.305	31406	0.303
13	none	0	38700	0.3	35842	0.309
14	none	1	40114	0.306	34808	0.293
15	none	2	37443	0.308	34932	0.309
16	none	3	37831	0.309	32689	0.298

Table S1. Muscle5 multiple sequence alignment analysis, related to Figure 4.
Column confidence is a measure of the reproducibility of each column, where 0 indicates the column is never found and 1 indicates it is found across all alignments. Dispersion is measured as the median dispersion of aligned letter pairs over the ensemble (D_LP), and the median dispersion of columns over the ensemble (D_Cols) (Robert Edgar, personal communication, 10 May 2023), where 0 is all the same and 1 is all different. Dispersion was extremely high. For the initial set of alignments: D_LP=0.5836 D_Cols=1.0000. After removal of rogue taxa: D_LP=0.5855 D_Cols=1.0000.