Structural screens identify candidate human homologs of insect chemoreceptors and cryptic *Drosophila* gustatory receptor-like proteins

Richard Benton[1,*] and Nathaniel J. Himmel[1,*]

[1]Center for Integrative Genomics
Faculty of Biology and Medicine
University of Lausanne
CH-1015
Lausanne
Switzerland

[*]Corresponding authors:

richard.benton@unil.ch
nathanieljohn.himmel@unil.ch

[1]These authors contributed equally to this work

Abstract

Insect Odorant receptors and Gustatory receptors define a superfamily of seven-transmembrane domain ligand-gated ion channels (termed here 7TMICs), with homologs identified across Animalia except Chordata. Previously, we used sequence-based screening methods to reveal conservation of this family in unicellular eukaryotes and plants (DUF3537 proteins) (Benton *et al*., 2020). Here we combine three-dimensional structure-based screening, *ab initio* protein folding predictions, phylogenetics and expression analyses to characterize additional candidate homologs with tertiary but little or no primary structural similarity to known 7TMICs, including proteins in disease-causing Trypanosoma. Unexpectedly, we identify structural similarity between 7TMICs and PHTF proteins, a deeply-conserved family of unknown function, whose human orthologs display enriched expression in testis, cerebellum and muscle. We also discover divergent groups of 7TMICs in insects, which we term the Gustatory receptor-like (Grl) proteins. Several *Drosophila melanogaster Grl*s display selective expression in subsets of taste neurons, suggesting that they are previously-unrecognized insect chemoreceptors. Although we cannot exclude the possibility of remarkable structural convergence, our findings support the origin of 7TMICs in a eukaryotic common ancestor, counter previous assumptions of complete loss of 7TMICs in Chordata, and highlight the extreme evolvability of this protein fold, which likely underlies its functional diversification in different cellular contexts.

Introduction

The insect chemosensory receptor repertoires of Odorant receptors (Ors) and Gustatory receptors (Grs) define a highly divergent family of ligand-gated ion channels, which underlie these animals' ability to respond to chemical cues in the external world (Benton, 2015; Joseph and Carlson, 2015; Robertson, 2019). Despite its vast size and functional importance, this family has long been an evolutionary enigma, displaying no resemblance to other classes of ion channels. Indeed, for many years, insect Ors and Grs were thought to be an invertebrate-specific protein class (Benton, 2006; Robertson et al., 2003). This view changed in the past decade, with the sequencing of a large number of genomes enabling the identification of homologs across animals (generally termed Gr-like (GRL) proteins)), including non-Bilateria (e.g., the sea anemone *Nematostella vectensis*), Hemichordata (e.g., the sea acorn *Saccoglossus kowalevskii*), various unicellular eukaryotes (e.g., the chytrid fungus *Spizellomyces punctatus* and the alga *Vitrella brassicaformis*) and Plantae (known as Domain of Unknown Function (DUF) 3537 proteins) (Benton, 2015; Benton et al., 2020; Robertson, 2015; Saina et al., 2015). For simplicity in nomenclature, we term here this broader superfamily (i.e., Ors, Grs, GRLs and DUF3537 proteins) as "seven transmembrane domain ion channels" (7TMICs), to distinguish them from unrelated 7TM G protein-coupled receptors (while acknowledging that in most cases we do not know yet whether they are ion channels). Despite extensive searching, 7TMIC homologs have not been identified in Chordata, leading to proposals that these proteins were lost at or near the base of the chordate lineage (Benton, 2015; Robertson, 2015; Saina et al., 2015).

A substantial challenge in identifying 7TMIC homologs is their extreme sequence divergence (as little as 8% amino acid identity). The inclusion of

proteins in this family relies primarily on the presence of topological features, notably seven TM domains and an intracellular N-terminus (Benton et al., 2020; Benton et al., 2006). Although insect Grs were originally recognized as possessing a short, conserved motif in transmembrane domain 7 (TM7) (described below) (Robertson, 2019; Scott et al., 2001), this motif is only partially or not at all conserved outside insects (Benton et al., 2020). For many protein families, the tertiary (three-dimensional) structure is generally more conserved than primary structure (Illergard et al., 2009; Murzin et al., 1995), and this property can offer an orthogonal strategy to validate homologous proteins. For the 7TMIC superfamily, the recent cryo-electronic microscope (cryo-EM) structures of homotetrameric complexes of insect Ors (Butterwick et al., 2018; Del Marmol et al., 2021) provide important experimental insight into the tertiary structure of these proteins (as well as mechanistic insights into how these ion channels function). In our previous study (Benton et al., 2020), we used *ab initio* structural predictions of candidate 7TMIC sequences to reinforce our proposals of homology despite extremely low amino acid identity.

The recent breakthroughs in accuracy (to atomic level) and speed (seconds-to-minutes per sequence) of protein structure predictions, notably by AlphaFold2 (Jumper et al., 2021; Varadi et al., 2022), have now enabled millions of protein models to be generated. Here we have exploited the unprecedented resource of the AlphaFold Protein Structure Database (Jumper et al., 2021; Varadi et al., 2022) and the DALI protein structure comparison algorithm (Holm, 2022), to screen for additional 7TMIC homologs by virtue of their tertiary structural similarity to experimentally-determined insect Or structures.

Results and Discussion

Tertiary structure-based screening for candidate 7TMIC homologs

Cryo-electronic microscopic (cryo-EM) structures of two insect Ors have been obtained: the fig wasp (*Apocrypta bakeri*) Or co-receptor (Orco) (Butterwick et al., 2018) (Figure 1A-B), which is a highly-conserved member of the repertoire across most insect species (Benton et al., 2006; Larsson et al., 2004) and MhOr5 from the jumping bristletail (*Machilis hrabei*), a broadly-tuned volatile sensor (Del Marmol et al., 2021). Despite sharing only 18% amino acid identity, these proteins adopt a highly similar fold (Del Marmol et al., 2021). As Orco shows higher sequence similarity to Grs – the ancestral family of insect chemosensory receptors from which Ors derived (Brand et al., 2018; Dunipace et al., 2001; Robertson et al., 2003) –we used *A. bakeri* Orco as the query structure in our analysis.

In our previous work (Benton et al., 2020), we generated *ab initio* protein models of Orco and candidate homologs in various unicellular eukaryotes using trRosetta (Yang et al., 2020) and RaptorX (Kallberg et al., 2012). We therefore first examined the AlphaFold2 structural model of *A. bakeri* Orco (Figure 1C) (Jumper et al., 2021; Varadi et al., 2022). This model displays striking qualitative similarity to the experimental structure (PDB 6C70 chain A) (Figure 1C). We assessed structural similarity quantitatively using two algorithms: first, using pairwise structural alignment in DALI (Holm, 2022), we extracted the resultant Z-score (the sum of equivalent residue-wise C_α-C_α distances between two proteins); second, we determined the template modeling (TM)-score from TM-align (Zhang

and Skolnick, 2004, 2005) (a measure of the global similarity of full-length proteins) (Table 1). These measures confirmed the visual impression that the modeled and experimental structures are almost identical (e.g., TM-score = 0.96, where 1 would be a perfect match). We extended our assessment of available (or newly-generated) AlphaFold2 models to other well-established members of the 7TMIC family from animals as well as more much more divergent unicellular 7TMIC homologs previously identified (Benton et al., 2020) (Supplementary file 1). Using the same quantitative assessments, these all displayed substantial tertiary structural similarity to *A. bakeri* Orco (Table 1), reinforcing our previous conclusions that these proteins form part of the same superfamily. Moreover, the observation that multiple distinct algorithms (AlphaFold2, trRosetta and RaptorX) predict the same global fold of these proteins strengthens confidence in the validity of *ab initio* structural models.

We proceeded to screen the AlphaFold Protein Structure Database for other proteins that are structurally similar to *A. bakeri* Orco using the hierarchical search function in DALI (Holm, 2022). This algorithm currently permits pairwise alignment of Orco to the complete predicted structural proteomes of 47 species – encompassing several vertebrates, invertebrates, plants, unicellular eukaryotes and prokaryotes – returning hits ordered by Z-score (Supplementary file 2). We focused on those hits with a Z-score of >10 (Figure 1D). This threshold successfully captured known 7TMICs, while removing a large number of proteins (generally with a much lower Z-score) that did not fulfil other criteria for structural similarity (as described below. Of the expected hits, within the *D. melanogaster* structural proteome we recovered all models of the members of the Or and Gr repertoires. From *C. elegans*, we found all members of the Gustatory receptor (GUR) family (Robertson et al., 2003) – including the photoreceptor LITE-1 (formerly GUR-2) (Edwards et al., 2008; Gong et al., 2016; Liu et al., 2010) – and the Serpentine receptor R (SRR) family (which are of unknown function, but with diverse neuronal and non-neuronal expression (Vidal et al., 2018)) (Figure 1D and Supplementary file 2). From the four plant species screened, all members of the DUF3537 family were successfully identified (Figure 1D and Supplementary file 2). Inspection of several models below our Z-score threshold indicated that the proteins (typically multipass membrane proteins) have likely spurious resemblance to subregions of Orco rather than displaying similarity in their overall fold.

As will be illustrated below for individual novel candidate 7TMIC homologs, other hits were subsequently analyzed for their fulfilment of several criteria: (i) the presence of seven predicted TM domains, (ii) a predicted intracellular location of the N-terminus and (iii) longer intracellular than extracellular loops (like insect Ors (Otaki and Yamamoto, 2003), while also recognizing that intracellular loops can vary enormously in length in homologs (Benton et al., 2020)). For hits that fulfilled these criteria, “reverse” searching of the *D. melanogaster* structural proteome with DALI was performed to verify that Ors and Grs were structurally the most similar proteins in this species (Supplementary file 3). We next qualitatively assessed the predicted tertiary structural similarity to *A. bakeri* Orco (Figure 1A-C) (Butterwick et al., 2018), verifying: (i) the characteristic packing of the TMs, (ii) the projection of the long TM4, TM5 and TM6 below the main bundle of helices (forming the “anchor” domain where most intersubunit contacts occur in complexes (Butterwick et al., 2018; Del Marmol et al., 2021)) and (iii) the exceptional splitting of TM7 into two subregions (TM7a, part of the anchor

domain, and TM7b, which lines the ion conduction pathway (Butterwick et al., 2018; Del Marmol et al., 2021)). Structures were also quantitatively compared to *A. bakeri* Orco, as described above (Table 1). As negative controls, we also performed comparisons with a variety of other multipass membrane proteins belonging to other superfamilies, including several with seven TMs (e.g., Rhodopsin, Frizzled and the Adiponectin receptor) (Table 1). The new candidate homologs all displayed quantitative measures of similarity that were within the range of previously identified 7TMIC homologs, and clearly superior to the scores of negative control proteins (Table 1). We now present these candidate homologs from different species and the potential evolutionary and biological implications for the 7TMIC family, bearing in mind the caveat that some of these may represent cases of structural convergence (discussed below).

Extending our previous discovery of 7TMICs in various single-celled eukaryotes (informally grouped here under the termed Protozoa) (Benton et al., 2020), we identified single proteins in two species belonging to the Trypanosomatida order: *Leishmania infantum* and *Trypanosoma brucei*, the causal agents in humans of trypanosomiasis (sleeping sickness) and visceral leishmaniasis (black fever), respectively (Figure 1D-F and Table 1). Beyond the 7TMIC-like protein fold (Figure 1E-F and Table 1), these proteins are characterized in their N-terminal regions by a Membrane Occupation and Recognition Nexus (MORN)-repeat domain, which is implicated in protein-protein interaction and possibly lipid binding (Sajko et al., 2020). BLAST searches identified homologous proteins only within trypanosomes (Figure 1G), consistent with our failure to recover these sequences in earlier primary structure-based screens for 7TMICs. We did not detect any structurally-related proteins to Orco in Prokaryota or Fungi (previously, fungal GRLs were only identified in chytrids (Benton et al., 2020), which are not currently surveyed via DALI). Together, these results reinforce our previous conclusion (Benton et al., 2020) that 7TMICs evolved in or prior to the last eukaryotic common ancestor, and provide a first example of fusion of this transmembrane protein domain with a distinct, cytoplasmic protein domain.

PHTF proteins are candidate vertebrate 7TMICs

Given previous lack of success in identifying homologs of 7TMICs within any chordate genome, we were intrigued that our screen recovered two hits from *Homo sapiens* (and orthologous proteins of the three other vertebrate species screened) (Figure 1D and Supplementary files 2-3). The human proteins, PHTF1 and PHTF2, are very similar to each other (54.1% amino acid identity) and have the characteristic topology of 7TMICs (Figure 2A). The next most similar vertebrate proteins to Orco had substantially lower DALI Z-scores than PHTFs and represented a variety of likely spurious matches (Supplementary file 2). The single *D. melanogaster* ortholog (Phtf) (Manuel et al., 2000) displays a similar topology to the vertebrate proteins (Figure 2A), and is the next most similar protein model to *A. bakeri* Orco after the *D. melanogaster* Grs, Ors and Grls (see next section) (Supplementary file 2). PHTF is an acronym of “Putative Homeodomain Transcription Factor”, a name originally proposed because of presumably artefactual sequence similarity of a short region around TM4 to homeodomain DNA-binding sequences (Raich et al., 1999); subsequent

histological and biochemical studies (discussed below) established that PHTF1 is an integral membrane protein (Oyhenart et al., 2003).

To visually compare AlphaFold2 models of PHTF orthologs with *A. bakeri* Orco, we masked the long (>300 amino acid) first intracellular loop (Figure 2A), whose structure is mostly unpredicted but contains a few α-helical regions, as well as the ~100-residue N-terminus (Figure 2B). This visualization revealed the clear similarity in the organization of the seven TM helical core of the protein, including the split TM7 (Figure 2B-C), which was verified by quantitative structural comparisons (Table 1).

In contrast to other, taxon-restricted members of the 7TMIC superfamily, highly conserved PHTF homologs were found across Eukaryota, including in Bilateria, Cnidaria and several unicellular species (Figure 2D and Supplementary file 5). Phylogenetic analyses of a representative PHTF protein sequence dataset reveal that there is a single eukaryotic PHTF clade (Figure 2E and Figure 2—figure supplements 1-3). Bayesian and maximum likelihood phylogenetics largely agree on the topology of this tree and suggest that the PHTF1-PHTF2 duplication occurred specifically in the jawed vertebrate lineage (Gnathostomata) (Figure 2E).

Previous tissue-specific RNA expression analysis by Northern blotting of *H. sapiens PHTF1* and *PHTF2* revealed enrichment in testis and muscle, respectively (Manuel et al., 2000). We confirmed and extended these conclusions by analyzing publicly-available bulk RNA-sequencing (RNA-seq) datasets: *PHFT1* is most abundantly detected in cerebellum and testis, and *PHTF2* in skeletal muscle and arteries (Figure 2F and Figure 2—figure supplement 4). *D. melanogaster Phtf* displays highly-enriched expression in the testis, and much lower expression in neural tissues in the FlyAtlas 2.0 bulk RNA-seq datasets (Figure 2F and Figure 2—figure supplement 5) (Krause et al., 2022), potentially indicating a closer functional relationship to PHTF1 than PHTF2. Higher resolution expression analysis of *Phtf* in male reproductive tissue, using the Fly Cell Atlas (Li et al., 2022), revealed the most prominent expression in developing spermatocytes and spermatids (Figure 2G). The transcript expression of *D. melanogaster Phtf* is concordant with detection of rat (*Rattus norvegicus*) PHTF1 protein from primary spermatocytes to the end of spermatogenesis, predominantly localized to the endoplasmic reticulum (Oyhenart et al., 2005b; Oyhenart et al., 2003). The N-terminal region of mouse (*Mus musculus*) PHTF1 associates with the testis-enriched FEM1B E3 ubiquitin ligase and is suggested to recruit it to the endoplasmic reticulum (Oyhenart et al., 2005a). Overexpression and/or knock-down studies of PHTF1 and PHTF2 in cell lines hint at roles in regulating cell proliferation and survival, and possible links to various cancers (Chi et al., 2020; Huang et al., 2015). However, the biological function of any PHTF homologs in any organism is unclear. Nevertheless, PHTFs represent the first candidate homologs of the insect chemosensory receptors from chordates, indicating they might not have been completely lost from this lineage as previously thought (Benton, 2015; Robertson, 2015); we suggest they also act as ion channels

Novel sets of candidate insect chemoreceptors

Within the hits of our screen of *D. melanogaster* protein structures, we noticed ten proteins that do not belong to the canonical Gr or Or families (Supplementary file

2). These proteins have a similar length and transmembrane topology as Grs and Ors (Figure 3A). Visual inspection and quantitative analyses confirmed that their predicted fold is very similar to that of *A. bakeri* Orco (Figure 3A and Table 1). As they almost completely lack other defining sequence features of these families (see below) – we named these Gustatory receptor-like (Grl) proteins, using the same gene cytogenetic-based nomenclature conventions of other chemosensory gene families (e.g., (*Drosophila* Odorant Receptor Nomenclature Committee, 2000)), with one exception (GrlHz, see below).

For seven *D. melanogaster* Grls, BLAST searches identified homologs only in drosophilids; for two others (Grl40a and Grl65a) we recovered drosophilid and other fly homologs (Supplementary file 6). By contrast, the Grl originally designated CG3831 has homologs across a wide range of Holozoa (i.e., animals and their closest single-celled, non-fungal relatives), including chordates (e.g., the lancelet *Branchiostoma floridae*) and single-cell eukaryotes (e.g., *Capsaspora owczarzaki*) (Figure 3—figure supplements 1-4), leading us to name it GrlHolozoa (GrlHz). A subset of GrlHz homologs bear a long N-terminal domain containing WD40 repeats, which form a structurally-predicted beta-propeller domain, which is typically involved in protein-protein interactions (Figure 3—figure supplement 1D) (Kim and Kim, 2020).

Given that nine of these Grls are restricted to flies, a reasonable hypothesis is that they evolved from fly Grs. To infer their evolutionary origins, we therefore examined sequence similarity of Grls with a representative set of Grs, as well as Ors and other animal (i.e., non-insect) Grls. We found that fly Grls share little or no obvious sequence similarity with any of these other 7TMICs, precluding confident standard phylogenetic analysis and leading us to use an all-to-all graph-based methodology, which does not require a multiple sequence alignment. This approach infers relationships between sequences based on pairwise sequence similarity. This analysis first generates an all-to-all sequence similarity network via MMseqs2, in which sequence families can be identified as clusters in a 2D projection (Figure 3B), and then a tree by recursive spectral clustering (Figure 3—figure supplement 5) (see Methods) (Matsui and Iwasaki, 2020; Steinegger and Soding, 2017). In the network, we observed that several of the Grls were intermingled in clusters (e.g., Grl62b/Gr62c and Grl36a/Grl43a), suggesting relatively recent common ancestry (Figure 3B and Supplementary file 6). These two clusters were recapitulated as clades in the Graph Splitting phylogeny (Figure 3—figure supplement 5). For Grl62a/b/c, the possibility of recent ancestry is consistent with the tandem genomic organization of the corresponding genes, which implies their evolution by gene duplication through non-allelic homologous recombination, similar to other families of chemosensory genes (Nei et al., 2008). None of the Grl clusters grouped with those of Ors, Grs or other animal Grls, rather connecting broadly, but weakly, with all other clusters (Figure 3B). Consistent with this clustering pattern, all Grls were placed near the presumed root of the Graph Splitting tree (Figure 3—figure supplement 5). This basal placement of Grls was inconsistent with their conservation only in flies, and is likely a phylogenetic artefact (see legend to Figure 3—figure supplement 5).

Although analysis of amino acid sequences did not provide evidence of ancestry between Grs and Grls, we noted that *Grl36a* was immediately adjacent (separate by 306 bp) to the *Gr36a/b/c* cluster in the *D. melanogaster* genome. This proximity suggested that *Grl36a* might have arisen by gene duplication of a *Gr36*-like ancestor. Indeed, *Grl36a* homologs across drosophilid species were

always found in tandem with *Gr36*-related genes in various arrangements (Figure 3C, Figure 3—figure supplement 6). To further investigate the hypothetical ancestry of *Grl36a* and *Gr36*, we first examined their gene structure. We incorporated into this analysis *Gr59c* and *Gr59d* homologs, which are closely-related to *Gr36a/b/c* even though they are distantly located in the genome (Robertson et al., 2003), as well as *Grl43a*, the most closely-related paralog to *Grl36a* (Figure 3B). The *Gr* family is characterized by the general, but not universal, conservation of three C-terminal, phase 0 introns (Robertson et al., 2003). *Gr36*, *Gr59c/d* and homologous non-drosophilid genes possess only one of these introns, which corresponds to the second ancestral *Gr* intron located just before the exon encoding TM7. Both *D. melanogaster Grl36a* and *Grl43a* also have a phase 0 intron immediately before the TM7-encoding exon, which aligns with the *Gr* intron position on a multiple protein sequence alignment (Figure 3D), suggesting that these *Grl* and *Gr* introns are homologous. We next examined the TM7 motifs in these Grs and Grls. The canonical TM7 motif of Grs is TYhhhhhQF, where h is a hydrophobic residue (Figure 3D) (Robertson, 2019; Scott et al., 2001). However, Gr36 and Gr59c/d share a variant motif, T(H/N)(S/A)hhhhQ(Y/F/W), and we observed a very similar motif in Grl36a and Grl43a (Figure 3D).

The genomic proximity of *Gr36* and *Grl36a*, and similarity in introns and TM7 motifs of these genes (as well as *Gr59c/d* and *Grl43a*) provides evidence that these genes have a relatively recent common ancestry within drosophilids. Phylogenetic analyses of this proposed clade support that a Grl36a/Grl43a clade is the sister clade to Gr36, and that this split occurred after the emergence of the Gr59c/d clade (Figure 3E, Figure 3—figure supplements 6-9). None of the other *Grl* genes are located adjacent to *Gr* genes, nor do the proteins possess a recognizable TM7 motif. Some other Grls might possess conserved introns of *Gr*s (e.g., *Grl40a* with the first ancestral intron, and *Grl46b* and *Grl65a* with the second ancestral intron (data not shown)), but we cannot conclude with confidence that these are homologous. Thus, the ancestry of most Grls remains unresolved. Nevertheless, the highly restricted taxonomic representation of nine of these Grls and their structural similarity to Grs support a model in which Grls have evolved and diverged rapidly from ancestral Grs.

To gain insight into the potential role(s) of Grls, we first examined their expression in tissue-specific bulk RNA-seq datasets from the FlyAtlas 2.0 (Krause et al., 2022). Most *Grl*s were expressed at very low (<1 fragment per kilobase of exon per million mapped fragments (FKPM)) or undetectable levels in essentially all tissues in these datasets, although *Grl36*b was detected in neuronal tissues (eye, brain, thoracicoabdominal ganglion) (Figure 2—figure supplement 5). The one exception was *GrlHz*, which was expressed (>8 FKPM) in various tissues (e.g., heart, ovary, testis, and larval fat body and garland cells (nephrocytes)). The unique expression and conservation properties of *GrlHz* suggest it might have a different function from other *Grl*s.

The lack of detection of transcripts for most *Grl*s in the FlyAtlas 2.0 suggested that these genes might have highly restricted cellular expression patterns. Given the structural similarity of Grls to Grs, we examined their expression in an RNA-seq dataset of the major taste organ (labellum; a tissue not specifically represented in the FlyAtlas 2.0) (Dweck et al., 2021). *D. melanogaste*r *Gr* genes display a wide range of expression levels in the labellar transcriptome, in part reflecting the breadth of expression in different classes of taste neurons.

For example, *Gr66a* and *Gr64f* – broadly-expressed markers for "bitter/aversive" and "sweet/appetitive" neuronal populations, respectively (Freeman and Dahanukar, 2015) – are detected at comparatively high levels (>5 FKPM) (Figure 3F). By contrast, many receptors expressed in subsets of these major neuron types (e.g., *Gr22e* for bitter and *Gr61a* for sweet (Freeman and Dahanukar, 2015)) are expressed at much lower levels (~1 FKPM). Similar to this latter type of *Gr*, transcripts for four *Grl*s were detected at >0.5 FKPM: *GrlHz*, *Grl62c*, *Grl62a* and *Grl36a* (Figure 3F). Importantly, within the Fly Cell Atlas dataset of the proboscis and maxillary palp (Li et al., 2022), three of these were specifically expressed in the cluster of cells corresponding to *Gr66a*-expressing bitter/aversive neurons (Figure 3G). The fourth, *GrlHz*, was very sparsely expressed in non-neuronal cell types, including hemocytes (data not shown). None of the other six *Grl*s were detectable in this dataset, consistent with their lower expression in the labellar bulk RNA-seq transcriptome (Figure 3F). Moreover, no *Grl* was detectably expressed in other chemosensory tissue transcriptomes (leg, wing or antenna) (data not shown). These observations raise the possibility that at least three Grls (Grl62c, Grl62a and Grl36a) are chemosensory receptors for aversive stimuli.

A hypothesis for the evolution of the 7TMIC superfamily

Two hypotheses could explain the similarities between well-established 7TMICs and the candidate homologs described in this work: homology (i.e., shared ancestry), and thus the existence of a unified 7TMIC superfamily, or convergent evolution of the 7TMIC structure in one or more lineages. We discuss the latter possibility in the following section. Here we consider a detailed hypothesis of a 7TMIC superfamily of single evolutionary origin. Because confident multiprotein alignment of all members was impossible, we used the same all-to-all graph-based approach as for insect Grls to generate a sequence similarity network (Supplementary file 10), and families were identified as clusters in a 2D projection (Figure 4A). We used the gross connectivity of clusters, and the presence or (putative) absence of these proteins across taxa (Figure 4B), to make inferences about the ancestry of these proteins.

In the sequence similarity network, clusters of Ors, Grs and non-insect animal GRLs (excluding GrlHz) were closely-located or intermingled, while insect Grl clusters were more distantly located from this grouping (Figure 4A). GrlHz formed a distinct cluster, but this connects only with the Or/Gr/Grl clusters (and not plant DUF3537 or PHTF clusters) (Figure 4A), suggesting that it descended from a Gr-like ancestor. Given that GrlHz was not detected outside of Holozoa (Figure 4B), the simplest hypothesis is that an ancestral holozoan had a 7TMIC gene that duplicated to produce an ancestral GrlHz and an ancestral Gr (Figure 4C). The diversity of Ors, Grs and Grls would then have resulted from taxon-specific diversification of a single, holozoan branch of a hypothetical 7TMIC superfamily (Figure 4C).

The plant DUF3537 protein cluster was relatively well-connected to the Or/Gr/Grl clusters (Figure 4A), consistent with previously-recognized sequence similarity between DUF3537 and Grs that supported their proposed shared ancestry (Benton, 2015; Benton et al., 2020). If this is correct, a DUF3537/Or/Gr/Grl ancestor must have been present in a common ancestor of plants (part of Diaphoretickes) and animals (part of Amorphea) (Figure 4B-C).

Unicellular eukaryotic 7TMICs were dispersed between Or/Gr/Grl and DUF3537 proteins (Figure 4A); the simplest hypothesis is that these are related to other 7TMICs in accordance with their species' taxonomy (e.g., SAR (stremenopiles, alveolates and Rhizaria) 7TMICs are more closely related to plant DUF3537 proteins than to animal Grs). Alternatively, the generally sparse conservation of unicellular eukaryotic 7TMICs might indicate horizontal gene transfer(s).

Finally, PHTF also forms a separate cluster (Figure 4A), and its broad taxonomic representation argues that the *PHTF* ancestral gene must also have been present in a common Amorphea-Diaphoretickes ancestor (Figure 4B-C). If there was a single ancestral 7TMIC, we hypothesize that this gene must have duplicated in a common eukaryotic ancestor to produce the distinct PHTF and Or/Gr/Grl/DUF3537 lineages (Figure 4C).

Concluding remarks

Exploiting recent advances in protein structure predictions, we have used a tertiary structure-based screening approach to identify new candidate members of the 7TMIC superfamily. While the founder members of this superfamily, insect Ors and Grs, were thought for many years to define an invertebrate-specific protein family (Benton, 2006; Robertson et al., 2003), there is now substantial evidence that these proteins originated in a eukaryotic common ancestor. We also counter previous assumptions that 7TMICs were completely lost in Chordata, through discovery of two lineages within this superfamily: PHTF and GrlHz. Finally, we have identified many previously-overlooked putative chemosensory receptors in *D. melanogaster* (and related flies), despite decades of study of this species' chemosensory systems.

Two important issues remain open. First, are all of the candidate 7TMICs homologous, or does shared tertiary structure reflect convergent evolution in protein folding in at least some cases? Doubts about homology stem, reasonably, from the extreme sequence divergence between 7TMICs to beyond the twilight zone of sequence similarity (Rost, 1999). However, sequence divergence over many millions of years of accumulated amino acid substitutions is well-appreciated in this superfamily (e.g., pairs of *D. melanogaster* Grs can display as little as 8% amino acid identity (Robertson et al., 2003)). Thus, sequence dissimilarity alone is not compelling evidence for structural convergence. Examples of convergent evolution of tertiary protein structures have been described (Alva et al., 2010; Alva et al., 2015; Tomii et al., 2012) but the vast majority of these are small protein domains or motifs, some of which might represent relics of the evolution of proteins from short peptide ancestors (Alva et al., 2015; Lupas et al., 2001). The core of the 7TMIC fold is >300 amino acids, and the question of homology or convergence is most akin to the unresolved, long-standing debate regarding the evolution of the 7TM G protein-coupled receptor fold of type I and type II rhodopsins (Mackin et al., 2014; Rozenberg et al., 2021). In the case of 7TMICs, if PHTFs and other family members are not homologous, their taxonomic representation indicates that structural convergence must have occurred in a eukaryotic common ancestor. While it might be impossible to definitively distinguish homology from convergence, both hypotheses have interesting implications for this protein fold: convergent evolution of at least some 7TMICs from several distinct origins would argue that the fold is an energetically-favorable packing of seven TMs; if the superfamily had

a single origin, this would further highlight the remarkable potential for sequence diversification while maintaining a common tertiary structure (Schaeffer and Daggett, 2011).

Second, what are the biological roles of different 7TMICs? One aspect of this question pertains to their mechanism of action, that is, whether they assemble in multimeric complexes to form ligand-gated ion channels, similar to insect Ors and Grs. The apparent presence of an anchor domain in all 7TMICs, where most inter-subunit contacts occur in Ors (Butterwick et al., 2018; Del Marmol et al., 2021), raises the possibility that complex formation is a common biochemical property. Whether they function as ligand-gated ion channels is not necessarily trivial to answer. Even for insect Grs – for which abundant evidence exists for their *in vivo* requirement in tastant-evoked neuronal activity (Chen and Dahanukar, 2020) – definitive demonstration of their chemical ligand-gated ion conduction properties has (with rare exceptions, e.g., (Morinaga et al., 2022)) been elusive. For PHTF or GrlHz proteins, for example, it is currently difficult to anticipate what might be relevant ligands and we cannot exclude that they have a completely different type of biological activity. Nevertheless, available expression data points to roles of different proteins in specific, but diverse cell types, including chemosensory neurons, (developing) spermatocytes and muscle. The discoveries in this work should stimulate interest in an even broader community of researchers to understand the evolution and biology of 7TMICs.

Acknowledgements

We are very grateful to Julia Santiago for advice on protein structure comparisons and instruction on Coot and Pymol. We acknowledge use of data from the Genotype-Tissue Expression (GTEx) Project, which is supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We thank Roman Arguello, Jamin Letcher, Julia Santiago and members of the Benton laboratory for comments on the manuscript. Research in R.B.'s laboratory is supported by the University of Lausanne, an ERC Advanced Grant (833548) and the Swiss National Science Foundation. N.J.H. is supported by a Human Frontier Science Program Long-Term Postdoctoral Fellowship (LT-0003/2022-L).

Author contributions

R.B. conceived the project, and performed structural screens/analyses and expression analyses. N.J.H. performed sequence-based homolog identification, phylogenetic and network analyses, and gene structure/synteny and protein motif analyses. R.B. and N.J.H. prepared the figures and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Methods

7TMIC candidate homolog identification

Structural screens for candidate 7TMIC homologs were performed with the AF-DB search tool in DALI (ekhidna2.biocenter.helsinki.fi/dali/) (Holm, 2022) using as query the *A. bakeri* Orco structure (PDB 6C70-A) (Butterwick et al., 2018). As of December 2022, this server permitted screening of the structural proteome of 47 phylogenetically-diverse species. Proteins whose structural models had a Z-score >10 were retained for further analysis. Candidate homologs from these screens were assessed first by using these as queries in DALI AF-DB searches of the *D. melanogaster* proteome to ensure Ors and Grs were the best "reverse" hits, and subsequently for secondary structural features using DeepTMHMM (dtu.biolib.com/DeepTMHMM/) (Hallgren et al., 2022) and Phobius (phobius.sbc.su.se/) (Kall et al., 2007). Of the newly identified *D. melanogaster* Grls, we note that three were initially classified as being members of the *Gr* repertoire (Grl36a (Gr36d), Grl43a (Gr43b) and Grl65a (Gr65a), but later excluded (Flybase (flybase.org/) and (Robertson et al., 2003)).

To identify sequences of candidate homologs from other species that were not screened with DALI AF-DB, PSI-BLAST searches against the NCBI refseq_protein database were performed, using the query sequences indicated in each figure and dataset. PSI-BLAST was run with an expected threshold of 1E-10 until convergence. BLASTP searches for Gr36/59 homologs were performed more permissively, using an expected threshold of 0.05.

Structure predictions and analysis

AlphaFold2 protein models (Jumper et al., 2021; Varadi et al., 2022) were downloaded from the AlphaFold Protein Structure Database (alphafold.ebi.ac.uk; release July 2022). For proteins for which structural predictions were not already available, we generated AlphaFold2 models using ColabFold (Mirdita et al., 2022). Positive and negative control protein structures were downloaded from the RCSB Protein Data Bank (PDB codes are indicated in Table 1). Pairwise structural similarities of protein models were quantitatively assessed with DALI (Holm, 2022) and TM-align (zhanggroup.org/TM-align/) (Zhang and Skolnick, 2005). Proteins were aligned to the same coordinate space with Coot (www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/) (Emsley et al., 2010) and visualized in PyMol v2.5.4.

Phylogenetic and network analyses

Sequence databases assembled using PSI-BLAST (see above) were first curated in a semi-automated pipeline. First, sequences annotated as "partial" or "low quality", or which contained ambiguous sequence characters (e.g., X), were removed. CD-HIT (cd-hit.org) (Fu et al., 2012; Li and Godzik, 2006) was used to cluster redundant sequences (100% amino acid identity). Using Phobius TM domain predictions, we removed sequences with fewer than four TMs (this number was chosen to allow for the different sensitivity of Phobius compared to DeepTMHMM). In the final PHTF database, we manually excluded a single sequence as a spurious hit (*Branchiostoma floridae* XP_035670545.1, zinc transporter ZIP10-like); this sequence sorted independently in first pass phylogenetic analyses (via FastTree2 (Price et al., 2010)), and a search via InterPro (ebi.ac.uk/interpro/) revealed that it had no obvious similarity to the other proposed homologs. The database of Gr39/Gr59 homologs was manually curated

due to its relatively small size and accurate automatic annotation by RefSeq; here, we excluded BLAST hits not annotated as Grs, and visually inspected a sequence alignment for good alignment.

To reduce the large curated sequence databases to a size that could be locally analyzed by both maximum likelihood and Bayesian phylogenetic methods, CD-Hit was used to cluster sequences by 70-90% sequence identity, using the longest sequence as the representative for phylogenetic analyses. The clustering used to generate each phylogeny is indicated in the corresponding figure legend.

We took two separate approaches to inferring ancestry. In initial analyses, when comparing insect Grls to Ors/Grs/non-insect GRLs (Figure 3B) or for the entire 7TMIC superfamily (Figure 4A), we observed that extremely low sequence similarity severely constrained our ability to generate meaningful multiple sequence alignments (data not shown). We therefore generated all-to-all sequence similarity networks using MMSeqs2 and inferred phylogenies from these networks by the Graph Splitting method (both implemented in gs2) (Matsui and Iwasaki, 2020). MMSeqs2, as implemented in gs2, employs high sensitivity to sequence similarity and is thus capable of networking non-homologous sequences via spurious sequence identity, should it be present. This method does not distinguish between homology and convergence. Rather, the purpose of this analysis was to make inferences about relatedness under the assumption that all sequences are homologous. In the networks, edge weights are E-values from MMSeqs2. In the Graph Splitting trees, branch support values were generated by the Edge Perturbation method (1000 replicates) with a transfer bootstrap expectation (Lemoine et al., 2018). For visualization of sequence similarity networks, recursive, same-to-same sequence comparisons (resulting in an E-value of 0) were removed using an R script.

For all other trees, multiple sequence alignments were generated by MAFFT. We made no *a priori* assumptions about the alignment, so used default settings. Phylogenetic trees were inferred by maximum likelihood and Bayesian methods. Maximum likelihood trees were generated by IQ-TREE (Minh et al., 2020), using the best model selected for each analysis by ModelFinder (Kalyaanamoorthy et al., 2017) according to the Bayesian Information Criterion, and with bootstrapping by UFBoot2 (1000 replicates) (Hoang et al., 2018). Bayesian trees were generated by MrBayes (Ronquist and Huelsenbeck, 2003) using a mixed amino acid substitution model (Markov chain Monte Carlo analyses run until standard deviation of split frequencies < 0.05, with 25% burn in). To generate the most parsimonious hypotheses of protein evolution, we used NOTUNG (Chen et al., 2000) to rearrange poorly supported branches and resolve polytomies in a species tree-aware fashion (i.e., favoring speciation to gene duplication/horizontal gene transfer in poorly supported branches and polytomies), using default weights/costs (gene duplication 1.5, transfers 3.0, gene loss 1.0). Branches were eligible for rearrangement at branch support values less than UFboot 95 or posterior probability 0.95. Species trees used for rearrangement were based on the NCBI Taxonomy Common Tree, with polytomies randomly resolved for each analysis using the ape (Paradis and Schliep, 2019) and phytools (Revell, 2012) packages. Strict consensus trees were generated by comparing the species tree-aware maximum likelihood and Bayesian trees via the consensus function in ape.

The 7TMIC sequence similarity network was visualized and annotated in

Cytoscape (Shannon et al., 2003). Trees were visualized and annotated in NOTUNG, iTOL (itol.embl.de/) (Letunic and Bork, 2007), and Adobe Illustrator. Consensus sequence illustrations were adapted from figures generated by WebLogo (weblogo.berkeley.edu/) (Crooks et al., 2004).

Synteny and intron mapping

The locations of *Grl36a*, *Grl43a*, *Gr36* and *Gr59c/d* genes in different drosophilids were surveyed using the NCBI Genome Data Viewer (ncbi.nlm.nih.gov/genome/gdv/) (Rangwala et al., 2021). Gene intron-exon structures were manually surveyed using publicly available predictions available on RefSeq (via the Genome Data Viewer) and FlyBase, and visualized in SnapGene. The relative positions of introns were assessed via multiple sequence alignment of the protein sequences; for this analysis, we assumed that that entire sequences could be aligned (global alignment), and thus computed the alignment using the G-INS-i (Needleman-Wunsch) option in MAFFT.

Expression analysis

H. sapiens PHTF1 and *PHTF2* tissue-specific RNA expression data were obtained from the GTEx Portal (GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2; gtexportal.org/home/datasets). Tissue/life-stage-specific RNA expression data of *Phtf* and *Grl* genes in *D. melanogaster* were downloaded from the Fly Atlas 2.0 (motif.mvls.gla.ac.uk/FlyAtlas2) (Krause et al., 2022) or, for the labellum, from (Dweck et al., 2021). *D. melanogaster* scRNA-seq data was from the Fly Cell Atlas (Li et al., 2022): proboscis/maxillary palp (10× stringent dataset) and testis/seminal vesicle (10× relaxed dataset), visualized as HVG tSNE or UMAP plots, respectively, in the Scope interface (scope.aertslab.org/#/FlyCellAtlas) (Davie et al., 2018).

References

Alva, V., Remmert, M., Biegert, A., Lupas, A.N., and Soding, J. (2010). A galaxy of folds. Protein Sci *19*, 124-130.

Alva, V., Soding, J., and Lupas, A.N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. Elife *4*, e09410.

Benton, R. (2006). On the ORigin of smell: odorant receptors in insects. Cell Mol Life Sci *63*, 1579-1585.

Benton, R. (2015). Multigene Family Evolution: Perspectives from Insect Chemoreceptors. Trends Ecol Evol *30*, 590-600.

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.

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.

Bergsten, J. (2005). A review of long-branch attraction. Cladistics *21*, 163-193.

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.

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.

Butterwick, J.A., Del Marmol, 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.

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol *7*, 429-447.

Chen, Y.D., and Dahanukar, A. (2020). Recent advances in the genetic basis of taste detection in *Drosophila*. Cell Mol Life Sci *77*, 1087-1101.

Chi, Y., Wang, H., Wang, F., and Ding, M. (2020). PHTF2 regulates lipids metabolism in gastric cancer. Aging (Albany NY) *12*, 6600-6610.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, L., Aibar, S., Makhzami, S., Christiaens, V., Bravo Gonzalez-Blas, C.*, et al.* (2018). A Single-Cell Transcriptome Atlas of the Aging *Drosophila* Brain. Cell *174*, 982-998 e920.

Del Marmol, J., Yedlin, M.A., and Ruta, V. (2021). The structural basis of odorant recognition in insect olfactory receptors. Nature *597*, 126-131.

Drosophila Odorant Receptor Nomenclature Committee (2000). A unified nomenclature system for the *Drosophila* odorant receptors. Cell *102*, 145-146.

Dunipace, L., Meister, S., McNealy, C., and Amrein, H. (2001). Spatially restricted expression of candidate taste receptors in the *Drosophila* gustatory system. Curr Biol *11*, 822-835.

Dweck, H.K., Talross, G.J., Wang, W., and Carlson, J.R. (2021). Evolutionary shifts in taste coding in the fruit pest *Drosophila suzukii*. Elife *10*, e64317.

Edwards, S.L., Charlie, N.K., Milfort, M.C., Brown, B.S., Gravlin, C.N., Knecht, J.E., and Miller, K.G. (2008). A novel molecular solution for ultraviolet light detection in *Caenorhabditis elegans*. PLOS Biol *6*, e198.

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

Freeman, E.G., and Dahanukar, A. (2015). Molecular neurobiology of *Drosophila* taste. Curr Opin Neurobiol *34*, 140-148.

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.

Gong, J., Yuan, Y., Ward, A., Kang, L., Zhang, B., Wu, Z., Peng, J., Feng, Z., Liu, J., and Xu, X.Z.S. (2016). The *C. elegans* Taste Receptor Homolog LITE-1 Is a Photoreceptor. Cell *167*, 1252-1263 e1210.

Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Almagro Armenteros, J.J., Marcatili, P., Nielsen, H., Krogh, A., and Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv, 2022.2004.2008.487609.

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

Holm, L. (2020). Using Dali for Protein Structure Comparison. Methods Mol Biol *2112*, 29-42.

Holm, L. (2022). Dali server: structural unification of protein families. Nucleic Acids Res, W210-W215.

Huang, X., Geng, S., Weng, J., Lu, Z., Zeng, L., Li, M., Deng, C., Wu, X., Li, Y., and Du, X. (2015). Analysis of the expression of PHTF1 and related genes in acute lymphoblastic leukemia. Cancer Cell Int *15*, 93.

Illergard, 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.

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.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A.*, et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589.

Kall, L., Krogh, A., and Sonnhammer, E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res *35*, W429-432.

Kallberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat Protoc *7*, 1511-1522.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods *14*, 587-589.

Kim, Y., and Kim, S.H. (2020). WD40-Repeat Proteins in Ciliopathies and Congenital Disorders of Endocrine System. Endocrinol Metab (Seoul) *35*, 494-506.

Krause, S.A., Overend, G., Dow, J.A.T., and Leader, D.P. (2022). FlyAtlas 2 in 2022: enhancements to the *Drosophila melanogaster* expression atlas. Nucleic Acids Res *50*, D1010-D1015.

Larsson, M.C., Domingos, A.I., Jones, W.D., Chiappe, M.E., Amrein, H., and Vosshall, L.B. (2004). *Or83b* encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. Neuron *43*, 703-714.

Lemoine, F., Domelevo Entfellner, J.B., Wilkinson, E., Correia, D., Davila Felipe, M., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature *556*, 452-456.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127-128.

Li, H., Janssens, J., De Waegeneer, M., Kolluru, S.S., Davie, K., Gardeux, V., Saelens, W., David, F., Brbić, M., Leskovec, J.*, et al.* (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. Science *375*, eabk2432.

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.

Liu, J., Ward, A., Gao, J.W., Dong, Y.M., Nishio, N., Inada, H., Kang, L.J., Yu, Y., Ma, D., Xu, T.*, et al.* (2010). *C. elegans* phototransduction requires a G protein-dependent cGMP pathway and a taste receptor homolog. Nature Neuroscience *13*, 715-U788.

Lupas, A.N., Ponting, C.P., and Russell, R.B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol *134*, 191-203.

Mackin, K.A., Roy, R.A., and Theobald, D.L. (2014). An empirical test of convergent evolution in rhodopsins. Mol Biol Evol *31*, 85-95.

Manuel, A., Beaupain, D., Romeo, P.H., and Raich, N. (2000). Molecular characterization of a novel gene family (PHTF) conserved from *Drosophila* to mammals. Genomics *64*, 216-220.

Matsui, M., and Iwasaki, W. (2020). Graph Splitting: A Graph-Based Approach for Superfamily-Scale Phylogenetic Tree Reconstruction. Syst Biol *69*, 265-279.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530-1534.

Mirdita, M., Schutze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat Methods *19*, 679-682.

Morinaga, S., Nagata, K., Ihara, S., Yumita, T., Niimura, Y., Sato, K., and Touhara, K. (2022). Structural model for ligand binding and channel opening of an insect gustatory receptor. J Biol Chem *298*, 102573.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Nei, M., Niimura, Y., and Nozawa, M. (2008). The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet *9*, 951-963.

Otaki, J.M., and Yamamoto, H. (2003). Length analyses of *Drosophila* odorant receptors. J Theor Biol *223*, 27-37.

Oyhenart, J., Benichou, S., and Raich, N. (2005a). Putative homeodomain transcription factor 1 interacts with the feminization factor homolog fem1b in male germ cells. Biol Reprod *72*, 780-787.

Oyhenart, J., Dacheux, J.L., Dacheux, F., Jegou, B., and Raich, N. (2005b). Expression, regulation, and immunolocalization of putative homeodomain transcription factor 1 (PHTF1) in rodent epididymis: evidence for a novel form resulting from proteolytic cleavage. Biol Reprod *72*, 50-57.

779 Oyhenart, J., Le Goffic, R., Samson, M., Jegou, B., and Raich, N. (2003). Phtf1 is
780 an integral membrane protein localized in an endoplasmic reticulum domain in
781 maturing male germ cells. Biol Reprod *68*, 1044-1053.
782 Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern
783 phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526-528.
784 Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately
785 maximum-likelihood trees for large alignments. PLOS ONE *5*, e9490.
786 Raich, N., Mattei, M.G., Romeo, P.H., and Beaupain, D. (1999). PHTF, a novel
787 atypical homeobox gene on chromosome 1p13, is evolutionarily conserved.
788 Genomics *59*, 108-109.
789 Rangwala, S.H., Kuznetsov, A., Ananiev, V., Asztalos, A., Borodin, E., Evgeniev,
790 V., Joukov, V., Lotov, V., Pannu, R., Rudnev, D.*, et al.* (2021). Accessing NCBI
791 data using the NCBI Sequence Viewer and Genome Data Viewer (GDV).
792 Genome Res *31*, 159-169.
793 Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology
794 (and other things). Methods in Ecology and Evolution *3*, 217-223.
795 Robertson, H.M. (2015). The Insect Chemoreceptor Superfamily Is Ancient in
796 Animals. Chemical Senses *40*, 609-614.
797 Robertson, H.M. (2019). Molecular Evolution of the Major Arthropod
798 Chemoreceptor Gene Families. Annu Rev Entomol *64*, 227-242.
799 Robertson, H.M., Warr, C.G., and Carlson, J.R. (2003). Molecular evolution of the
800 insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc Natl
801 Acad Sci USA *100 Suppl 2*, 14537-14542.
802 Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic
803 inference under mixed models. Bioinformatics *19*, 1572-1574.
804 Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng *12*,
805 85-94.
806 Rozenberg, A., Inoue, K., Kandori, H., and Beja, O. (2021). Microbial Rhodopsins:
807 The Last Two Decades. Annu Rev Microbiol *75*, 427-447.
808 Saina, M., Busengdal, H., Sinigaglia, C., Petrone, L., Oliveri, P., Rentzsch, F.,
809 and Benton, R. (2015). A cnidarian homologue of an insect gustatory receptor
810 functions in developmental body patterning. Nat Commun *6*, 6243.
811 Sajko, S., Grishkovskaya, I., Kostan, J., Graewert, M., Setiawan, K., Trubestein,
812 L., Niedermuller, K., Gehin, C., Sponga, A., Puchinger, M.*, et al.* (2020).
813 Structures of three MORN repeat proteins and a re-evaluation of the proposed
814 lipid-binding properties of MORN repeats. PLOS ONE *15*, e0242677.
815 Schaeffer, R.D., and Daggett, V. (2011). Protein folds and protein folding. Protein
816 engineering, design & selection : PEDS *24*, 11-19.
817 Scott, K., Brady, R., Jr., Cravchik, A., Morozov, P., Rzhetsky, A., Zuker, C., and
818 Axel, R. (2001). A chemosensory gene family encoding candidate gustatory and
819 olfactory receptors in *Drosophila*. Cell *104*, 661-673.
820 Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin,
821 N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment
822 for integrated models of biomolecular interaction networks. Genome Res *13*,
823 2498-2504.
824 Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein
825 sequence searching for the analysis of massive data sets. Nat Biotechnol *35*,
826 1026-1028.
827 Tomii, K., Sawada, Y., and Honda, S. (2012). Convergent evolution in structural
828 elements of proteins investigated using cross profile analysis. BMC Bioinf *13*, 11.

829 Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G.,
830 Yuan, D., Stroe, O., Wood, G., Laydon, A.*, et al.* (2022). AlphaFold Protein
831 Structure Database: massively expanding the structural coverage of protein-
832 sequence space with high-accuracy models. Nucleic Acids Res *50*, D439-D444.
833 Vidal, B., Aghayeva, U., Sun, H., Wang, C., Glenwinkel, L., Bayer, E.A., and
834 Hobert, O. (2018). An atlas of *Caenorhabditis elegans* chemoreceptor expression.
835 PLOS Biol *16*, e2004218.
836 Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D.
837 (2020). Improved protein structure prediction using predicted interresidue
838 orientations. Proceedings of the National Academy of Sciences of the United
839 States of America *117*, 1496-1503.
840 Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of
841 protein structure template quality. Proteins *57*, 702-710.
842 Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment
843 algorithm based on the TM-score. Nucleic Acids Res *33*, 2302-2309.
844

Figure and table legends

Figure 1. Structure-based screening for 7TMIC homologs.
(A) Top view of a cryo-EM structure of the homotetramer of Orco from *A. bakeri* (PDB 6C70 (Butterwick et al., 2018)), in which one subunit has a spectrum coloration (N-terminus (blue) to C-terminus (red)). The ion channel pore is formed at the interface of the four subunits. A side view is show below. The anchor domain, comprising the cytoplasmic projections of TM4-6 and TM7a, forms most of the inter-subunit interactions in Ors (Butterwick et al., 2018; Del Marmol et al., 2021).
(B) Top: output of transmembrane topology predictions of DeepTMHMM (Hallgren et al., 2022) for *A. bakeri* Orco. Bottom: schematic of the membrane topology of an Orco monomer, with the same spectrum coloration as in (A) (adapted from (Benton et al., 2020)). Note that the seventh predicted helical region is divided into two in the cryo-EM structure: TM7a (located in the cytosol) and TM7b (located in the membrane).
(C) Comparisons of side and top views of the cryo-EM structure of an *A. bakeri* Orco subunit (6C70-A) (left) and an AlphaFold2 protein structure prediction of *A. bakeri* Orco. Helical regions are numbered in the top views. Note the model contains the extracellular loop 2 (EL2) and intracellular loop 2 (IL2) regions that were not visualized in the cryo-EM structure. Quantitative comparisons of structures are provided in Table 1.
(D) Summary of the results of the screen for Orco/Or-like protein folds in the AlphaFold Protein Structure Database for the indicated species using DALI (Holm, 2022). The threshold of DALI Z-score >10 was informed by inspection of the results of the screen (see Results). Raw output of the screen are provided in Supplementary file 2.
(E) Top: transmembrane topology predictions of the single screen hits from the Trypanosoma species *Leishmania infantum* and *Trypanosoma brucei brucei*. Bottom: AlphaFold2 structural models of these proteins, displayed as in (C). The long N-terminal region contains tandem MORN repeats and sequence of unknown structure (gray); these are masked in the top view of the models.
(F) Visual comparison of the *L. infantum* GRL1 AlphaFold2 model (the N-terminal region is masked) with the *A. bakeri* Orco structure, aligned with Coot (Emsley et al., 2010). Quantitative comparisons of structures are provided in Table 1.
(G) Consensus phylogeny of putative trypanosome homologs. The primary sequence database was assembled using *Leishmania infantum* GRL1 (XP_001464500.1) and *Trypanosoma brucei brucei* GRL1 (XP_845058.1) as query sequences (highlighted in bold). Branch support values refer to maximum likelihood UFboot/Bayesian posterior probabilities. Note that although the *T. cruzi* homolog (XP_803355.1) was not identified in the original DALI screen, visual inspection of the corresponding AlphaFold2 model (A0A2V2WL40) revealed the same global fold. The sequence database, alignment and trees are available in Supplementary file 4.

Figure 2. PHTF proteins are candidate vertebrate 7TMICs.
(A) DeepTMHMM-predicted transmembrane topology of PHTF proteins.
(B) Top: AlphaFold2 predicted structure of *H. sapiens* PHTF1; in the image on the right the long N-terminal region (NTR) and intracellular loop 1 (IL1) are highlighted in blue; these sequences contain a few predicted helical regions but

are of largely unknown structure. Bottom: visual comparison of the *H. sapiens* PHTF1 AlphaFold2 structure (in which the NTR and IL1 are masked) with the *A. bakeri* Orco structure.
(C) AlphaFold2 structures of PFTH proteins in which the NTR and IL1 are masked. Quantitative comparisons of these structures to the cryo-EM Orco structure are provided in Table 1.
(D) Major taxa/species in which a PHTF homolog was identified (see sequence databases in Supplementary file 5). Silhouette images in this and other figures are from Phylopic (phylopic.org/).
(E) Phylogenies of a representative set of PHTF sequences. The sequence database was constructed using the *D. melanogaster* and *H. sapiens* PHTF query sequences. Top left: maximum likelihood phylogeny (JTT+R10 model) and Bayesian phylogeny. The scale bars represent average number of substitutions per site. Bottom left: phylogenies where weakly supported branches (<95/0.95) have been rearranged and polytomies resolved in a species tree-aware manner. Right: strict consensus of the species tree-aware phylogenies. There is a single eukaryotic PHTF clade and the PHTF1-2 split occurred in the jawed vertebrate lineage. However, this interpretation relies on the rearrangement of the weakly supported jawless vertebrate PHTF branch. Therefore, an alternative, but weakly supported, hypothesis is that the duplication occurred in a common vertebrate ancestor, and a single PHTF copy was lost in jawless vertebrates. Select branch support values are present on key branches and refer to maximum likelihood UFboot/Bayesian posterior probabilities. Asterisks indicate that branch support was below the threshold for species-aware rearrangement. The fully annotated trees are available in Figure 2—figure supplements 1-3, and the sequence database, clustered database, alignment, and trees are available in Supplementary file 5.
(F) Summary of tissue-enriched RNA expression of *H. sapiens PHTF1* and *PHTF2* (data are from the GTex Portal; the fully annotated dataset is provided in Figure 2—figure supplement 4) and *D. melanogaster Phtf* (data from the Fly Atlas 2.0; the fully annotated dataset is provided in Figure 2—figure supplement 5).
(G) Left: Uniform Manifold Approximation and Projection (UMAP) representation of RNA-seq datasets from individual cells of the *D. melanogaster* testis and seminal vesicle generated as part of the Fly Cell Atlas (10× relaxed dataset) (Li et al., 2022) colored for expression of *Phtf*. Simplified annotations of cell clusters displaying the highest levels of *Phtf* expression are adapted from (Li et al., 2022); unlabeled clusters represent non-germline cell types of the testis.

Figure 3. Insect Grls are highly divergent, candidate chemosensory receptors.
(A) Proposed nomenclature of *D. melanogaster* Grls (with original gene name and cytological location in parentheses), with corresponding DeepTMHMM-predicted transmembrane topologies and AlphaFold2 structural models. Note that TM7 is not predicted for Grl36b and Grl58a by DeepTMHMM, but is predicted – with the characteristic TM7a/7b split – in the structural model (as well as predicted by Phobius (data not shown)). Quantitative comparisons of these structures to the cryo-EM Orco structure are provided in Table 1.
(B) Sequence similarity network of Grls, Grs and Ors (including Orco). The network was generated using an all-to-all comparison made by MMSeqs2 as implemented by gs2. The connections represent E-values where the weakest

connections (arbitrarily defined as edge weights >1) are colored in lighter gray. Lack of connection between two nodes indicates that those two sequences could not be identified as having any significant sequence similarity under the most sensitive MMSeqs2 settings. Nodes and edges are arranged in a prefuse force-directed layout. The sequence database (detailed in Figure 3—figure supplement 5), sequence similarity network, and a subsequent Graph Splitting tree are available in Supplementary file 6. The Graph Splitting tree is visualized in Figure 3—figure supplement 5; however, we do not place high confidence in the phylogenetic accuracy of the tree due to the likely effects of long branch attraction. The evolution of GrlHz is described in Figure 3—figure supplement 1, with detailed phylogenies in Figure 3—figure supplements 2-4 and corresponding data in Supplementary file 9.

(C) Schematic of the gene arrangement of *Grl36a* and *Gr36* homologs in drosophilids. Color coding reflects relatedness with respect to major speciation and gene duplication events; colors match the phylogenetic tree branches in Figure 3—figure supplement 6B-C. The *Drosophila* subgenus entirely lacks Gr36 homologs (see Figure 3—figure supplement 6).

(D) Alignment of the C-terminal region of *D. melanogaster* Orco, Gr64a, select insect Gr36/Gr59 homologs and *D. melanogaster* Grl36a and Grl43a, extracted from a larger alignment available in Supplementary file 7. The black bar shows the common location of a phase 0 intron, which is presumably homologous in different sequences. The canonical TM7 motif of the Gr family (represented as relative amino acid frequencies extracted from WebLogo) is shown above the sequence, and the variant motifs of different Gr or Grl ortholog groups are shown below.

(E) Phylogenies of Gr36, Gr59c/d, Grl36a, Grl43a and homologous non-drosophilid sequences (color-coded as in (D)). The sequence database was assembled using *D. melanogaster* Gr36a, Grl36a, and Grl43a as the query sequences. Top left: maximum likelihood phylogeny (JTT+F+R7 model) and Bayesian phylogeny. The scale bars represent average number of substitutions per site. Bottom left: phylogenies where weakly supported branches (<95/0.95) have been rearranged and polytomies resolved in a species tree-aware manner. Right: strict consensus of the species tree-aware phylogenies. These analyses support that Gr36 and Grl36a/43a are sister clades, which likely split after Gr59c/d diverged from the ancestral lineage. Sequences are colored as in (D). Select branch support values are present on key branches and refer to maximum likelihood UFboot and Bayesian posterior probabilities, in this order. Asterisks indicate that branch support was below the threshold for species-aware rearrangement. A simplified schematic of gene duplication and loss is illustrated in Figure 3—figure supplement 6F. The fully annotated trees are available in Figure 3—figure supplements 7-9, and the sequence database, alignment, and trees are available in Supplementary file 8.

(F) Histogram of *Gr* and *Grl* expression levels in adult proboscis and maxillary palps determined by bulk RNA-seq. Mean values ±SD of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) are plotted; n = 3 biological replicates. Data is from (Dweck et al., 2021).

(G) Left: t-distributed stochastic neighbor embedding (tSNE) representation of RNA-seq datasets from individual cells of the *D. melanogaster* proboscis and maxillary palp – generated as part of the Fly Cell Atlas (10× stringent dataset) (Li et al., 2022) – colored for expression of the indicated genes. *Gr64f* and *Gr66a* are

broad markers of "sweet/appetitive" and "bitter/aversive" gustatory sensory neurons, respectively. Transcripts for three Grls are detected in subsets of bitter/aversive neurons. Annotations of cell clusters are adapted from (Li et al., 2022); unlabeled clusters represent other non-gustatory sensory neuron or non-neuronal cell types of this tissue.

Figure 4. A hypothesis for the evolution of the 7TMIC superfamily.
(A) Sequence similarity network of the 7TMIC superfamily, generated using the same Ors and Grs from Figure 3B, unicellular eukaryotic Grls from (Benton et al., 2020), and sequence databases assembled using the following query sequences: *N. vectensis* GRL1, *D. melanogaster* Grls and Phtf, *H. sapiens* PHTF1 and PHTF2, *A. thaliana* DUF3537, *C. elegans* SRRs and trypanosome GRLs. The network was generated and visualized as in Figure 3B. The sequence database, sequence similarity network, and a subsequent Graph Splitting tree are available in Supplementary file 10. The Graph Splitting tree is visualized in Figure 4—figure supplement 1.
(B) Presence and absence of 7TMICs across taxa: "other animal Grl" refers to Grls in non-insect animal species previously identified by primary sequence similarity (Benton, 2015; Robertson, 2015; Saina et al., 2015), and nematode SRRs. The dashed branch represents several collapsed paraphyletic clades.
(C) Model of 7TMIC superfamily evolution. The dashed branches represent several collapsed paraphyletic clades and speciation events. The trypanosome 7TMICs are unplaced, due to the currently unresolved taxonomy of trypanosomes (Burki et al., 2020).

Table 1. Quantitative structural comparisons of candidate 7TMIC homologs.
Summary of % amino acid identity, DALI Z-score and TM-align TM-score of the indicated experimentally-determined or *ab initio*-predicted structures of 7TMIC homologs (or negative-control, unrelated proteins) compared to *A. bakeri* Orco. The cryo-EM structure chain A (6C70-A) (Butterwick et al., 2018) was used as the query in all comparisons. Protein models are provided in Supplementary file 1. Note the nomenclature of unicellular eukaryotic 7TMICs is tentative; identical names (e.g., GRL1) do not imply orthology. Typically, a Z-score >20 indicates that the two proteins being compared are definitely homologous, 8-20 that they are probably homologous, and 2-8 is a "gray area" influenced by protein size and fold (Holm, 2020). TM-scores of 0.5-1 indicate that the two proteins being compared adopt generally the same fold, while TM-scores of 0-0.3 indicate random structural similarity (Zhang and Skolnick, 2004, 2005).

Supplementary Figures

Figure 2—figure supplement 1. Fully annotated phylogenetic trees for PHTF homologs.
Sequences are from the protein sequence database generated using *D. melanogaster* Phtf and *H. sapiens* PHTF1/2, and are representatives of clusters of 90% sequence identity. For maximum likelihood, the tree was generated using a JTT+R10 substitution model. Branch support values for maximum likelihood (UFboot) and Bayesian analyses (posterior probability) are shown at the branches. The sequence database, clustered database, alignment and trees are available in Supplementary file 5.

Figure 2—figure supplement 2. **Fully annotated species-aware trees for PHTF homologs.**
Trees are based on the maximum likelihood (left) and Bayesian (right) trees. Branches without support values were eligible for rearrangement. Tree files are available in Supplementary file 5.

Figure 2—figure supplement 3. **Strict consensus of the species-aware trees for PHTF homologs.**
The tree file is available in Supplementary file 5.

Figure 2—figure supplement 4. **Tissue-specific RNA expression of *H. sapiens PHTF1* and *PHTF2***
Plot of RNA expression levels (transcripts per million (TPM)) from the indicated tissues is from the GTEx Portal (GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2).

Figure 2—figure supplement 5. Tissue-specific RNA expression of *D. melanogaster Grl*s and *Phtf*
Heatmap plot of the expression of *D. melanogaster Phtf* and *Grl*s in the indicated tissues/life stages/sexes determined by bulk RNA-seq; fragments per kilobase of exon per million mapped fragments (FPKM) values are shown; data are from the Fly Atlas 2.0 (Krause et al., 2022).

Figure 3—figure supplement 1. **Evolution of GrlHz, a family of Grl 7TMIC not restricted to flies.**
(A) Major taxa/species for which a GrlHz homolog was recovered.
(B) Phylogenies of a representative set of GrlHz sequences (clustered by 70% sequence identity). The sequence database was assembled using *D. melanogaster* GrlHz as the query sequence. Top: Maximum likelihood phylogeny and Bayesian phylogeny. The scale bars represent average number of substitutions per site. Bottom: Phylogenies where weakly supported branches (<95/0.95) have been rearranged and polytomies resolved in a species tree-aware manner. Right: Strict consensus of the species tree-aware phylogenies. The sequence database, clustered database, alignment, and trees are available in Supplementary file 9. The fully annotated trees are visualized in Figure 3—figure supplements 2-4.
(C) Left: the single holozoan copy hypothesis of GrlHz evolution. Under this scenario, a single GrlHz is widely conserved across Holozoa, but has been independently duplicated/lost several times in various taxa. Right: the two-paralog hypothesis of GrlHz evolution. As both the maximum likelihood and Bayesian phylogenies evidence two GrlHz clades, and because some species have two substantially divergent GrlHz sequences, it is possible that there was a gene duplication event early in the evolution of Holozoa.
(D) Examples of GrlHz structures. Of 196 representative sequences, 31 sequences (mostly from Hymenoptera and Lepidoptera) bear N-terminal WD40 repeats.

Figure 3—figure supplement 2. Fully annotated phylogenetic trees for GrlHz homologs.

For maximum likelihood, the tree was generated using a JTT+F+R9 substitution model. Branch support values for maximum likelihood (UFboot) and Bayesian analyses (posterior probability) are shown at the branches. Tree files are available in Supplementary file 9.

Figure 3—figure supplement 3. Fully annotated species-aware trees for GrlHz homologs.
Trees are based on the maximum likelihood (left) and Bayesian (right) trees. Branches without support values were eligible for rearrangement. Tree files are available in Supplementary file 9.

Figure 3—figure supplement 4. **Strict consensus of the species-aware trees for GrlHz homologs.**
The tree file is available in Supplementary file 9.

Figure 3—figure supplement 5. **Fully annotated Graph Splitting tree for newly identified Grls, Ors and Grs.**
Key Edge Perturbation support values are visible on branches; full annotation is available in Supplementary file 6. The primary sequence databases were assembled using each of the *D. melanogaster* Grls as query sequences. *D. melanogaster* Or and Gr sequences were manually collected from FlyBase. Sequences from *M. hrabei* (jumping bristletail), *Thermobia domestica* (firebrat), *Ladona filva* (dragonfly), and *Ephemera danica* (green drake mayfly) were added, following the proposal that canonical Ors may have diversified after the emergence of Neoptera (most winged insects) (Brand et al., 2018). 2,498 additional sequences were collected using the *Nematostella vectensis* GRL1 query sequence (XP_048580785.1); the PSI-BLAST searches were stopped at four iterations, as the search had substantially recovered insect Gr sequences, and further searches returned tens of thousands of sequences. The basal placement of the Grls is unusual given their conservation in flies, as this would suggest they diversified in a common animal ancestor and that the Grls were lost in all animal taxa except flies. This hypothesis seems unlikely given the extreme number of independent gene loss events this would require, and we therefore suspect that this tree topology represents a phylogenetic error, for example, long branch attraction (Bergsten, 2005). The inset shows major collapsed clades, where the tip node is sized proportionally to the number of sequences collapsed. The tree file is available in Supplementary file 6

Figure 3—figure supplement 6. The evolution of Gr36, Gr59, Grl36a and Grl43a.
(A) Schematic of the gene arrangement of *Grl36a* and *Gr36* homologs in drosophilids, with colors matching trees in B and C. This panel is reproduced from Figure 3C.
(B) Species-aware Bayesian phylogeny of Grl36. The sequence database, alignment and trees (including a similar maximum likelihood tree) are available in Supplementary file 8.
(C) Species tree-aware Bayesian phylogeny of Gr36. The sequence database, alignment and trees (including a similar maximum likelihood tree) are available in Supplementary file 8.
(D) Phylogenies of Gr36, Gr59, Grl36a, Grl43a and other homologous

sequences. The sequence database was assembled using *D. melanogaster* Gr36a, Grl36a and Grl43a as query sequences. Top: Maximum likelihood phylogeny and Bayesian phylogeny. Bottom: Phylogenies where weakly supported branches (<95/0.95) have been rearranged and polytomies resolved in a species tree-aware manner. The sequence database, alignment and trees are available in Supplementary file 8.
(E) Strict consensus of the species tree-aware phylogenies. These analyses support that Gr36 and Grl36a/43a are sister clades, which likely split after the Gr59 split. The tree file is available in Supplementary file 8.
(F) Proposed model of Gr36, Gr59, Grl36a and Grl43a evolution.

Figure 3—figure supplement 7. Fully annotated phylogenetic trees for Gr36, Gr59, Grl36a and Grl43a homologs.
For maximum likelihood, the tree was generated using a JTT+F+R7 substitution model and is rooted. Branch support values for maximum likelihood (UFboot) and Bayesian analyses (posterior probability) are shown at the branches. Non-drosophilid sequences assumed to be the outgroup. Tree files are available in Supplementary file 8.

Figure 3—figure supplement 8. **Fully annotated species-aware trees for Gr36, Gr59, Grl36a and Grl43a homologs.**
Trees are based on the maximum likelihood (left) and Bayesian (right) trees. Branches without support values were eligible for rearrangement. Tree files are available in Supplementary file 8.

Figure 3—figure supplement 9. **Strict consensus of the species-aware trees for Gr36, Gr59, Grl36a and Grl43a homologs.**
Although the consensus tree has a polytomy near the emergence of Gr59, this is strictly due to disagreement as to whether the lone *Scaptodrosophila* sequence is a Gr59 homolog or an outgroup to all other *Drosophila*/*Sophophora* sequences shown here. The tree file is available in Supplementary file 8.

Figure 4—figure supplement 1. Graph Splitting tree for the proposed 7TMIC superfamily.
Key Edge Perturbation support values are visible on branches. The inset shows major collapsed clades, where the triangular tip is sized proportionally to the number of sequences collapsed. This tree suggests a different branching pattern than the hypothesis in Figure 4C, consistent with a more complex duplication/loss history for the 7TMIC superfamily. However, as in Figure 3—figure supplement 5, we suspect long branch attraction is present in this analysis, at least for the fly and nematode Grls. The tree file is available in Supplementary file 10.

Supplementary files

Supplementary file 1. AlphaFold2 models.
Models of proteins analyzed in this work, either downloaded from the AlphaFold Protein Structure Database or, where not already available, predicted using the AlphaFold2 algorithm implemented in ColabFold (Mirdita et al., 2022). The four-letter code in the filename represents the first letter of the genus and the first three letters of the species (e.g., “Dmel” = *D. melanogaster*); species names are

given in full in the figures.

Supplementary file 2. DALI screen search results.
Individual text files represent the output of the DALI AF-DB search using *A. bakeri* ORCO chain A (PDB 6C70-A) as query and the structural proteome dataset of the indicated species. The four-letter codes in the file names and job titles is as described for Supplementary file 1.

Supplementary file 3. Reverse DALI search results.
Individual text files represent the output of the DALI AF-DB search using the indicated query candidate 7TMICs from *Trypanosoma* (GRL1), *D. melanogaster* (Grls) or *H. sapiens* (PHTF1/2) and the structural proteomic dataset of *D. melanogaster*.

Supplementary file 4. Trypanosome 7TMIC sequence database, alignment and phylogenetic trees.

Supplementary file 5. PHTF sequence database, clustered sequence database, alignment, phylogenetic trees, species-aware trees and consensus tree.

Supplementary file 6. Or, Gr Or and Grl sequence database, all-to-all sequence similarity network, sequence similarity network annotation file and Graph Splitting tree.
In the sequence similarity network file, the first column is the source node, the second column the target node and the third column the E-value edge weight. The node identifiers can be matched to identifiers in the annotation file.

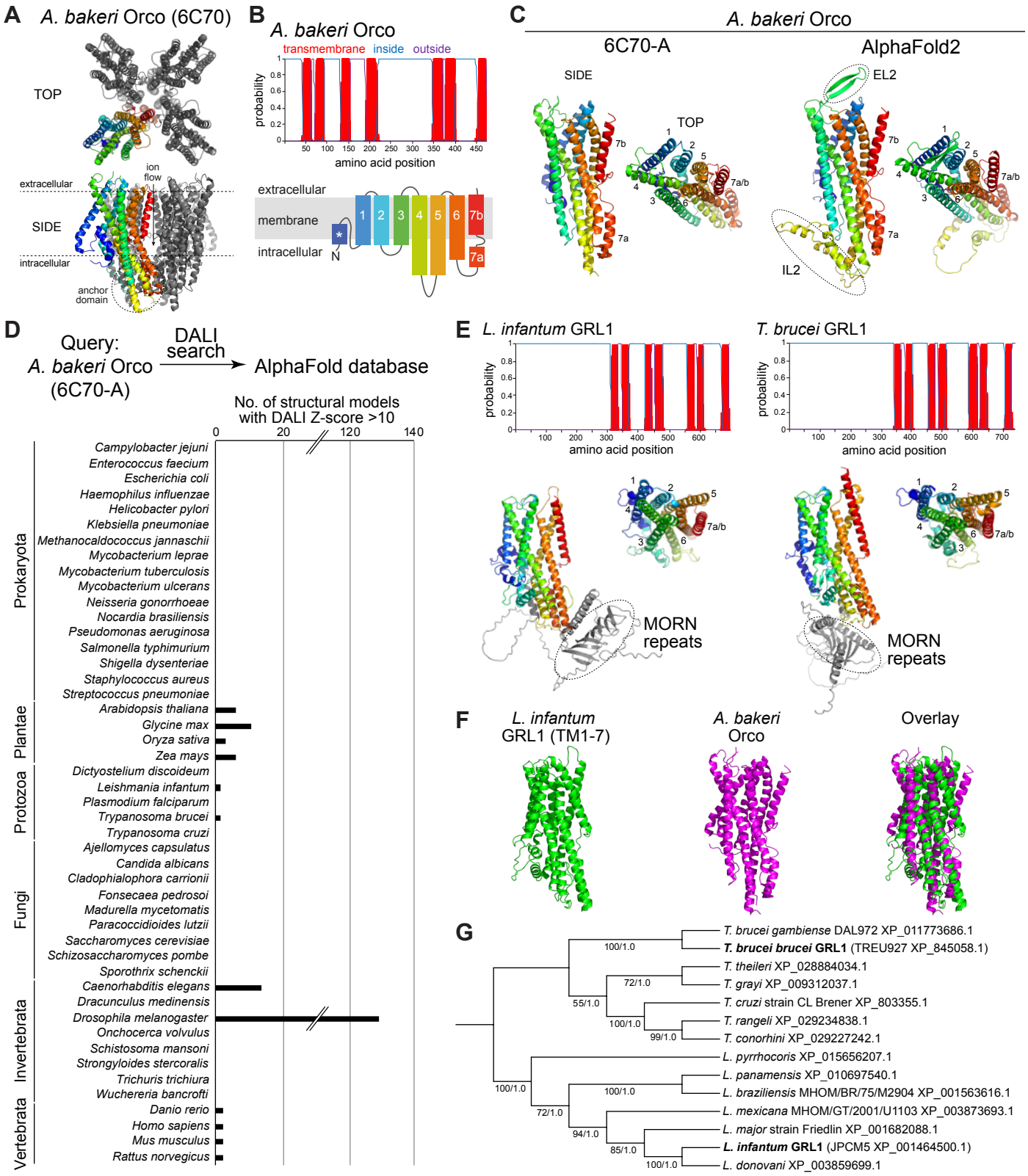
Supplementary file 7. Alignment used for illustrating intron and motif conservation in Grl36a and Grl43a.

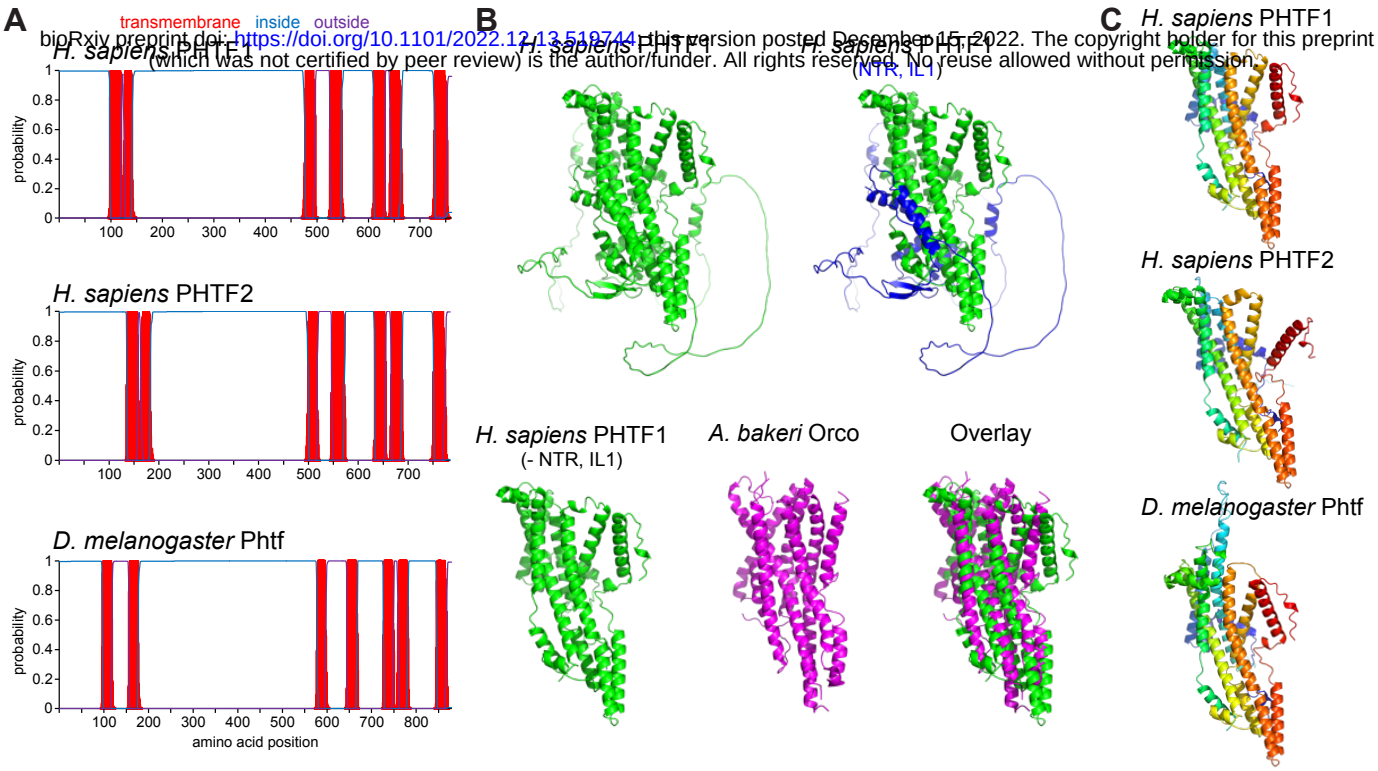
Supplementary file 8. Gr36, Gr59, Grl36a and Grl43a sequence database, alignment, phylogenetic trees, species-aware trees and consensus tree, as well as sequence databases, alignments, phylogenetic trees and species-aware trees for Figure 3—figure supplement 6.

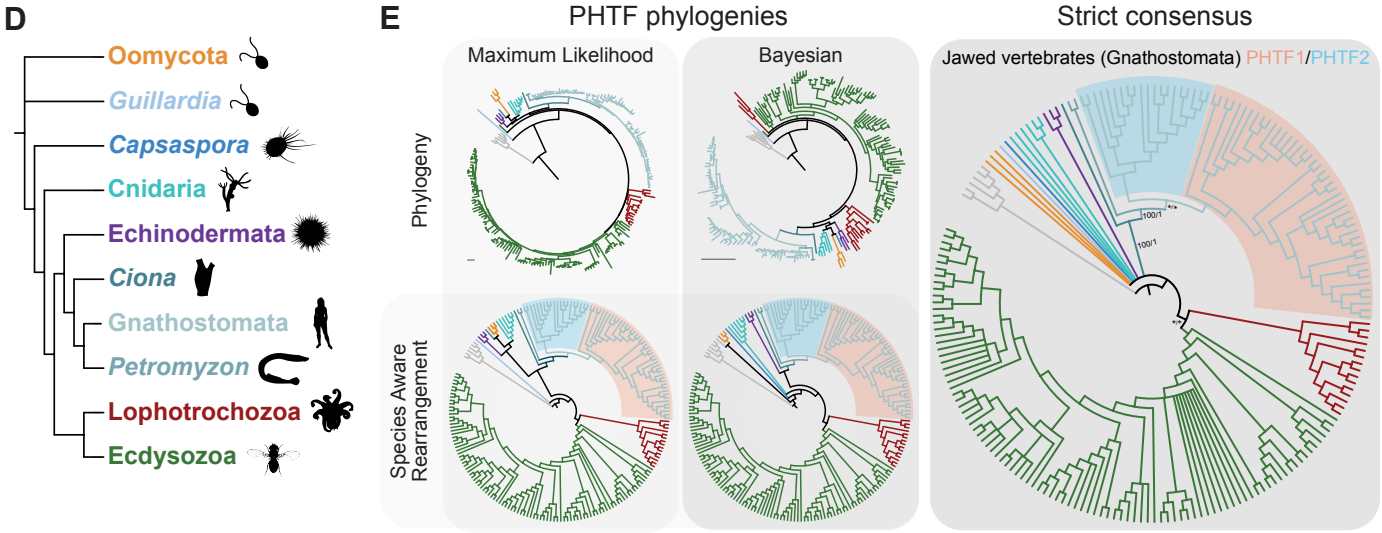
Supplementary file 9. GrlHz sequence database, alignment, phylogenetic trees, species-aware trees and consensus tree.

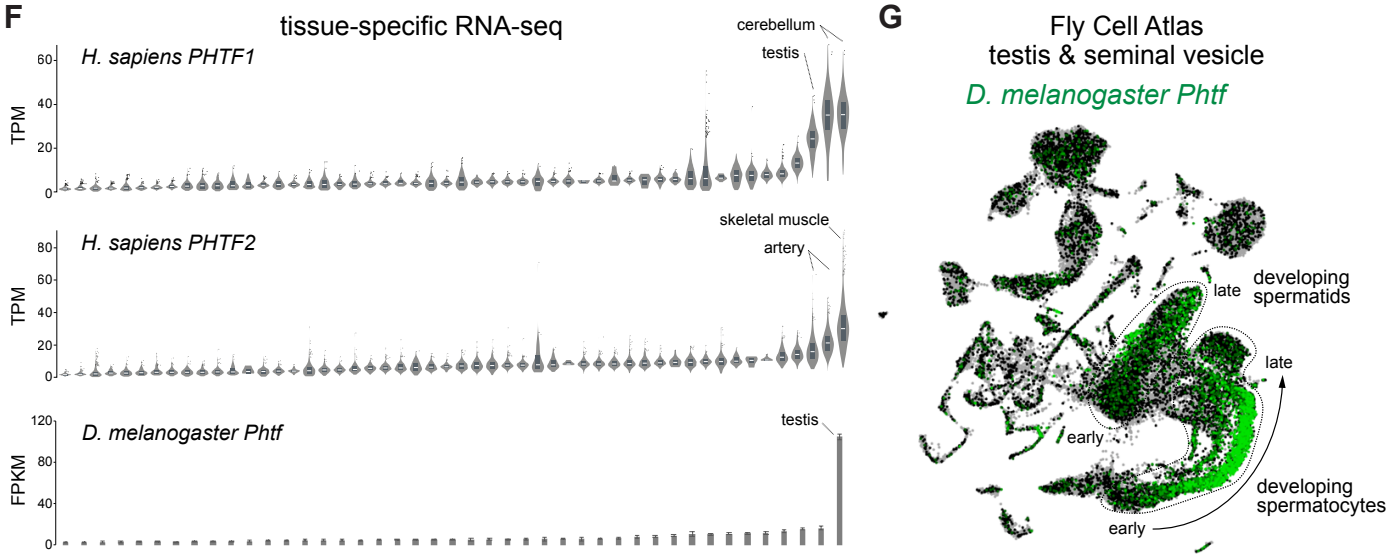
Supplementary file 10. 7TMIC superfamily sequence database, all-to-all sequence similarity network, sequence similarity network annotation file and Graph Splitting tree.
In the sequence similarity network file, the first column is the source node, the second column the target node and the third column the E-value edge weight. The node identifiers can be matched to identifiers in the annotation file.

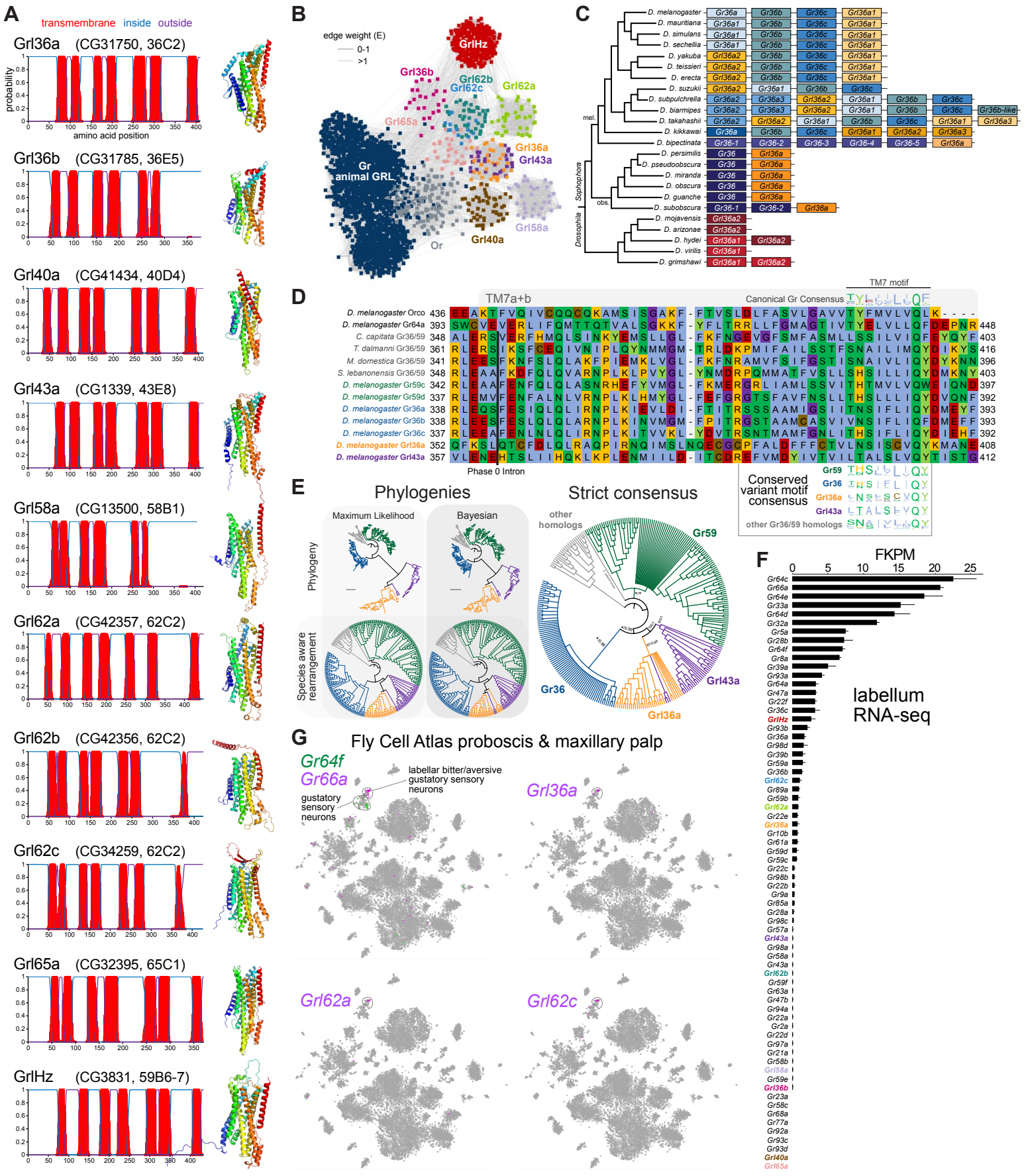
Supplementary file 11. All uncurated PSI-BLAST sequence databases.
Each of the FASTA filenames is formatted as follows (with the exception of the *D. melanogaster* Or and Gr sequences, which were collected manually from FlyBase): ProteinFamily-QuerySpecies-QuerySequence.fasta.











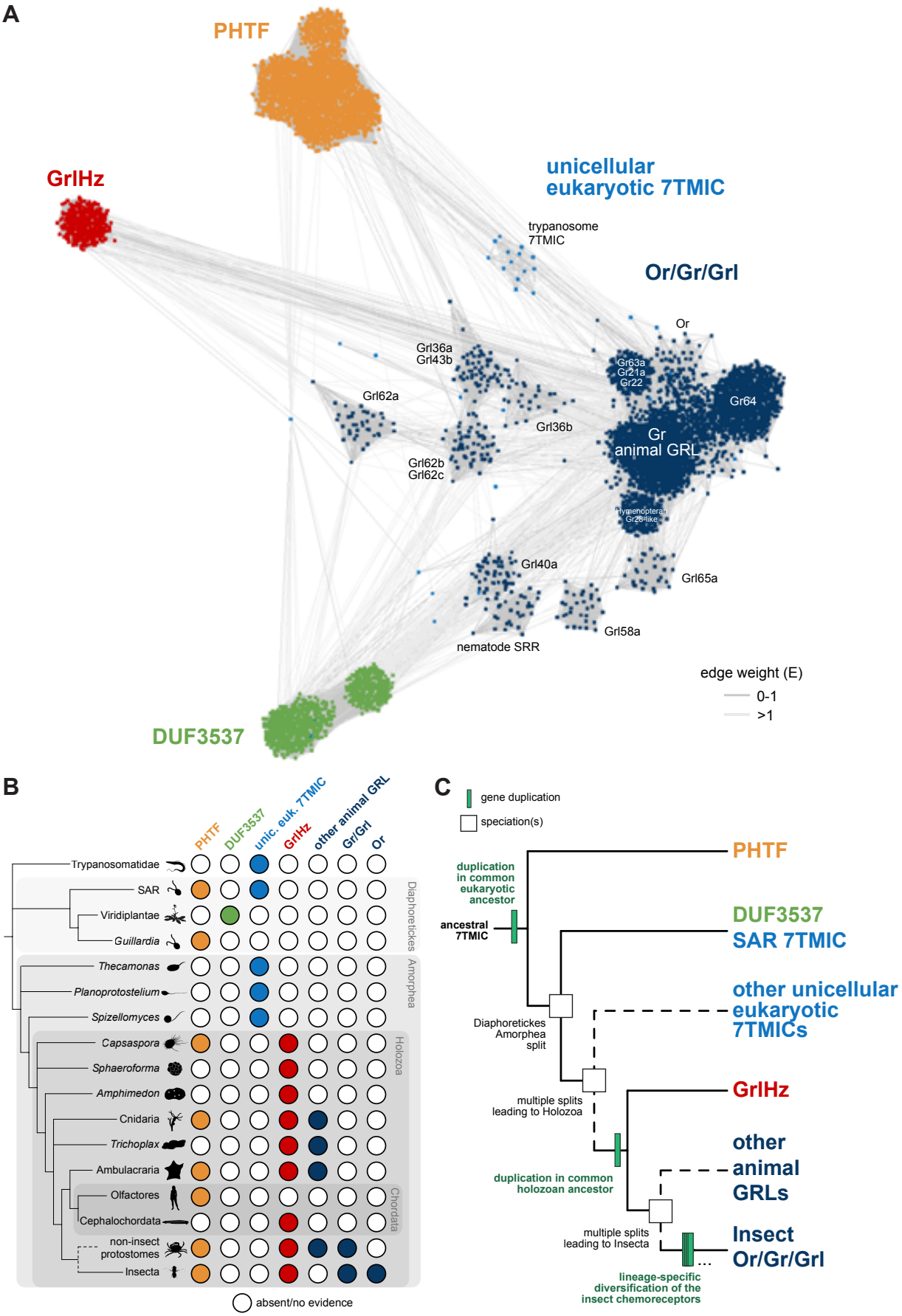


Table 1. Quantitative structural comparisons of candidate 7TMIC homologs

				Comparison to *A. bakeri* Orco (6C70-A)		
Category	**Protein**	**Model or PDB**	**Method or algorithm**	**Amino acid identity (%)**	**DALI Z-score**	**TM-align TM-score**
Positive controls (known 7TMIC)	*A. bakeri* Orco	61b81_unrelaxed_rank_1_model_2	AlphaFold2	100	50.7	0.96
	M. hrabei Or5	7LIC-A	cryo-EM	19	36.3	0.81
	D. melanogaster Gr64a	AF-P83293-F1-model_v4	AlphaFold2	13	29.6	0.79
	N. vectensis GRL1	AF-A7S7G0-F1-model_v4	AlphaFold2	10	31.3	0.78
Unicellular eukaryotic 7TMIC	*T. trahens* GRL1	AF-A0A0L0DUY0-F1-model_v3	AlphaFold2	9	23.2	0.71
	T. trahens GRL2	AF-A0A0L0DQC1-F1-model_v3	AlphaFold2	12	25.3	0.70
	T. trahens GRL3	AF-A0A0L0D5B5-F1-model_v3	AlphaFold2	14	13.1	0.50
	T. trahens GRL4	AF-A0A0L0D5H0-F1-model_v3	AlphaFold2	9	9.9	0.53
	T. trahens GRL5	AF-A0A0L0DD38-F1-model_v3	AlphaFold2	10	12.2	0.56
	T. trahens GRL6	AF-A0A0L0DJ52-F1-model_v3	AlphaFold2	8	15.6	0.57
	V. brassicaformis GRL1	AF-A0A0G4FIT4-F1-model_v3	AlphaFold2	10	9.1	0.47
	V. brassicaformis GRL2	AF-A0A0G4ECU2-F1-model_v3	AlphaFold2	11	14.4	0.57
	V. brassicaformis GRL3	AF-A0A0G4FWI7-F1-model_v3	AlphaFold2	14	23.8	0.74
	V. brassicaformis GRL4	AF-A0A0G4EU86-F1-model_v3	AlphaFold2	10	18.5	0.70
	V. brassicaformis GRL5	AF-A0A0G4FBY6-F1-model_v3	AlphaFold2	10	18.5	0.68
	V. brassicaformis GRL6	AF-A0A0G4G8W6-F1-model_v3	AlphaFold2	8	21.4	0.70
	M. pusilla GRL1	AF-C1MGH9-F1-model_v3	AlphaFold2	12	11.3	0.60
	C. primus GRL1	AF-A0A5B8MFA4-F1-model_v3	AlphaFold2	10	18.1	0.71
	L. infantum GRL1	AF-A4HWQ9-F1-model_v3	AlphaFold2	6	13.5	0.64
	T.brucei GRL1	AF-Q57U78-F1-model_v3	AlphaFold2	9	13.4	0.62
Fly Grl	*D. melanogaster* Grl36a	AF-Q8INZ1-F1-model_v3	AlphaFold2	9	19.5	0.67
	D. melanogaster Grl36b	AF-Q8INY2-F1-model_v3	AlphaFold2	8	15.2	0.62
	D. melanogaster Grl40a	AF-Q0E8M7-F1-model_v3	AlphaFold2	8	19.5	0.66
	D. melanogaster Grl43a	AF-Q9V4Q0-F1-model_v3	AlphaFold2	10	19.9	0.69
	D. melanogaster Grl58a	AF-Q9W2A4-F1-model_v3	AlphaFold2	8	15.0	0.60
	D. melanogaster Grl62a	AF-B7Z0I0-F1-model_v3	AlphaFold2	8	19.4	0.69
	D. melanogaster Grl62b	AF-B7Z0I1-F1-model_v3	AlphaFold2	11	19.1	0.66
	D. melanogaster Grl62c	AF-Q6ILZ2-F1-model_v3	AlphaFold2	10	17.2	0.63
	D. melanogaster Grl65a	AF-Q8IQ72-F1-model_v3	AlphaFold2	11	25.9	0.74
	D. melanogaster GrlHz	AF-Q9W1W8-F1-model_v3	AlphaFold2	7	22.5	0.74
PHTF	*H. sapiens* PHTF1	AF-Q9UMS5-F1-model_v3	AlphaFold2	7	12.9	0.63
	H. sapiens PHTF2	AF-Q8N3S3-F1-model_v3	AlphaFold2	8	12.0	0.62
	D. melanogaster Phtf	AF-Q9V9A8-F1-model_v3	AlphaFold2	5	11.8	0.63
Negative controls (non-7TMIC)	*B. taurus* Rhodopsin	1F88-A	X-ray crystal	9	2.1	0.31
	C. reinhardtii ChR2	6EID-A	X-ray crystal	7	3.6	0.27
	H. sapiens Frizzled4	6BD4	X-ray crystal	8	4.0	0.34
	H. sapiens AdipR	5LXG	X-ray crystal	2	3.6	0.29
	E. coli GlpG	2XOV	X-ray crystal	5	3.5	0.27
	M. musculus TRPV3	6LGP-D	cryo-EM	10	2.7	0.27
	M. musculus PIEZO	6BPZ-B	cryo-EM	5	4.0	0.27
	B. taurus CNGA/CNGB	7O4H-A	cryo-EM	9	2.8	0.24