

RESEARCH ARTICLE

# Evolution of Pentameric Ligand-Gated Ion Channels: Pro-Loop Receptors

Mariama Jaiteh<sup>‡</sup>, Antoine Taly, Jérôme Héning\*

Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, CNRS and Université Paris Diderot, Paris, France

<sup>‡</sup> Current address: Uppsala Biomedicinska Centrum, Uppsala, Sweden

\* [henin@ibpc.fr](mailto:henin@ibpc.fr)



**OPEN ACCESS**

**Citation:** Jaiteh M, Taly A, Héning J (2016) Evolution of Pentameric Ligand-Gated Ion Channels: Pro-Loop Receptors. PLoS ONE 11(3): e0151934. doi:10.1371/journal.pone.0151934

**Editor:** Stéphanie Bertrand, Laboratoire Oceanologique de Banyuls sur Mer, FRANCE

**Received:** October 7, 2015

**Accepted:** March 7, 2016

**Published:** March 17, 2016

**Copyright:** © 2016 Jaiteh et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** Agence Nationale de la Recherche grant PentaGate (<http://www.agence-nationale-recherche.fr/?Projet=ANR-13-BSV8-0020>) benefitted all three authors by funding missions for AT and JH and a stipend for MJ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Pentameric ligand-gated ion channels (pLGICs) are ubiquitous neurotransmitter receptors in Bilateria, with a small number of known prokaryotic homologues. Here we describe a new inventory and phylogenetic analysis of pLGIC genes across all kingdoms of life. Our main finding is a set of pLGIC genes in unicellular eukaryotes, some of which are metazoan-like Cys-loop receptors, and others devoid of Cys-loop cysteines, like their prokaryotic relatives. A number of such “Cys-less” receptors also appears in invertebrate metazoans. Together, those findings draw a new distribution of pLGICs in eukaryotes. A broader distribution of prokaryotic channels also emerges, including a major new archaeal taxon, Thaumarchaeota. More generally, pLGICs now appear nearly ubiquitous in major taxonomic groups except multicellular plants and fungi. However, pLGICs are sparsely present in unicellular taxa, suggesting a high rate of gene loss and a non-essential character, contrasting with their essential role as synaptic receptors of the bilaterian nervous system. Multiple alignments of these highly divergent sequences reveal a small number of conserved residues clustered at the interface between the extracellular and transmembrane domains. Only the “Cys-loop” proline is absolutely conserved, suggesting the more fitting name “Pro loop” for that motif, and “Pro-loop receptors” for the superfamily. The inferred molecular phylogeny shows a Cys-loop and a Cys-less clade in eukaryotes, both containing metazoans and unicellular members. This suggests new hypotheses on the evolutionary history of the superfamily, such as a possible origin of the Cys-loop cysteines in an ancient unicellular eukaryote. Deeper phylogenetic relationships remain uncertain, particularly around the split between bacteria, archaea, and eukaryotes.

## Introduction

Pentameric ligand-gated ion channels (pLGICs) mediate fast synaptic transmission in the nervous system of animals with bilateral symmetry (Bilateria), where they are ubiquitous and known as Cys-loop receptors [1–4]. Each receptor is a fivefold symmetric or pseudosymmetric

transmembrane assembly of protein subunits surrounding a central pore that is selective of either cations or anions.

The idea that pLGICs are ancient enough to predate eukaryotes, and that they may have prokaryotic relatives was formulated in 1990 by Cockcroft et al. [5]. No pLGIC was known outside metazoan Cys-loop receptors until the discovery of their prokaryotic homologues by Tasneem et al. [6]. Since prokaryotic receptors lack the eponymous cysteine residues, the Cys-loop family was then superseded by the superfamily that became known as pLGICs. Two prokaryotic pLGICs have been cloned and characterized functionally [7, 8] and were the choice models for high-resolution structural studies [9–13] that paved the way for the recent successes with animal Cys-loop receptors [14–16].

The currently documented taxonomic distribution of pLGICs is that established by Tasneem et al. [6], that is, Cys-loop receptors are ubiquitous in Bilateria and other pLGICs are found sporadically in several bacterial taxons and one archaeal genus. Subsequent work by the same group [17] refined the methods but did not fundamentally alter that picture, while Corringer et al. [18] reported that cursory searches of genome databases revealed more bacterial genes likely to belong to the superfamily. Tasneem et al. interpreted the taxonomic distribution as indicating a complex evolutionary history involving multiple lateral transfers and frequent gene loss. The case of non-bilaterian metazoans has received relatively little attention, presumably due to the scarcity of genome data, although pLGICs have been documented in the cnidarians *Hydra* [19] and *Nematostella* [20, 21]. While the discovery of prokaryotic homologues has provided a fascinating glimpse on the evolutionary origin of animal pLGICs, many questions remain open, among which the ancestry of animal Cys-loop receptors, the reasons for the sparse yet broad distribution of pLGICs in prokaryotes, and the biological roles of those microbial proteins, none of which has been studied *in vivo*.

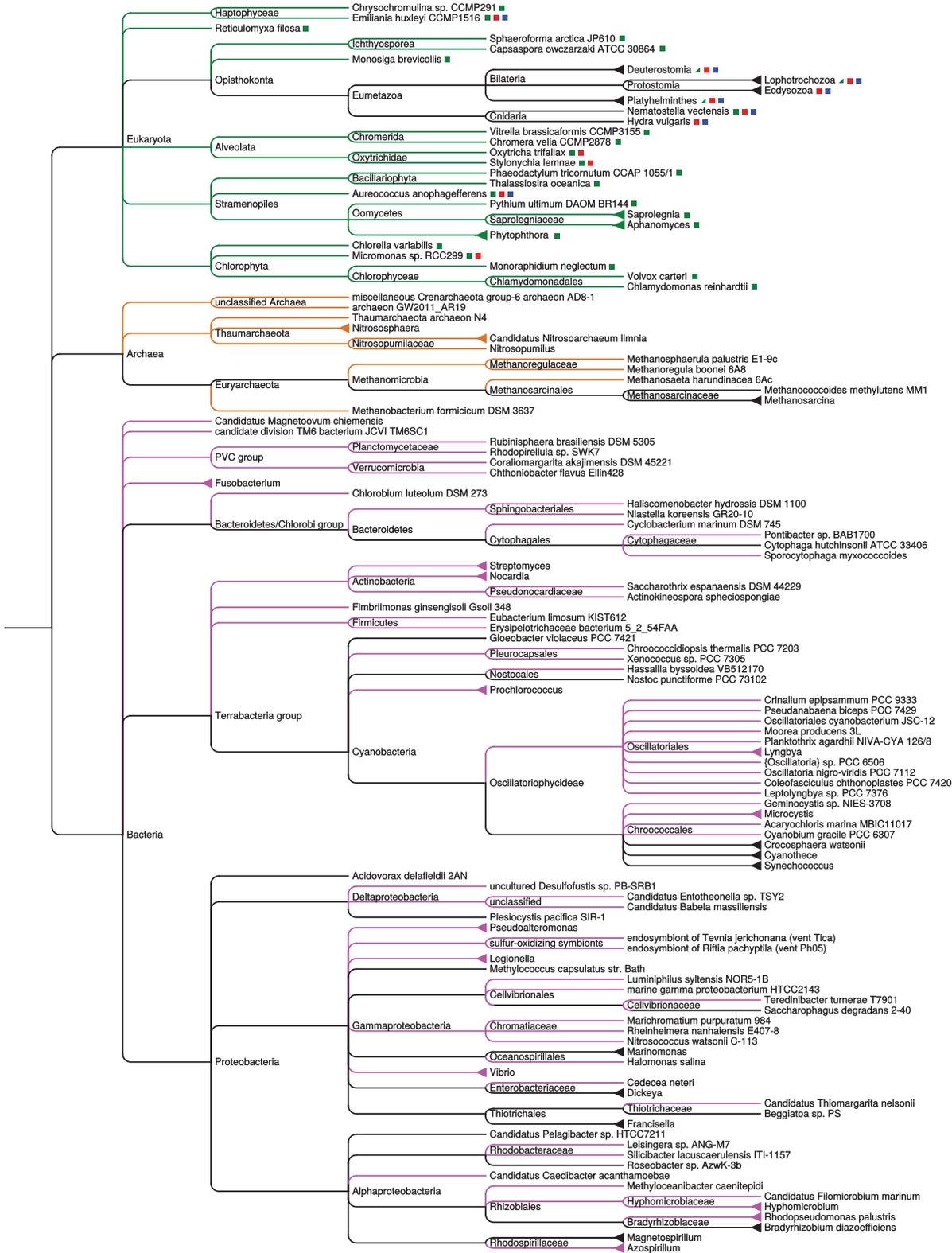
Here we exploit the ever-growing body of genomic data to expand our knowledge of pLGICs throughout the tree of life, focusing on taxonomic groups where they are less well characterized. We extend the inventory of members of the superfamily across all kingdoms of life by performing remote homology searches in protein sequence databases. As metazoan members are better known, we focus on unicellular organisms, and find previously unreported pLGICs in unicellular eukaryotes (called “protists” below for brevity), as well as a broader distribution of prokaryotic channels including a major new archaeal taxon. We also detect and investigate metazoan pLGICs lacking the eponymous Cys-loop cysteines. We construct a multiple sequence alignment of a broadly distributed set of pLGIC sequences, and use it to derive a maximum-likelihood phylogenetic tree that suggests new hypotheses on the evolutionary history of this ancient superfamily. Finally, we discuss the questions that remain unsolved, particularly in light of the difficulty of inferring ancient evolutionary relations based on limited phylogenetic signal.

## Results

### Taxonomic distribution

We find predicted pLGICs in new prokaryotic organisms including several Archaea, and perhaps more surprisingly, in a number of unicellular eukaryotes. Compared with the work of Tasneem et al. [6], we find pLGICs in 61 new bacterial genera, 10 new archaeal genera, and 22 protist genera. This new taxonomic distribution of pLGICs is illustrated in Fig 1, wherein each pLGIC-possessing taxon is placed within a tree of life and colored branches indicate new taxons.

The main novel finding is the existence of pLGIC genes in a wide array of protists. Those include Stramenopiles (a large group including marine organisms such as diatoms, and



**Fig 1. Phylogenetic tree of organisms whose genomes contain predicted Pro-loop receptors.** Colored branches represent taxa that were not discussed previously in the pLGIc literature (green: unicellular eukaryotes, orange: archaea, magenta: bacteria). Colored squares next to eukaryotic taxa indicate the types of pLGIcs present (green: Cys-less, blue: cationic-type Cys-loop, red: anionic-type Cys-loop); the half green squares next to metazoan taxa indicate the presence of Cys-less pLGIcs in some species. The tree is extracted from NCBI Taxonomy.

doi:10.1371/journal.pone.0151934.g001

Oomycetes), Chlorophyta (green algae such as *Chlamydomonas* and *Chlorella*), Opisthokonta (the group that also includes animals and Fungi), and Alveolata. Noteworthy, we detect no pLGICs in multicellular plants (relatives of Chlorophyta) or Fungi (Opisthokonta). Thus, metazoans remain the only multicellular organisms known to possess pLGICs. Within Opisthokonta, pLGIC genes are found in the Choanoflagellate *Monosiga brevicollis*, one of the closest known relatives of metazoans, [22] and *Capsaspora owczarzaki*, in the slightly more distant taxon Filasterea [23]. In contrast, the one other complete genome from a Choanoflagellate, *Salpingoeca rosetta*[24], yields no hit.

The new bacterial pLGIC genes belong to species in many taxons, and diverse ecological niches: marine, soil, plant pathogens, and a few human pathogens. Bacterial species associated with humans include *Erysipelotrichaceae* bacterium, which was isolated from the gut of patients with Crohn's disease; human pathogens *Francisella tularensis*, *Fusobacterium varium*, and *Legionella drancourtii*.

Many new archaeal genomes have been sequenced since the work of Tasneem et al. Whereas the two pLGICs of *Methanosarcina* pinpointed by that work were taxonomically isolated and offered little basis for evolutionary interpretation, this larger set supports a more substantial analysis of molecular phylogeny. We find pLGIC genes in new archaeal species in Euryarchaeota and in a number of Thaumarchaeota [25]. Taxonomic coverage of those groups is very difficult to measure, as most hits come from datasets that are not annotated as whole genomes. Notably, we find no hit in the phylum Crenarchaeota, despite the presence of dozens of genomes (complete or partial) in databases. Finally, as this work was being completed, the taxon Lokiarchaeota was described based on metagenomic data, and presented as a link between archaea and eukaryotes [26]. Extending our search to the set of predicted proteins for the candidate species *Archaeon loki* did not yield significant hits.

Our representative set of metazoan genomes includes two non-bilaterian species: *Hydra vulgaris* and *Nematostella vectensis* (both Cnidaria). Both have genes in the anionic and cationic groups of Cys-loop receptors. *Hydra* has two Cys-loop receptors, automatically annotated as  $\gamma$ -aminobutyric acid (GABA<sub>A</sub>, Uniprot ID T2MFN2) and nicotinic acetylcholine receptor (nAChR, T2MCD8). A taxonomically unrestricted search in UniprotKB yields no other examples of non-bilaterian pLGICs. In particular, there is no hit in the genomes of basal metazoans Placozoa [27], Porifera [28], and Ctenophora [29]. Of these three phyla, only Ctenophora possess a nervous system, which has been proposed to have evolved independently from those of Bilateria and Cnidaria, which share a common origin [29].

pLGICs of both *Hydra*[19] and *Nematostella*[20] have been discussed before. The authors who reported the *Hydra* genome found the expression pattern of the nAChR gene to be compatible with a function in neuromuscular signaling [19]. They find the neuromuscular junction of *Hydra* to possess many, but not all, of the molecular components found in the equivalent bilaterian system. Neither the GABA<sub>A</sub>-like receptor of *Hydra*, nor the pLGICs of *Nematostella* have been characterized experimentally.

## Multiple sequence alignment

The complete multiple sequence alignment is provided as [S1 Dataset](#). It contains 561 protein sequences: 218 metazoan (among which 69 Cys-less), 193 eubacterial, 24 archaeal and 126 from protists. The alignment has 3405 positions, discarding segments at the N and C-termini that do not align with the pLGIC domains. The variable M3M4 cytoplasmic region accounts for 1170 gap-rich positions. The “core” of the alignment includes about 320 well-aligned positions (a typical pLGIC sequence length without the cytoplasmic loop), interspersed with about 1915 poorly aligned and gap-rich positions. Despite the high noise level, significant structural

elements emerge as less gapped blocks, most notably the four transmembrane helices and some beta-strands of the extracellular domain (ECD).

A sub-alignment of 15 sequences is presented in [Fig 2](#). Several regions of the ECD appear poorly conserved, including some hydrophilic regions of the ECD that, in known structures, contribute to the  $\beta$ -sandwich core. This highlights the contrast between a tightly conserved fold and divergent sequences. In contrast, transmembrane segments align readily, constrained by the reduced alphabet of hydrophobic amino-acid residues. Aligning the M4 segment poses a challenge, as it is surrounded in N-term by the variable or absent cytoplasmic domain, and sometimes in C-term by sequence regions that do not belong to the common pLGIC architecture. As a result, predicted M4 helices only appear aligned when using the local alignment-based, iterative algorithms of MAFFT (E-INS-i and L-INS-i). A sequence profile for the balanced dataset is provided as [S1 Fig](#).

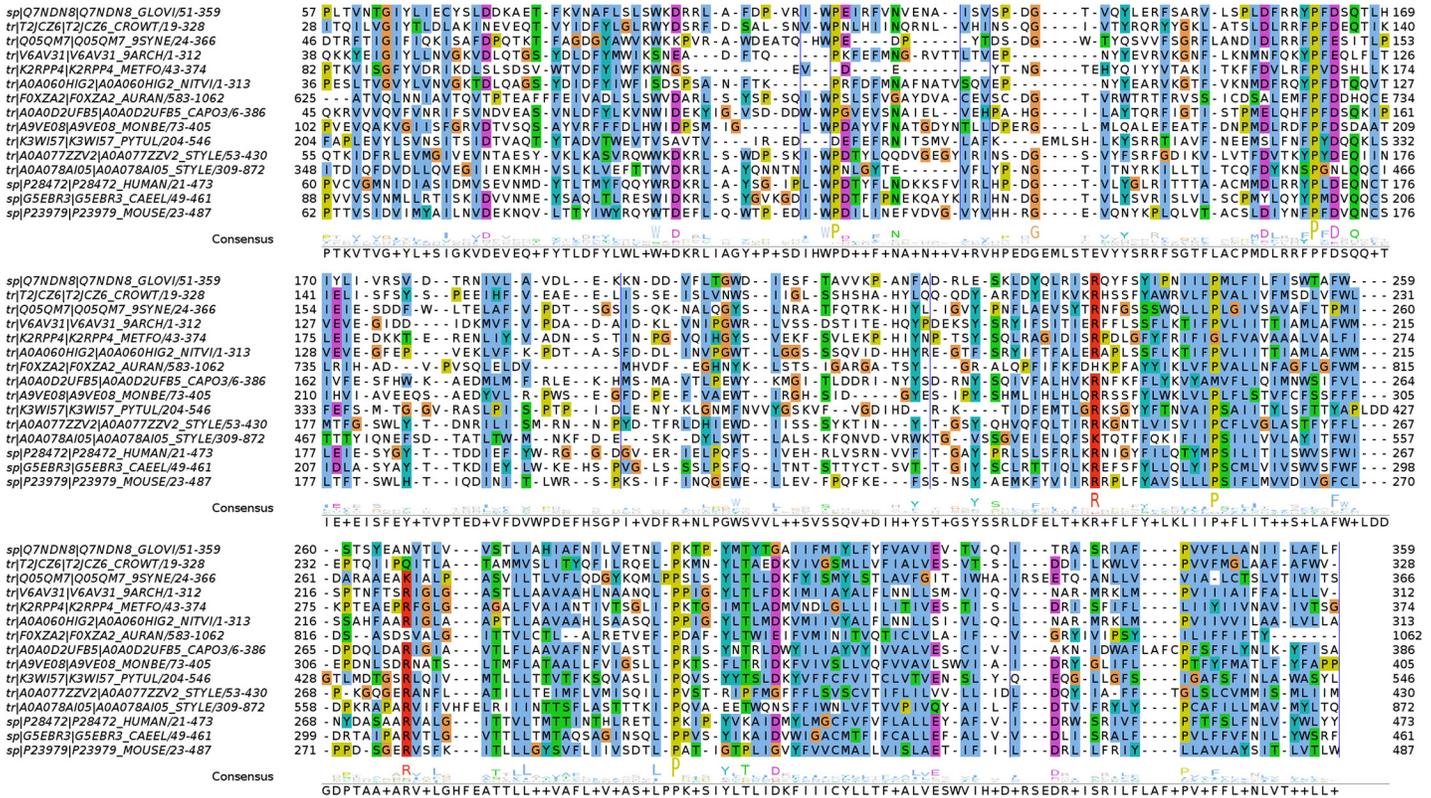
### Conserved motifs and notable sequence features

The most conserved residues across the superfamily are listed in [Table 1](#) and their three-dimensional arrangement is pictured in [Fig 3](#). It is evident from the figure that these residues are clustered at the interface between ECD and TMD, within the coupling pathway between ligand binding and pore opening in known pLGICs. At the level of the superfamily, no conservation linked to more specific function emerges, reflecting the functional diversity of both ligand-binding and ion translocation. Furthermore, well-studied pLGICs including those of prokaryotes have demonstrated a high degeneracy of the sequence to fold to function relationship, with divergent sequences giving rise to a remarkably conserved fold, supporting a common functional pattern of ligand or pH-gated, anion or cation-selective transport.

The highly conserved [F/Y]PxD motif forms the tip of the  $\beta 6$ - $\beta 7$  loop (“Cys loop”), and is the most recognizable signature of the pLGIC superfamily. In known structures, the aspartate residue forms a salt bridge with the equally conserved arginine of the pre-M1 linker. The proline residue is the single most conserved residue in our alignment ([Table 1](#)), and the only absolutely conserved residue throughout the superfamily.

The pair of cysteines flanking the Cys loop and forming a disulfide bridge are conserved in metazoans, and were noted by Tasneem et al. to be absent from bacterial sequences, where they are typically replaced with a polar residue for the first cysteine and with a hydrophobic residue for the second. We find that protist pLGICs fall under two separate categories: most contain a prokaryotic-like, Cys-less loop, but four species possess pLGICs featuring a *bona fide* Cys loop (annotations in [Fig 1](#)). A more precise relationship between those categories is elicited in the phylogenetic analysis described below.

We find Cys-less pLGIC sequences in invertebrate Metazoa from our representative set. To complete this surprising finding, we ran a taxonomically unrestricted search that yielded more Cys-less channels in metazoan species, including one cnidarian and one echinoderm, several nematodes, platyhelminths, annelids, and mollusks (bivalvia and gastropoda). Cys-less Pro-loop receptors are not found in any vertebrate species, and only in two chordates: the cephalochordate *Branchiostoma floridae* (amphioxus or lancelet) and the tunicate *Oikopleura dioica*. One Cys-less metazoan pLGIC was identified in the nematode *Dirofilaria immitis* [30] even before the discovery of prokaryotic pLGICs, yet that finding has not been widely publicized in the community. This channel (Uniprot ID Q70GM3) possesses a unique variant of the Pro loop sequence, with a YPFE motif (with E instead of the more common D) shifted by one residue towards the N-terminus. This trait is shared by four other nematode pLGICs, which form a molecular clade (see phylogeny below).



**Fig 2. Subset of large multiple sequence alignment.** Contains 11 novel pLGIC sequences from bacteria, archaea, and eukaryotes. GLIC, nematode GluCl, mouse serotonin, and human GABA<sub>A</sub> β3 receptors are included for comparison. Residues are colored by type according to the ClustalX scheme. Unconserved residues are hidden and indicated by blue, vertical lines. Species names are abbreviated in the figure. Bacteria: *Gloeobacter violaceus*, *Crocospira watsonii*, *Synechococcus sp.*, archaea: *Thaumarchaeota archaeon*, *Methanobacterium formicicum*, *Nitrososphaera viennensis*, eukaryotes: *Capsaspora owczarzewski*, *Monosiga brevicollis*, *Pythium ultimum*, *Stylonychia lemnae*.

doi:10.1371/journal.pone.0151934.g002

**Table 1. Summary of most conserved amino-acid residues throughout pLGICs in all taxons.**

position	residue	frequency (%)	motif	substitutions	GLIC	GABA <sub>A</sub> β3	5HT3
β6–β7	F	71 <sup>a</sup>	F/Y/Px <sup>b</sup> D	Y 25%	119	143	142
β6–β7	P	100	F/Y/PxD	-	120	144	143
β6–β7	D	93	F/Y/PxD <sup>c</sup>	E 3%	122	146	145
β9	W	71 <sup>a</sup>	ECD core <sup>d</sup>	F 16%, Y 11%	160	186	187
pre-M1	R	94	salt bridge <sup>c</sup>	-	192	216	218
M1 14 <sup>e</sup>	P	77 <sup>a</sup>	M1 kink	G 5%	204	228	230
M2M3	P	88	M2M3 loop	G 3%	247	273	274

The table lists residues with conservation above 80% in our balanced set of 561 pLGIC sequences throughout the tree of life. Substitutions above 2% frequency with similar residues are indicated. To prevent bias due to fragmentary sequences, this conservation measure excludes gaps. All positions listed have a gap content below 2%. Residue numbering is given for three pLGICs crystallographic structures: *Gloeobacter* GLIC, human GABA<sub>A</sub> β3 and mouse serotonin receptors.

<sup>a</sup> Included because similar residues are more than 80% conserved.

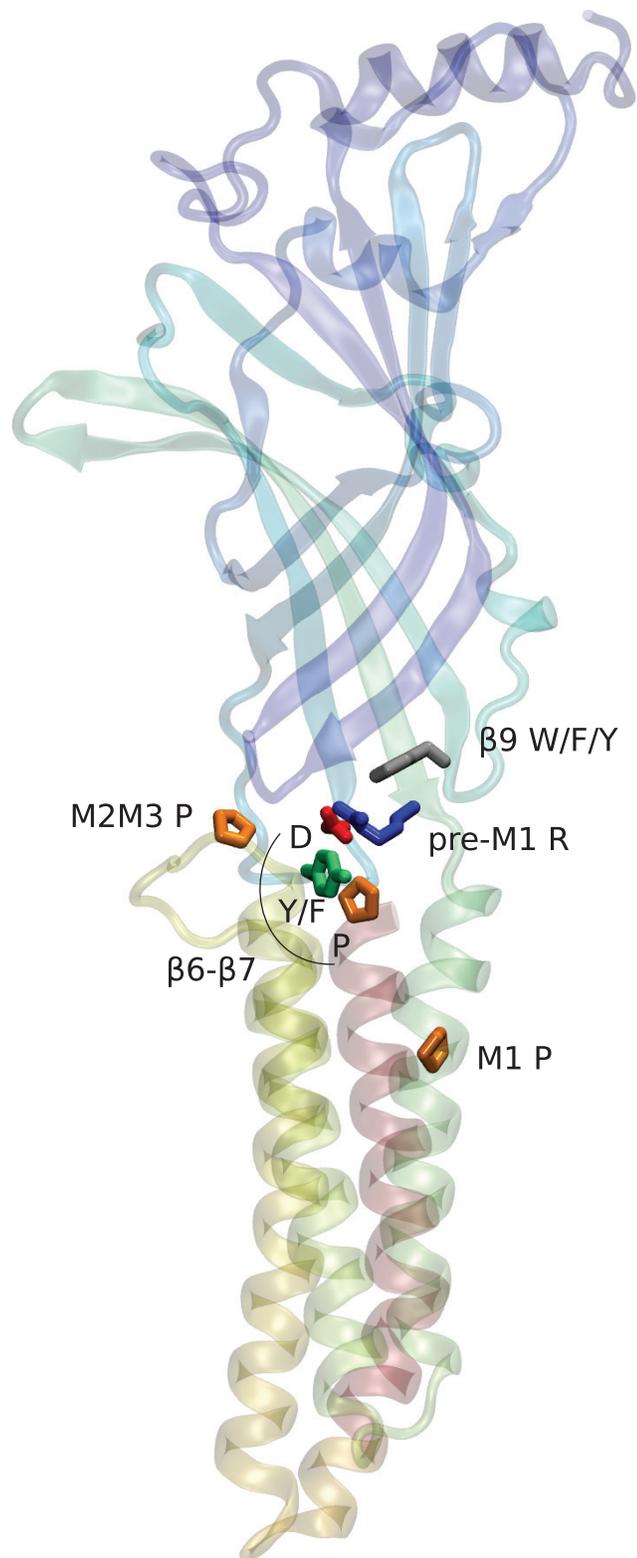
<sup>b</sup> Packed between aromatic M1 2', β6–β7 proline, and bulky apolar M2 22'.

<sup>c</sup> These two residues form a salt bridge.

<sup>d</sup> This aromatic sidechain is buried in the hydrophobic core of the β-sandwich fold of the ECD.

<sup>e</sup> See our proposed prime notation for M1 and M3 in the TM domain section below.

doi:10.1371/journal.pone.0151934.t001



**Fig 3. Location of the most conserved residues within the structure of a pLGIC subunit.** One subunit of the homomeric GABA<sub>A</sub> β3 receptor shown as cartoon, colored from blue to red along the sequence. Conserved residues listed in Table 1 are shown as sticks and colored by residue type (orange: Pro, grey: Phe, green: Tyr, red: Asp, blue: Arg).

doi:10.1371/journal.pone.0151934.g003

The ECD displays a number of partially conserved aromatic positions. Based on known structures, some belong to the hydrophobic core of  $\beta$ -sandwich structure, while others are expected to participate in the ligand-binding site where they might form cation- $\pi$  interactions. The latter are less conserved than the former, consistent with the fact that the ligand-binding motifs are more variable than the core architecture of the ECD, although the known set of binding site structures and corresponding ligands may only represent a fraction of the functional repertoire of pLGICs in unicellulars.

The most conserved position within TM helices is a proline in M1 that forms a kink in known structures. This kink has been shown to undergo a structural transition during the gating and desensitization processes in nicotinic receptors. [31]

Charged residues in pre-M2 position have long been documented to form a charge selectivity filter [32, 33]. In our dataset, the cationic residue at M2 0' is found with 79% conservation (R and K combined). The previous two residues that form the classic metazoan motifs are more variable in prokaryotic sequences. M2 has one other relatively conserved position: Leu 9' (75%), which can be substituted with a Phe (13%).

### Transmembrane domain: prime numbering for M1 and M3

Homologous residues in transmembrane helix M2 are conventionally compared across pLGICs using a standard prime numbering. The lack of a similar convention for M1 and M3 is impractical, as evidenced by published work on a glycine receptor [35] referring to an M1 residue as -26', in M2 notation. Most prokaryotic pLGICs as well as some mammalian cases (5HT3) show indels within the M1M2 linker, making that notation non-constant across receptor families. The prime numbering used for M3 by Auerbach and coworkers [34] is defined for nAChR subunits only.

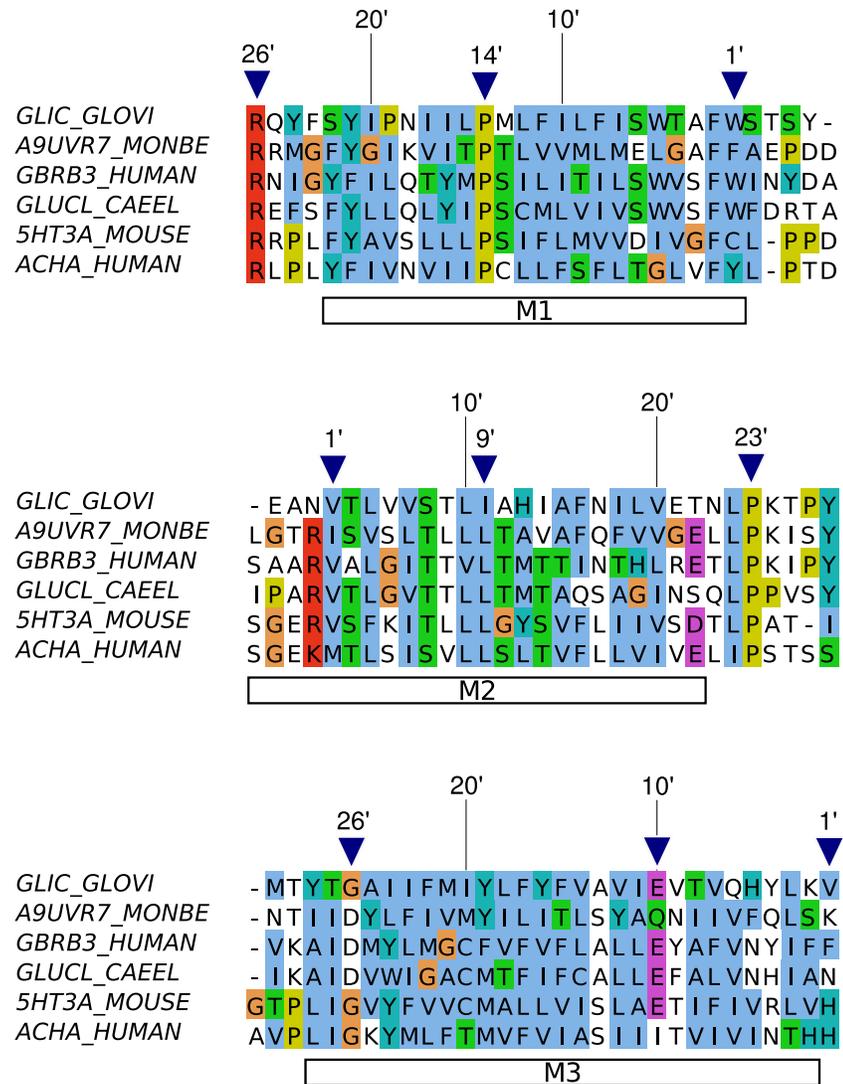
Despite significant variability in transmembrane helices M1 and M3, we find that the alignments are robust enough to propose universal prime numberings for those two helices as well (Fig 4). We follow the convention of numbering residues from cytoplasmic to extracellular side.

In M1, we give the partly conserved C-term aromatic residue (GLIC W217) number 1'. Conserved residues include F 2', P 14', and extending beyond the helix, R 26' (pre-M1 linker).

The M2M3 loop is well-known to be critical, and would benefit from a universal residue notation. Cationic channels typically display an insertion at the end of that segment, just before M3 0'. M2 numbering may be extended until position 25' unambiguously, after which indels appear in some sequences.

For helix M3, we follow the convention started by Cadugan and Auerbach [34] for nicotinic receptor  $\alpha$  subunits. As helix M3 is relatively divergent across families, we anchor the numbering by referring to motifs present in individual groups of pLGICs. The initial T/P/K position (GLIC T253, Torpedo nAChR $\alpha$  P272, GABA $_A$   $\beta$ 3 K279) takes number 29'. Position 24' is aromatic-rich (65% in balanced set, 89% in metazoan sequences only). Typical metazoan anionic channels have: K/T 29', D 26', C 20', F 15', E 10'. Typical metazoan cationic channels have P 29', G 26', Y/F 24', polar 4'.

In eukaryotes, most acetylcholine receptors have a Cys at M1 10', while most anionic channels have one at M3 10'. In GABA $_A$   $\alpha$  and  $\gamma$  subunits, both are present, and predicted to be ideally positioned to form a potential disulfide link between M1 and M3. [36] The recent crystal structure of a GABA $_A$   $\beta$ 3 homopentamer did confirm the structural predictions of that study, yet unfortunately did not provide a test of the hypothetical disulfide bridge, as  $\beta$ 3 subunits lack the M1 cysteine.



**Fig 4. Prime numbering scheme for transmembrane helices M1 to M3.** A sequence alignment for a set of pLGICs is shown annotated with a prime numbering convention in each helix, starting on the cytoplasmic side. Sequences are labeled with their abbreviated gene and species names, or Uniprot identifier in the case of the predicted pLGIC from the protozoan *Monosiga brevicollis*. The figure shows the existing convention for M2, generalizes that proposed for M3 in nAChR  $\alpha$  subunits [34], and proposes a new convention for M1. Triangles indicate 1' positions as well as conserved residues that may help anchoring other sequences.

doi:10.1371/journal.pone.0151934.g004

The sequences for helix M4, although they broadly align together, show too little conservation for a global numbering to be reliable.

### Phylogeny of pLGICs

The maximum-likelihood prediction yields an unrooted tree. One approach to root it is to reconcile the molecular phylogeny and the phylogeny of species with the program Notung, minimizing a penalty reflecting the number of gene losses and duplications requested by each choice of root. The results from that analysis are ambiguous, as many edges receive the best root score. However, those likely roots lie within the same broad region of the tree, comprised of the deep connections between prokaryotic branches. Although the precise position of the

root remains uncertain, this does not affect our conclusions below. The resulting unrooted tree is available online for download as well as flexible, searchable visualization at: <http://itol.embl.de/shared/jhenin>.

Eukaryotic pLGICs are predicted as monophyletic, although with low statistical support. That is mostly reflected in one clade of bacteria whose position in the tree is unstable, and which appears either close to or within the eukaryotic group depending on small variations in the data, in particular alignment filtering. We opted for a loose filtering scheme following the conclusions of Tan et al. suggesting that stricter alignments may reduce the accuracy of inferred phylogenies.[37] The same lack of statistical support was apparent in the tree published by Tasneem et al., only made more striking by a considerably smaller dataset wherein most relationships were *deep* relationships and therefore difficult to ascertain.

Eukaryotic proteins form two notable clades: Cys-less channels, including a group of Cys-less channels of invertebrate metazoans; and a Cys-loop clade. The Cys-less clade includes genes from protists and one group of Cys-less channels from metazoans, which is dominated by Lophotrochozoa but also includes a group of sequences from Platyhelminths, and a sister group of 5 sequences from Deuterostomia (Branchiostoma and Strongylocentrotus), together with an isolated sequence from Nematostella.

The Cys-loop group includes two subclades, one of which contains the anionic channels of metazoans, as well as protist Cys-loop receptors with a sequence signature for anion selectivity in the pore-lining M2 segment, the second contains cationic channels of metazoans and protist sequences with the equivalent signature for cation selectivity. Thus, the grouping of multicellular and unicellular genes within the Cys-loop clades is congruent with a very simple, 3-amino-acid sequence signature in the transmembrane domain (with the exception of one *Nemstostella* sequence, colored black in Fig 5, which has neither the typical cationic, nor the typical anionic signature). Henceforth we will refer to those two groups as anionic-type and cationic-type Cys-loop receptors, respectively, although the ionic selectivity of their unicellular members is predicted from a simple sequence signature and phylogenetic grouping with bilaterian receptors, rather than characterized experimentally.

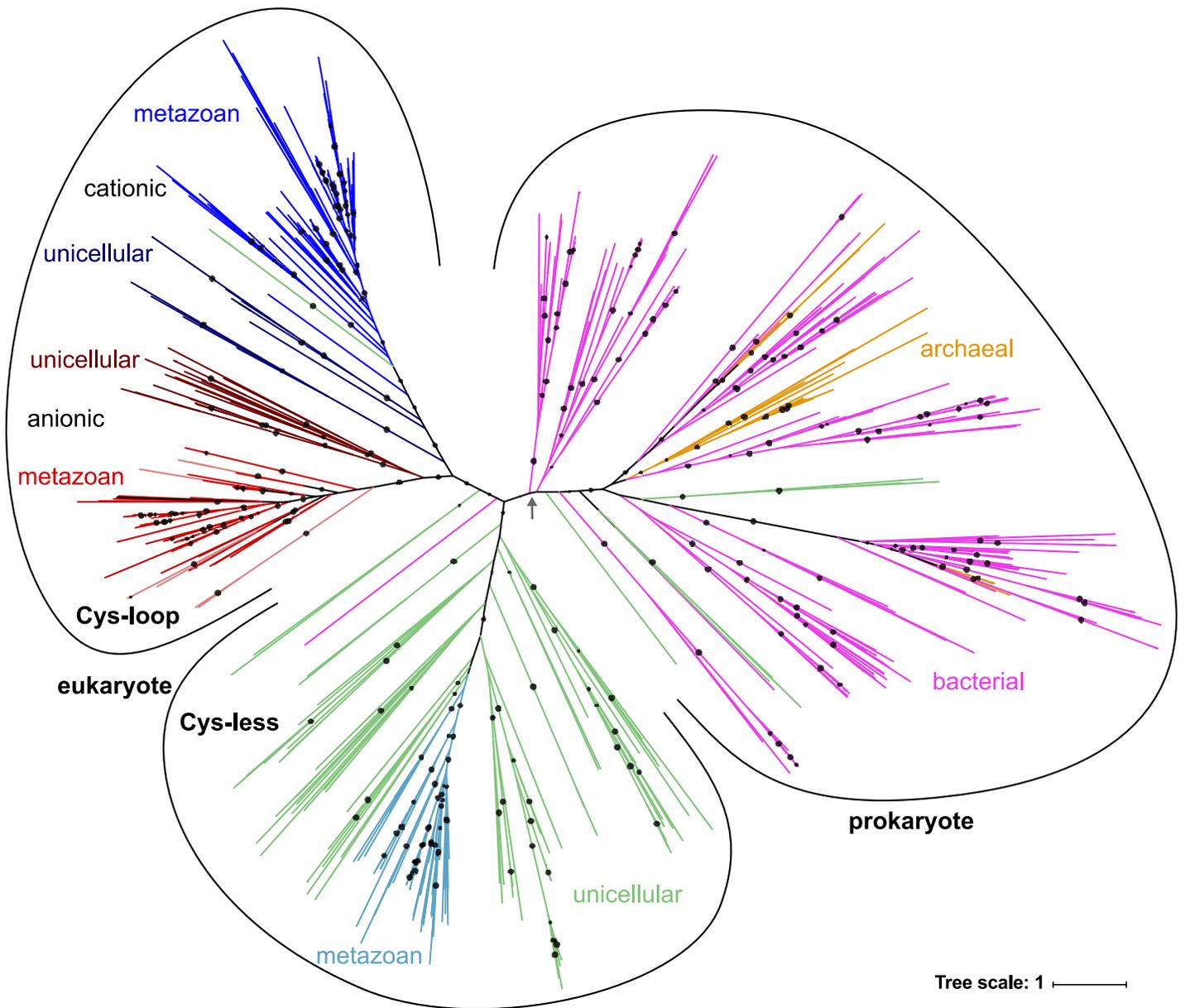
A few Cys-less sequences are also found within the anionic Cys-loop clade: we refer to them as “secondary” Cys-less pLGICs. They mostly belong to two groups of nematode genes, one of which includes the unusual Pro loop sequence of *Dirofilaria immitis*[30]. Isolated cases of secondary Cys-less sequences are two sequences in the spider *Stegodyphus mimosarum* and one in *Oikopleura dioica*. Note that a search for Cys-less sequences in all Metazoan genomes yielded no Cys-less pLGICs in other tunicate genomes, including that of *Ciona intestinalis*. A group of two Cys-less sequences from ciliate protozoa of the family Oxytrichidae (*Stylonychia lemnae* and *Oxytricha trifallax*) appear within the cationic Cys-loop clade.

Within prokaryotes, a clade of sequences belonging to Thaumarchaeota emerges, while other archaeal sequences are more interspersed with their bacterial relatives. Seven Cys-less sequences from protists appear within the prokaryotic branches (pale green in Fig 5), although always with SH support below 90%, indicating uncertain placement. Conversely, one isolated sequence from *Synechococcus* groups with eukaryotic sequences, and the bipartition separating it from other prokaryotes and Cys-less eukaryotes has a low SH support value of 68%.

## Discussion

### Pro-loop receptors

Confirming previous results [6, 17], we find that the proline residue forming the tip of the  $\beta 6$ – $\beta 7$  loop is the single most conserved position in the superfamily. This degree of conservation implies that it is subject to high evolutionary pressure: it is presumably essential in maintaining



**Fig 5. Inferred phylogenetic tree of the Pro-loop superfamily.** Branch colors represent a combination of taxonomy and sequence features: magenta: eubacteria, orange: archaea, pale green: Cys-less pLGICs of protists, cyan: Cys-less pLGICs of metazoans, blue: cationic-type Cys-loop of metazoans, dark blue: cationic-type Cys-loop of protists, red: anionic-type Cys-loop of metazoans, dark red: anionic-type Cys-loop of protists, pale red: anionic-type Cys-less of metazoans. Circles indicate SH support above 90%. The gray arrow indicates two bacterial branches of poorly-defined position, which sometimes group with eukaryotic sequences.

doi:10.1371/journal.pone.0151934.g005

a conformation of the loop that is functionally critical. Based on a high-resolution structure of GLIC and a reinterpretation of crystallographic data on other pLGICs, this proline has been argued to be in a *cis* configuration in all known structures [12]. This suggests that the superfamily's absolutely conserved, defining feature is not just a proline, but specifically a *cis*-proline residue, in line with a precise requirement on the local conformation of the loop. Rendon et al. [17] already noted this conservation and suggested renaming the Cys loop p-loop, however, the phrase has not gained adoption since, perhaps because of possible confusion with similarly

named motifs. More closely mimicking the name “Cys loop”, however, we propose that “Pro loop” would be an accurate and unambiguous name for both the loop and for the superfamily.

The Pro loop as a whole is a highly conserved motif at the core of the conserved cluster within the ECD/TMD interface (Fig 3 and Table 1). This interface features two conserved prolines, two aromatic residues, and a pair forming a salt bridge between the Pro loop and the short  $\beta$ 10–M1 linker. An explanation for the high selection pressure on these residues is that they maintain the subtle mechanical contact between the two domains in a way compatible with signal transduction across the interface. The fact that this conservation transcends receptors families with different functions is consistent with the persistence of signal transduction in chimeric receptors that mix those families [38–41].

The present results confirm that the cystine bridge closing the  $\beta$ 6– $\beta$ 7 (“Cys”) loop is a secondary feature of pLGICs in some eukaryotes (including most metazoa and all vertebrates), rather than a primitive characteristic of the superfamily. The few reports of experiments investigating the role of these residues find them to be essential. Based on mutagenesis on the  $\alpha$  subunit of *Torpedo* nAChR, Mishina et al. [42] hypothesized as early as 1985 that the pair of cysteine residues is “essential for maintaining the proper conformation of the extracellular region of the AChR molecule”, as it abolished binding of bungarotoxin to the surface of injected *Xenopus* oocytes. A mutagenesis study on glycine receptor subunit  $\alpha$ 1 indicated that mutation of one cysteine (198) to serine abolishes receptor expression to the cell surface, while mutation of the second cysteine (209) allows some expression to the cell surface, but still abolishes glycine-induced whole cell currents and strychnine binding [43]. In contrast, cysteine cross-linking experiments on glycine [44], GABA<sub>A</sub> [45, 46], and nicotinic [47] receptors found that a reducing agent had no significant functional effect on wild-type receptors, implying either that the Cys-loop disulfide is too stable to be reduced under the conditions of those studies, or that such a reduction has little impact on the fully folded receptors. In support of the second hypothesis, the fold of the ECD in GLIC and ELIC is essentially identical to that of Cys-loop receptors and stable in the absence of a disulfide. The hypothesis that the disulfide is necessary for native folding of Cys-loop receptors is neither validated nor contradicted by the existence of GLIC and ELIC, as bacteria have a different protein expression machinery; however, GLIC can be readily expressed in eukaryotic cells such as *Xenopus* oocytes [7]. The present data reinforces that notion, with the finding of native Cys-less pLGICs in eukaryotes (many protists and a few invertebrates). Although those are predicted proteins without biochemical or biophysical characterization, at least some of these sequences can be expected to yield functional proteins. If indeed the pair of cysteines is essential for folding Cys-loop receptors, then Cys-less pLGICs must have an alternate mechanism that forms and stabilizes the very same native fold. Thus, the appearance and subsequent conservation of the bridging cysteines remains partly unexplained. We postulate that functional Cys-less mutants of Cys-loop receptors could be designed based on Cys-less sequences, perhaps including additional hydrophobic residues stabilizing the core of the ECD.

## Congruency and incongruencies of molecular and species phylogenies

Some taxonomic clades are reflected in the molecular phylogeny, although often with exceptions. Those congruent groups include eukaryotes, and within those, metazoa, as three separate sub-clades of larger groups of eukaryotic pLGICs: cationic-type Cys-loop, anionic-type Cys-loop, and Cys-less.

Within unicellular eukaryotes, the distribution of different clades of pLGICs shows little correlation with cladistic groupings (Fig 1). We find no Cys-loop genes in Oomycetes, compatible

with an early gene loss in that group. Other groups have too few representatives in our dataset to draw any conclusion.

Although Opisthokonta are the closest extant relatives of Metazoa, we find no Cys-loop gene in unicellular members of that taxon, but only Cys-less pLGICs; conversely, Cys-loop receptors are found in a wide range of eukaryotic taxons, pointing to either an early origin followed by frequent loss, or very widespread lateral transfers, or both. Some sequences of *Mono-siga* do appear at the base of the Cys-less group of eukaryotic pLGICs, placing them close to the anionic group of Metazoan Cys-loop receptors. We note that the phylogenetic position of 2 out of 4 such sequences has poor statistical support.

The case of *Emiliania* has been pointed out as unusual: it is a “pan genome” [48] assembled from many strains, which explains the extraordinary number of pLGIC genes it contains (34 after excluding fragments), compared with any other genome outside Metazoa. Those sequences exhibit limited taxonomic congruency: although many of them form clades, others group with those of different protists. This may reflect either ancient divergence of those groups, or the uncertainty of the predicted deep relationships, highlighted by the large dataset from *Emiliania*.

One group of eukaryotic-like bacterial sequences including the GLIC and ELIC channels (gray arrow in Fig 5) shows an unstable position, appearing within the eukaryotic branch in some phylogenies inferred from intermediate data (not shown). Accordingly, its grouping in the tree presented here has poor statistical support.

Of note, the ten pLGIC sequences from Thaumarchaeota form a statistically robust clade, suggestive of a single gene acquired by an ancestor of that group. Taxonomic coverage of that group is, however, partial, indicating gene loss. These Thaumarchaeota sequences are part of a clade of 16 sequences from Archaea. Eight other archaeal genes are split into several groups among bacterial relatives, possibly reflecting acquisition through several independent lateral gene transfers.

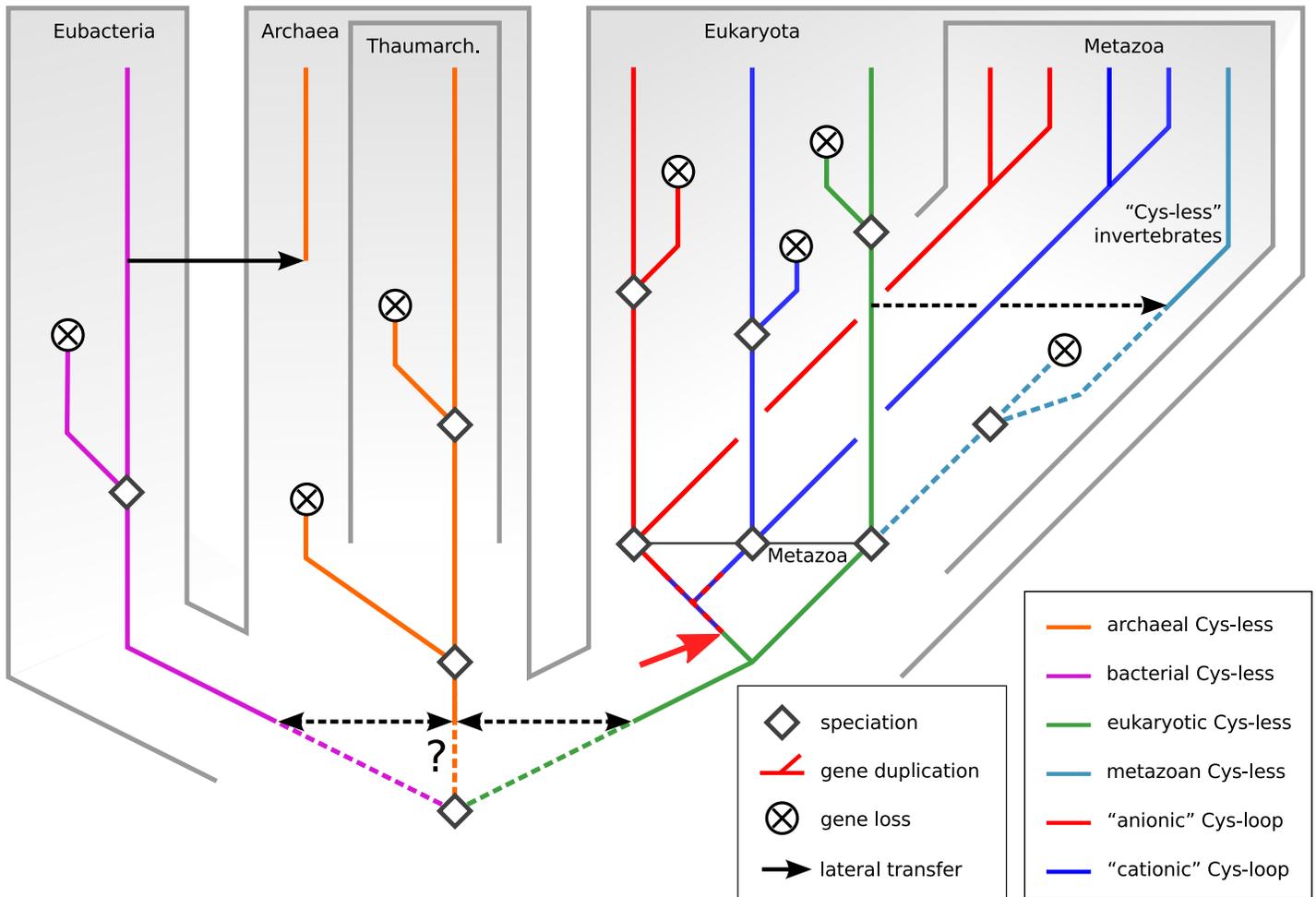
## Tentative evolutionary history

Fig 6 presents a speculative, schematic evolutionary history of the pLGIC superfamily based on the molecular phylogeny of Fig 5 and including alternate hypotheses for some unresolved questions.

The scattered presence of pLGICs in Bacteria across several major phyla is intriguing. It points to either massive gene loss happening in most clades, or a number of lateral transfer events. It remains unclear in what taxon pLGICs originated, for three reasons: their sparse distribution in extant species, the uncertainty on deep phylogenetic relationships, and the resulting difficulty of precisely rooting the complete gene tree.

The origin of archaeal pLGICs is equally unclear, but again, requires a combination of lateral transfer and gene loss events given the sparse distribution and taxonomically incongruent phylogeny. The main novel information is the presence of pLGICs in 6 species of the relatively recent group Thaumarchaeota [25], whose pLGICs form a molecular clade.

The result that eukaryotic pLGICs are monophyletic is compatible with two simple hypotheses for their origin: either a single speciation event, namely the appearance of eukaryotes, or a single lateral transfer event to an ancestral eukaryote. Whichever of these two events occurred was followed by early duplication and differentiation into the extant pLGIC families, as well as frequent loss of some of those differentiated types of receptors (all of them, in some unicellular species). The high rate of loss in unicellulars might be explained by contraction of unicellular genomes [49]. The presence of both the cationic and anionic Cys-loop clades in unicellular



**Fig 6. Schematic, putative molecular phylogeny of the Pro-loop superfamily.** Each taxonomic category represents many species. Dashed lines indicate alternate hypotheses. Speciation events followed by short branches ending with gene loss depict the many unicellular lineages that have likely lost their Pro-loop receptors. In the case of metazoans, the cyan line indicates the clade of Cys-less Pro-loop receptors that survives in a few extant invertebrates. The red arrow indicates appearance of the Cys loop, presumably in an ancient unicellular eukaryote. Dashed red-blue lines describe the ancestral Cys-loop receptor, which may have been anionic or cationic-type.

doi:10.1371/journal.pone.0151934.g006

eukaryotes pushes back the appearance of these clades to an early stage of eukaryotic evolution, before the appearance of metazoans.

This finding solves one question raised by Tasneem et al. a decade ago [6]: did eukaryotes acquire pLGICs “well before the emergence of the animal lineage”, followed by frequent loss in other eukaryotes, or was there “a single precursor for all the animal sequences”, from which “the massive radiation of the Cys-loop receptors occurred only after the animals branched off”? Those authors accurately added that this could only be answered after more eukaryotic sequences become available. Although the present data, is compatible with a single ancestor gene for all eukaryotic pLGICs, it indicates that some of the diversity of Cys-loop receptors—namely, the anionic and cationic branches—predates the appearance of metazoans.

The recruitment of pLGICs into synapses in Cnidaria and Bilateria was linked with an evolutionary pattern of successive gene loss and gene expansion events [50]. In those metazoans, a dependence on rapid synaptic transmission may have enshrined the corresponding pLGICs and prevented the pattern of drastic gene loss that prevails in unicellulars. Still, it appears that

some Bilateria did lose individual families of pLGICs, such as GABA(A) receptors in the case of *Schistosoma mansoni*[51].

Among the protists identified, the Choanoflagellate *Monosiga brevicollis*, one of the closest known relatives of metazoans, prompts new hypotheses about the origin of metazoan pLGICs. One such hypothesis is that extant metazoan pLGICs have direct ancestors in protists of the clade Holozoa. Unfortunately, the phylogenetic positions of *Monosiga* pLGICs have poor statistical support, not allowing for clear conclusions. We should note that discussing the evolution of individual genes in protists is all the more difficult that the deep phylogeny of eukaryotes is subject of active research, and their taxonomy somewhat unstable [52].

We find that “true” Cys-loop receptors form a monophyletic group within pLGICs, consistent with a character acquired once and conserved since. There are few cases indicating secondary loss of the cysteines, and conversely they are not found outside the Cys-loop clade. We cannot say whether all such genes were inherited linearly from a very ancient common ancestor, or appeared more recently, evolved in a more limited lineage, and were acquired by distant eukaryotic species through lateral transfer. All three pLGIC groups (Cys-less, anionic-type and cationic-type) are present in *Emiliana* (Haptophyte), *Aureococcus* (Heterokont), Ciliates, and Opisthokonta, including metazoans (Fig 1). Based on the present data, it is not possible to place the origin of Cys-loop receptors with certainty any later in the history of eukaryotes than the common ancestor of all those groups, which may well be the common ancestor of all extant eukaryotes.

Cys-less Pro-loop receptors are unexpected in metazoans. Most of them belong to Lophotrochozoa and form a statistically well-supported clade together with Cys-less protist channels, and a few are “secondary” Cys-less pLGICs within the Cys-loop clade. Two possible histories can be proposed for the Cys-less eukaryotic clade: 1) They derive from an ancestral receptor that was lost in other metazoans, possibly because it did not undergo the same selection pressure as Cys-loop receptors, for example due to a function outside the nervous system. 2) These metazoans with Cys-less pLGICs acquired the Cys-less form through a lateral transfer event from a unicellular eukaryote. The biology of such channels is not documented, as only one member is mentioned in the literature [30] with a report that an RNAi experiment yielded “no obvious phenotype”. The channel in question belongs to the nematode *Diriofilaria immitis* and is predicted to belong to the anionic Cys-loop clade, which implies that it probably results from a secondary loss of the cysteine residues. Based on that result, together with their evolutionary grouping with protist proteins, we predict that Cys-less invertebrate pLGICs could have a non-synaptic expression pattern.

Incongruencies between the present molecular phylogeny and that of species suggest several lateral transfers in unicellular organisms. This notion is compatible with the ecology of many of the pLGIC-containing unicellular species, many of which share biotopes, or are even symbionts or parasites of larger organisms. Microbial organisms with pLGICs include many marine species, among which some of the most abundant species of photosynthetic plankton, both prokaryotic (*Prochlorococcus* and *Pelagibacter*) and eukaryotic (*Emiliana*). *Teredinibacter turnerae* is an endosymbiont of mollusks [53]; *Chlorella* is an endosymbiont of freshwater unicellular organisms such as *Paramecium bursaria*. In general green algae of the *Trebouxiophyceae* class encompass endosymbionts of many organisms including mollusks. One species of *Synechococcus* was recently shown to accumulate in cells of the oyster *Crassostrea gigas* in a manner compatible with endosymbiosis [54]. Cases of photosymbiosis (symbiosis with a photosynthetic guest) exist between Platyhelminth and green algae [55]. Such symbiosis relationships with metazoans raise the possibility that Cys-less receptors may have been absent from the metazoan common ancestor, and were instead acquired through an ancient lateral transfer event.[56]

## Evolutionary biology of pLGICs in unicellular organisms

Among Cys-loop receptors, the presence of two clades including respectively the cationic and anionic members suggest a classification of protist Cys-loop receptors into a cationic-like and an anionic-like group. In contrast, the present phylogenetic study gives no indication on the potential ion selectivity of Pro-loop channels outside the Cys-loop clade, whether in eukaryotic or prokaryotic organisms. As the number of known pLGICs in unicellulars grows, it becomes more striking that we know no biological role for any of them. The *in vivo* function of GLIC, although by far the best-characterized prokaryotic pLGICs, is unknown. This is largely explained by the difficulty to grow *Gloeobacter* in the laboratory. The ELIC channel is sensitive to GABA; it has been hypothesized that its function is linked to degradation of amino-acids in plant roots by *Dickeya dadantii* [8].

Chimeric receptors [38, 40, 41] and reverse-selectivity mutants [33, 57] have shown that the respective structures of the ECD and TMD can dictate ligand and ion selectivity independently, which implies that these two domains may, in principle, respond independently to evolutionary pressure regarding either ion or ligand specificity. The conserved core pictured in Fig 3 may then be seen as a universal adapter that allows these two modular structural elements to communicate after perhaps two billion years of independent evolution. In practice, the phylogeny of Cys-loop receptors reflects an early differentiation of anionic and cationic channels, followed by that of large clades of receptors with unique ligands in early Bilateria or earlier [58]. Ligand sensitivity shows more evolutionary mobility than ion selectivity: indeed, cases of convergent evolution of ligand sensitivity has been discussed in some Cys-loop receptors. Dent mentions nematode acetylcholine-gated chloride channels, whose neurotransmitter binding site “represents a unique structural solution to the problem of binding acetylcholine” [21]. Cases of homoplasy have also been reported for glutamate sensitivity of GluCl receptors [20, 59].

The M2 helix, like the ECD, is variable, but in metazoans it exhibits either of the GE[K/R] and A[R/K] motifs, which correlate with ionic selectivity (respectively cationic and anionic) [4]. For prokaryotic and protist sequences, this location in the alignment is in general not sufficient to predict ionic selectivity, as the residues of the two motifs are often substituted, only the final, basic residue being conserved.

## Limitations of this study

The high divergence of the superfamily makes both the search for homologues and subsequent sequence alignment challenging. Homologous sequences could exist in the currently sequenced genomes and yet evade detection for two reasons: genome assembly errors resulting in the transcript being incorrectly predicted or not at all; or a high degree of divergence making similarity to other pLGICs undetectable. The main argument pointing to the robustness of the search is the mostly similar sets discovered by Psi-BLAST and by HMM-based methods such as HMMer and InterPro.

While the Pro-loop motif Y/FPxD is found to be the single best marker of the superfamily, it could be argued that this reasoning is circular, as that motif contributes significantly to the statistical model built and used to detect pLGICs. It is hence entirely possible that pLGIC sequences divergent enough to be missing this motif would defeat remote homology searches and escape detection entirely, although a handful of sequences are detected although they lack the motif. One would expect to find a hint of this phenomenon in “gray area” sequences that are neither similar enough to known pLGICs, nor different enough from them to conclusively decide whether they belong to the superfamily. In practice, the edge cases are mostly fragments, or appear to combine fragments of pLGIC sequence with unrelated fragments, and may reflect unreliable genome assembly rather than actual gene variants. One limitation of a full-length

search including the transmembrane domain is that sequences with a similar membrane topology appear close to the detection threshold due to the lower sequence complexity of membrane-spanning segments. The requirement of a full-length match allows such cases to be eliminated.

The reliability of the inferred molecular phylogeny is limited by the evolutionary diversity of this set of sequences, and their ancient divergence. First, long divergence may lead to long-branch attraction. Second, mutation saturation may occur in the less conserved regions, leading to loss of phylogenetic signal. Finally, rooting the tree is made difficult by differences in evolutionary rates, which makes mid-point rooting meaningless, as well as a nonlinear evolutionary history. The most likely root lies deep within prokaryotic branches, yet root placement is approximate. Moreover, the relationship between the Pro-loop receptors of Archaea, Eubacteria, and Eukaryota is as unclear as the phylogenetic relationship between those clades themselves, that is, the placement of the root of the tree of life [60].

Due to the difficulty of aligning many divergent sequences, in this work we intentionally discarded terminal parts of the sequence before and after the two signature domains of Pro-loop receptors. Furthering the work started by Tasneem et al. [6] of studying the domain architectures of all members of the superfamily could lead to much insight into the roles of these proteins in unicellular organisms, and possibly non-bilaterian animals. Still, we note that the association of the classic pLGIC architecture with other domains has only been evidenced at the genome level so far: any biological conclusions would require experimental validation.

Our molecular phylogeny displays incongruencies with the species phylogeny, which carries some uncertainty itself. It could be argued that this reflects not just a non-linear evolutionary history but the dynamic nature of prokaryotic genomes, which challenges their very representation as a “tree of Life” [49].

## Perspectives for future work

These results leave many questions open, the deepest of which is the ancient evolutionary history of pLGICs. In what taxon did they appear? How were they acquired by ancient eukaryotes?

Co-evolution analysis may point to evolutionary networks: binding sites, interfaces between subunits or between TM helices, signaling network between the ECD and TM domains. At the genome level, it would be interesting to search for co-evolution of pLGIC genes with genes involved in the nervous system in animals.

Since the TMD and ECD seem to dictate ion and ligand specificity, respectively, recombining these segments among paralogous genes may have been an evolutionary path to functionally novel receptors. This could be detected as incongruencies between phylogenies inferred separately from each domain, although isolated TMD sequences are likely to yield poor phylogenetic signal, making the analysis challenging.

The number of whole sequenced genomes for unicellular eukaryotes is still small. As more data becomes available, it should become possible to confirm or falsify the hypotheses put forth here.

## Methods

### Searching protein sequence databases for pLGIC homologues

In the spirit of Rendon et al. [17], we tested three independent approaches: iterative BLAST (Psi-BLAST [61]) and two methods based on hidden Markov models: HMMer [62], and Interpro [63]. Searches were run on the Uniprot SwissProt and TrEMBL databases, and NCBI Refseq and NR, using searches in larger and less well-curated databases to complete results from the smaller ones.

Psi-BLAST was found to be the most sensitive method by Rendon et al. We ran an iterative psi-BLAST [61] search (max. target sequences:5000, BLOSUM 45 matrix, e-value threshold: 1). Due to the over-representation of metazoan proteins in databases, unrestricted searches result in Position-Specific Scoring Matrices (PSSM) that are, in practice, characteristic of metazoan sequences. To avoid such a bias, the search was restricted by excluding the kingdom Metazoa.

We stopped at the third iteration due to a sharp increase in the number of false positives (detected as explained below). To mitigate the low specificity of the psi-BLAST search, we used two methods that were expected to be more specific: a domain-based search in InterPro [64] and an HMM-based search with HMMERsearch [65].

The HMM-based search was run using the HMMer server (hmmer.janelia.org) [62]. The initial input was the modest-sized alignment of metazoan and bacterial pLGIC sequences published by Tasneem et al. An HMM profile was created from this alignment and used to screen the Uniprot database.

As protein databases contain many thousands of metazoan pLGIC sequences, an unrestricted search yields a very large dataset that is heavily biased toward the Metazoa kingdom. To avoid that imbalance, we restricted the metazoan search space to a set of 31 representative metazoan species (S1 Table). The resulting dataset still contained 2400 metazoan sequences, and only about 300 sequences outside of Metazoa. A second HMM was constructed by randomly pruning the metazoan data down to 200 sequences, which formed a set of 500 sequences when joined with all microorganism sequences. This sub-sampled dataset is more balanced with respect to taxonomic distribution, which was expected to reduce biases in the alignment or sequence profiles due to the over-representation of Metazoa. The less focused HMM obtained by the alignment of this dataset was used as query in a new search with HMMsearch on Uniprot database with the Metazoa kingdom excluded.

A final search performed on the larger NR database provided hits that represented new species: only those hits were added to the Uniprot-derived database.

Separately, a domain-based search was performed using three InterPro signatures [63] common to the all pLGICs: the family signature IPR006201 (Neurotransmitter-gated ion-channel), and both individual domain signatures IPR006202 (Neurotransmitter-gated ion-channel ligand-binding domain) and IPR006029 (Neurotransmitter-gated ion-channel transmembrane domain). Hits from that search not retrieved by HMMer, typically because they belonged to subsets of Uniprot that were not available for scanning by HMMer, were retrieved, validated, and added to the dataset.

## Hit validation

A hit validation procedure was performed, combining reverse BLAST, a topological prediction, and pairwise alignment with the sequences of GLIC and GluCl. A hit to be validated as a true positive if it fulfilled at least two of the following criteria:

- A reverse BLAST starting from that sequence returns known pLGIC sequences in the first hits (keeping the same parameters as in the initial psi-BLAST search).
- The membrane as predicted by TMHMM [66] is compatible with a pLGIC subunit, or at least a topology compatible with a truncated pLGIC sequence. To account for the uncertainty of the topology prediction, and based on tests on known true positives, we used a threshold score of 0.2 to consider a residue as potentially part of a transmembrane helix, taking into account the fact that pore-facing helix M2 is less hydrophobic and thus not consistently detected as transmembrane.
- the sequence aligns well with GLIC, GluCl and other known pLGICs.

Our final dataset contains all sequences from microorganisms sequences and a representative group of 31 metazoan species. This still yields a large number of metazoan sequences (2088), mainly because animal genomes typically contain many pLGIC paralogues. For some applications, metazoan sequences were sub-sampled as described above to form a more balanced set of 500 sequences when joined with all microorganism sequences. This sub-sampled dataset is more balanced with respect to taxonomic distribution, which was expected to reduce biases in the alignment or sequence profiles. This set was complemented with all Cys-less metazoan sequences that were detected by a regular expression test on all hits, finding regions that matched the Pro-loop motif, but lacked one or both of the cysteines.

## Multiple sequence alignments

Due to the low quality of some sequences from uncurated databases, including sequences of very different lengths, some manual filtering was needed to enhance the quality of the alignment. Sequences that contained only small homologous fragments or large foreign insertions within conserved regions were removed. Extraneous domains at the N and C termini were truncated, either before a segment that aligned with the ECD of known pLGICs, or after the predicted fourth TM helix. This yielded sequences of more comparable lengths, and improved alignments.

Several alignment methods were assessed on our datasets: M-Coffee [67], MAFFT L-INS-i and E-INS-i [68], Muscle [69], HMMalign [66], and Promals3D [70]. The final alignment was produced with MAFFT version 7.266 [68], using the E-INS-i algorithm, which makes minimal assumptions about the sequence set, with up to 1000 steps of iterative refinement, and the BLOSUM62 matrix, which has been shown to exhibit better mathematical consistency than BLOSUM30 (no violations of the triangle inequality) [71]. Alignments were visualized using Jalview [72], annotating the alignment with the topological prediction from TMHMM [66] to validate the alignment of transmembrane segments.

Alignments was validated by combining different assessment criteria:

- comparison with pairwise alignments of well-known pLGICs, including structural alignments;
- lack of indels in key secondary structural elements, notably transmembrane helices;
- proper alignment of transmembrane helices as predicted by TMHMM; helix M4 proved particularly sensitive to alignment parameters.

## Phylogenetic inference

The alignment to be used for phylogenetic inference was filtered using TRIMAL [73] using the *gappyout* setting.

When developing the analysis workflow and working with intermediate datasets, we used FastTree 2 [74] to infer near-maximum-likelihood phylogenies extremely rapidly (less than half a minute for our data, on a single core of a desktop workstation).

After experimenting with filtering by TRIMAL [73] and GBlocks [75], and aware that strict filtering may worsen the accuracy of phylogenetic inference [37], we opted for a loose (default) filtering by Noisy [76], which aims at removing homoplastic positions that would drive phylogenetic inference into error. As the resulting alignment was still highly gapped and contained positions that were not reliably aligned, we filtered out sites with more than 50% gaps.

Assessment of the dataset by ProtTest [77] found the LG amino-acid substitution model [78] with gamma-distributed rates and the observed amino-acid frequencies (LG+G+F) to be optimal for this dataset. The maximum likelihood tree was inferred by RAxML, using LG+G+F, using subtree pruning and regrafting moves (SPR). Branch support was estimated by the Shimodaira x2013;Hasegawa approximate likelihood ratio test (SH-aLRT)[79] as implemented in RAxML, because ML bootstrap is known to be overly conservative [79].

Notung 2.6 [80] was used to reconcile the gene tree with the phylogenetic tree of pLGIC-possessing species, extracted from the tree of life of NCBI Taxonomy. The approach aims at placing the root of the gene tree to minimize the number of gene duplications and losses in the overall evolutionary model.

Phylogenetic trees of pLGIC genes were visualized and rendered in interactive Tree of Life [81]. Species trees were rendered with Archaeopteryx [82].

## Supporting Information

**S1 Dataset. Multiple alignment of pLGIC sequences from metazoans, archae, bacteria, and unicellular eukaryotes.**

(ZIP)

**S1 Fig. Sequence profile logo for the alignment of the balanced set of pLGIC sequences including prokaryotes and eukaryotes.**

(PDF)

**S1 Table. List of the 31 representative metazoan species with their taxid and Uniprot species identifier.**

(PDF)

## Acknowledgments

We are grateful to Drs. Alessandra Carbone (Université Pierre et Marie Curie), Céline Brochier-Armanet (Université Claude Bernard), Grace Brannigan (Rutgers University), and Pierre-Jean Corringer and Francis Biville (Institut Pasteur) for fruitful discussions. This work was supported by a grant from the French Agence Nationale de la Recherche (Project PentaGate).

## Author Contributions

Conceived and designed the experiments: JH. Performed the experiments: MJ JH. Analyzed the data: MJ AT JH. Contributed reagents/materials/analysis tools: MJ JH. Wrote the paper: MJ AT JH.

## References

1. Le Novère N, Changeux JP. Molecular evolution of the nicotinic acetylcholine receptor: an example of multigene family in excitable cells. *J Mol Evol.* 1995; 40(2):155–172. doi: [10.1007/BF00167110](https://doi.org/10.1007/BF00167110) PMID: [7699721](https://pubmed.ncbi.nlm.nih.gov/7699721/)
2. Ortells MO, Lunt GG. Evolutionary history of the ligand-gated ion-channel superfamily of receptors. *Trends Neurosci.* 1995 Mar; 18(3):121–127. PMID: [7754520](https://pubmed.ncbi.nlm.nih.gov/7754520/)
3. Changeux JP, Edelstein SJ. Allosteric mechanisms of signal transduction. *Science.* 2005; 308(5727):1424–1428. doi: [10.1126/science.1108595](https://doi.org/10.1126/science.1108595) PMID: [15933191](https://pubmed.ncbi.nlm.nih.gov/15933191/)
4. Corringer PJ, Poitevin F, Prevost MS, Sauguet L, Delarue M, Changeux JP. Structure and pharmacology of pentameric receptor channels: from bacteria to brain. *Structure.* 2012; 20(6):941–956. doi: [10.1016/j.str.2012.05.003](https://doi.org/10.1016/j.str.2012.05.003) PMID: [22681900](https://pubmed.ncbi.nlm.nih.gov/22681900/)
5. Cockcroft VB, Osguthorpe DJ, Barnard EA, Friday AE, Lunt GG. Ligand-gated ion channels. Homology and diversity. *Mol Neurobiol.* 1990; 4(3–4):129–169. doi: [10.1007/BF02780338](https://doi.org/10.1007/BF02780338) PMID: [1725701](https://pubmed.ncbi.nlm.nih.gov/1725701/)

6. Tasneem A, Iyer LM, Jakobsson E, Aravind L. Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. *Genome Biol.* 2004; 6(1):R4. doi: [10.1186/gb-2004-6-1-r4](https://doi.org/10.1186/gb-2004-6-1-r4) PMID: [15642096](https://pubmed.ncbi.nlm.nih.gov/15642096/)
7. Bocquet N, Prado de Carvalho L, Cartaud J, Neyton J, Le Poupon C, Taly A, et al. A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. *Nature.* 2007; 445(7123):116–119. doi: [10.1038/nature05371](https://doi.org/10.1038/nature05371) PMID: [17167423](https://pubmed.ncbi.nlm.nih.gov/17167423/)
8. Zimmermann I, Dutzler R. Ligand activation of the prokaryotic pentameric ligand-gated ion channel ELIC. *PLoS Biol.* 2011; 9(6):e1001101. doi: [10.1371/journal.pbio.1001101](https://doi.org/10.1371/journal.pbio.1001101) PMID: [21713033](https://pubmed.ncbi.nlm.nih.gov/21713033/)
9. Hilf RJC, Dutzler R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature.* 2008; 452(7185):375–379. doi: [10.1038/nature06717](https://doi.org/10.1038/nature06717) PMID: [18322461](https://pubmed.ncbi.nlm.nih.gov/18322461/)
10. Bocquet N, Nury H, Baaden M, Poupon CL, Changeux JP, Delarue M, et al. X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature.* 2009; 457(7225):111–114. doi: [10.1038/nature07462](https://doi.org/10.1038/nature07462) PMID: [18987633](https://pubmed.ncbi.nlm.nih.gov/18987633/)
11. Hilf RJC, Dutzler R. Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel. *Nature.* 2009; 457(7225):115–118. doi: [10.1038/nature07461](https://doi.org/10.1038/nature07461) PMID: [18987630](https://pubmed.ncbi.nlm.nih.gov/18987630/)
12. Sauguet L, Poitevin F, Murail S, Van Renterghem C, Moraga-Cid G, Malherbe L, et al. Structural basis for ion permeation mechanism in pentameric ligand-gated ion channels. *EMBO J.* 2013; 32(5):728–741. doi: [10.1038/emboj.2013.17](https://doi.org/10.1038/emboj.2013.17) PMID: [23403925](https://pubmed.ncbi.nlm.nih.gov/23403925/)
13. Spurny R, Billen B, Howard RJ, Brams M, Debaveye S, Price KL, et al. Multisite binding of a general anesthetic to the prokaryotic pentameric *Erwinia chrysanthemi* ligand-gated ion channel (ELIC). *J Biol Chem.* 2013; 288(12):8355–8364. doi: [10.1074/jbc.M112.424507](https://doi.org/10.1074/jbc.M112.424507) PMID: [23364792](https://pubmed.ncbi.nlm.nih.gov/23364792/)
14. Hibbs RE, Gouaux E. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature.* 2011; 474(7349):54–60. doi: [10.1038/nature10139](https://doi.org/10.1038/nature10139) PMID: [21572436](https://pubmed.ncbi.nlm.nih.gov/21572436/)
15. Miller PS, Aricescu AR. Crystal structure of a human GABAA receptor. *Nature.* 2014; 512(7514):270–275. doi: [10.1038/nature13293](https://doi.org/10.1038/nature13293) PMID: [24909990](https://pubmed.ncbi.nlm.nih.gov/24909990/)
16. Hassaine G, Deluz C, Grasso L, Wyss R, Tol MB, Hovius R, et al. X-ray structure of the mouse serotonin 5-HT3 receptor. *Nature.* 2014; 512(7514):276–281. doi: [10.1038/nature13552](https://doi.org/10.1038/nature13552) PMID: [25119048](https://pubmed.ncbi.nlm.nih.gov/25119048/)
17. Rendon G, Kantorovitz MR, Tilson JL, Jakobsson E. Identifying bacterial and archaeal homologs of pentameric ligand-gated ion channel (pLGIC) family using domain-based and alignment-based approaches. *Channels (Austin).* 2011; 5(4):325–343. doi: [10.4161/chan.5.4.16822](https://doi.org/10.4161/chan.5.4.16822)
18. Corringer PJ, Baaden M, Bocquet N, Delarue M, Dufresne V, Nury H, et al. Atomic structure and dynamics of pentameric ligand-gated ion channels: new insight from bacterial homologues. *J Physiol.* 2010; 588(Pt 4):565–572. doi: [10.1113/jphysiol.2009.183160](https://doi.org/10.1113/jphysiol.2009.183160) PMID: [19995852](https://pubmed.ncbi.nlm.nih.gov/19995852/)
19. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of *Hydra*. *Nature.* 2010; 464(7288):592–596. doi: [10.1038/nature08830](https://doi.org/10.1038/nature08830) PMID: [20228792](https://pubmed.ncbi.nlm.nih.gov/20228792/)
20. Kehoe J, Buldakova S, Acher F, Dent J, Bregestovski P, Bradley J. Aplysia cys-loop glutamate-gated chloride channels reveal convergent evolution of ligand specificity. *J Mol Evol.* 2009; 69(2):125–141. doi: [10.1007/s00239-009-9256-z](https://doi.org/10.1007/s00239-009-9256-z) PMID: [19554247](https://pubmed.ncbi.nlm.nih.gov/19554247/)
21. Dent JA. The evolution of pentameric ligand-gated ion channels. *Adv Exp Med Biol.* 2010; 683:11–23. doi: [10.1007/978-1-4419-6445-8\\_2](https://doi.org/10.1007/978-1-4419-6445-8_2) PMID: [20737785](https://pubmed.ncbi.nlm.nih.gov/20737785/)
22. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature.* 2008; 451(7180):783–788. doi: [10.1038/nature06617](https://doi.org/10.1038/nature06617) PMID: [18273011](https://pubmed.ncbi.nlm.nih.gov/18273011/)
23. Suga H, Chen Z, de Mendoza A, Sebe-Pedros A, Brown MW, Kramer E, et al. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat Commun.* 2013; 4:2325. doi: [10.1038/ncomms3325](https://doi.org/10.1038/ncomms3325) PMID: [23942320](https://pubmed.ncbi.nlm.nih.gov/23942320/)
24. Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biol.* 2013; 14(2):R15. doi: [10.1186/gb-2013-14-2-r15](https://doi.org/10.1186/gb-2013-14-2-r15) PMID: [23419129](https://pubmed.ncbi.nlm.nih.gov/23419129/)
25. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol.* 2008; 6(3):245–252. doi: [10.1038/nrmicro1852](https://doi.org/10.1038/nrmicro1852) PMID: [18274537](https://pubmed.ncbi.nlm.nih.gov/18274537/)
26. Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature.* 2015; 521(7551):173–179. doi: [10.1038/nature14447](https://doi.org/10.1038/nature14447) PMID: [25945739](https://pubmed.ncbi.nlm.nih.gov/25945739/)
27. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, et al. The *Trichoplax* genome and the nature of placozoans. *Nature.* 2008; 454(7207):955–960. doi: [10.1038/nature07191](https://doi.org/10.1038/nature07191) PMID: [18719581](https://pubmed.ncbi.nlm.nih.gov/18719581/)

28. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*. 2010 Aug; 466:720–726. doi: [10.1038/nature09201](https://doi.org/10.1038/nature09201) PMID: [20686567](https://pubmed.ncbi.nlm.nih.gov/20686567/)
29. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, et al. The ctenophore genome and the evolutionary origins of neural systems. *Nature*. 2014; 510(7503):109–114. doi: [10.1038/nature13400](https://doi.org/10.1038/nature13400) PMID: [24847885](https://pubmed.ncbi.nlm.nih.gov/24847885/)
30. Yates DM, Wolstenholme AJ. *Dirofilaria immitis*: identification of a novel ligand-gated ion channel-related polypeptide. *Exp Parasitol*. 2004; 108(3–4):182–185. doi: [10.1016/j.exppara.2004.08.003](https://doi.org/10.1016/j.exppara.2004.08.003) PMID: [15582516](https://pubmed.ncbi.nlm.nih.gov/15582516/)
31. Purohit P, Chakraborty S, Auerbach A. Function of the M1  $\pi$ -helix in endplate receptor activation and desensitization. *J Physiol*. 2015 Apr; doi: [10.1113/JP270223](https://doi.org/10.1113/JP270223) PMID: [25929452](https://pubmed.ncbi.nlm.nih.gov/25929452/)
32. Imoto K, Methfessel C, Sakmann B, Mishina M, Mori Y, Konno T, et al. Location of a delta-subunit region determining ion transport through the acetylcholine receptor channel. *Nature*. 1986; 324(6098):670–674. doi: [10.1038/324670a0](https://doi.org/10.1038/324670a0) PMID: [2432430](https://pubmed.ncbi.nlm.nih.gov/2432430/)
33. Corringer PJ, Bertrand S, Galzi JL, Devillers-Thiéry A, Changeux JP, Bertrand D. Mutational analysis of the charge selectivity filter of the alpha7 nicotinic acetylcholine receptor. *Neuron*. 1999; 22(4):831–843. doi: [10.1016/S0896-6273\(00\)80741-2](https://doi.org/10.1016/S0896-6273(00)80741-2) PMID: [10230802](https://pubmed.ncbi.nlm.nih.gov/10230802/)
34. Cadugan DJ, Auerbach A. Conformational dynamics of the alphaM3 transmembrane helix during acetylcholine receptor channel gating. *Biophys J*. 2007; 93(3):859–865. doi: [10.1529/biophysj.107.105171](https://doi.org/10.1529/biophysj.107.105171) PMID: [17513382](https://pubmed.ncbi.nlm.nih.gov/17513382/)
35. Scott S, Lynch JW, Keramidas A. Correlating structural and energetic changes in glycine receptor activation. *J Biol Chem*. 2015; 290(9):5621–5634. doi: [10.1074/jbc.M114.616573](https://doi.org/10.1074/jbc.M114.616573) PMID: [25572390](https://pubmed.ncbi.nlm.nih.gov/25572390/)
36. Hénin J, Salari R, Murlidaran S, Brannigan G. A Predicted Binding Site for Cholesterol on the GABAA Receptor. *Biophys J*. 2014; 106(9):1938–1949. doi: [10.1016/j.bpj.2014.03.024](https://doi.org/10.1016/j.bpj.2014.03.024) PMID: [24806926](https://pubmed.ncbi.nlm.nih.gov/24806926/)
37. Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, et al. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst Biol*. 2015 Sep; 64(5):778–791. doi: [10.1093/sysbio/syv033](https://doi.org/10.1093/sysbio/syv033) PMID: [26031838](https://pubmed.ncbi.nlm.nih.gov/26031838/)
38. Eiselé JL, Bertrand S, Galzi JL, Devillers-Thiéry A, Changeux JP, Bertrand D. Chimaeric nicotinic-serotonergic receptor combines distinct ligand binding and channel specificities. *Nature*. 1993; 366(6454):479–483. doi: [10.1038/366479a0](https://doi.org/10.1038/366479a0) PMID: [8247158](https://pubmed.ncbi.nlm.nih.gov/8247158/)
39. Grutter T, de Carvalho LP, Dufresne V, Taly A, Edelstein SJ, Changeux JP. Molecular tuning of fast gating in pentameric ligand-gated ion channels. *Proc Natl Acad Sci U S A*. 2005; 102(50):18207–18212. doi: [10.1073/pnas.0509024102](https://doi.org/10.1073/pnas.0509024102) PMID: [16319224](https://pubmed.ncbi.nlm.nih.gov/16319224/)
40. Duret G, Van Renterghem C, Weng Y, Prevost M, Moraga-Cid G, Huon C, et al. Functional prokaryotic-eukaryotic chimera from the pentameric ligand-gated ion channel family. *Proc Natl Acad Sci U S A*. 2011; 108(29):12143–12148. doi: [10.1073/pnas.1104494108](https://doi.org/10.1073/pnas.1104494108) PMID: [21730130](https://pubmed.ncbi.nlm.nih.gov/21730130/)
41. Mnatsakanyan N, Nishtala SN, Pandhare A, Fiori MC, Goyal R, Pauwels JE, et al. Functional Chimeras of GLIC Obtained by Adding the Intracellular Domain of Anion- and Cation-Conducting Cys-Loop Receptors. *Biochemistry*. 2015; 54(16):2670–2682. doi: [10.1021/acs.biochem.5b00203](https://doi.org/10.1021/acs.biochem.5b00203) PMID: [25861708](https://pubmed.ncbi.nlm.nih.gov/25861708/)
42. Mishina M, Tobimatsu T, Imoto K, Tanaka K, Fujita Y, Fukuda K, et al. Location of functional regions of acetylcholine receptor alpha-subunit by site-directed mutagenesis. *Nature*. 1985; 313(6001):364–369. doi: [10.1038/313364a0](https://doi.org/10.1038/313364a0) PMID: [3838201](https://pubmed.ncbi.nlm.nih.gov/3838201/)
43. Rajendra S, Vandenberg RJ, Pierce KD, Cunningham AM, French PW, Barry PH, et al. The unique extracellular disulfide loop of the glycine receptor is a principal ligand binding element. *EMBO J*. 1995; 14(13):2987–2998. PMID: [7621814](https://pubmed.ncbi.nlm.nih.gov/7621814/)
44. Lobo IA, Trudell JR, Harris RA. Cross-linking of glycine receptor transmembrane segments two and three alters coupling of ligand binding with channel opening. *J Neurochem*. 2004; 90(4):962–969. doi: [10.1111/j.1471-4159.2004.02561.x](https://doi.org/10.1111/j.1471-4159.2004.02561.x) PMID: [15287902](https://pubmed.ncbi.nlm.nih.gov/15287902/)
45. Jansen M, Akabas MH. State-dependent cross-linking of the M2 and M3 segments: functional basis for the alignment of GABAA and acetylcholine receptor M3 segments. *J Neurosci*. 2006; 26(17):4492–4499. doi: [10.1523/JNEUROSCI.0224-06.2006](https://doi.org/10.1523/JNEUROSCI.0224-06.2006) PMID: [16641228](https://pubmed.ncbi.nlm.nih.gov/16641228/)
46. Bali M, Jansen M, Akabas MH. GABA-induced intersubunit conformational movement in the GABAA receptor alpha 1M1-beta 2M3 transmembrane subunit interface: experimental basis for homology modeling of an intravenous anesthetic binding site. *J Neurosci*. 2009; 29(10):3083–3092. doi: [10.1523/JNEUROSCI.6090-08.2009](https://doi.org/10.1523/JNEUROSCI.6090-08.2009) PMID: [19279245](https://pubmed.ncbi.nlm.nih.gov/19279245/)
47. Mnatsakanyan N, Jansen M. Experimental determination of the vertical alignment between the second and third transmembrane segments of muscle nicotinic acetylcholine receptors. *J Neurochem*. 2013; 125(6):843–854. doi: [10.1111/jnc.12260](https://doi.org/10.1111/jnc.12260) PMID: [23565737](https://pubmed.ncbi.nlm.nih.gov/23565737/)

48. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton *Emiliana* underpins its global distribution. *Nature*. 2013; 499(7457):209–213. doi: [10.1038/nature12221](https://doi.org/10.1038/nature12221) PMID: [23760476](https://pubmed.ncbi.nlm.nih.gov/23760476/)
49. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*. 2014; 12:66. doi: [10.1186/s12915-014-0066-4](https://doi.org/10.1186/s12915-014-0066-4) PMID: [25141959](https://pubmed.ncbi.nlm.nih.gov/25141959/)
50. Liebeskind BJ, Hillis DM, Zakon HH. Convergence of ion channel genome content in early animal evolution. *Proc Natl Acad Sci U S A*. 2015 Feb; 112(8):E846–E851. doi: [10.1073/pnas.1501195112](https://doi.org/10.1073/pnas.1501195112) PMID: [25675537](https://pubmed.ncbi.nlm.nih.gov/25675537/)
51. Dufour V, Beech RN, Wever C, Dent JA, Geary TG. Molecular Cloning and Characterization of Novel Glutamate-Gated Chloride Channel Subunits from *Schistosoma mansoni*. *PLoS Pathog*. 2013; 9(8): e1003586. doi: [10.1371/journal.ppat.1003586](https://doi.org/10.1371/journal.ppat.1003586) PMID: [24009509](https://pubmed.ncbi.nlm.nih.gov/24009509/)
52. Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, Patterson DJ, et al. Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet*. 2006; 2(12):e220. doi: [10.1371/journal.pgen.0020220](https://doi.org/10.1371/journal.pgen.0020220) PMID: [17194223](https://pubmed.ncbi.nlm.nih.gov/17194223/)
53. Distel DL, Morrill W, MacLaren-Toussaint N, Franks D, Waterbury J. *Teredinibacter turnerae* gen. nov., sp. nov., a dinitrogen-fixing, cellulolytic, endosymbiotic gamma-proteobacterium isolated from the gills of wood-boring molluscs (Bivalvia: Teredinidae). *Int J Syst Evol Microbiol*. 2002; 52(Pt 6):2261–2269. PMID: [12508896](https://pubmed.ncbi.nlm.nih.gov/12508896/)
54. Avila-Poveda OH, Torres-Ariño A, Girón-Cruz DA, Cuevas-Aguirre A. Evidence for accumulation of *Synechococcus elongatus* (Cyanobacteria: Cyanophyceae) in the tissues of the oyster *Crassostrea gigas* (Mollusca: Bivalvia). *Tissue Cell*. 2014; 46(5):379–387. doi: [10.1016/j.tice.2014.07.001](https://doi.org/10.1016/j.tice.2014.07.001) PMID: [25109761](https://pubmed.ncbi.nlm.nih.gov/25109761/)
55. Bailly X, Laguerre L, Correc G, Dupont S, Kurth T, Pfannkuchen A, et al. The chimerical and multifaceted marine acoel *Symsagittifera roscoffensis*: from photosymbiosis to brain regeneration. *Front Microbiol*. 2014; 5:498. doi: [10.3389/fmicb.2014.00498](https://doi.org/10.3389/fmicb.2014.00498) PMID: [25324833](https://pubmed.ncbi.nlm.nih.gov/25324833/)
56. Dunning Hotopp JC, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Muñoz Torres MC, et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*. 2007; 317(5845):1753–1756. doi: [10.1126/science.1142490](https://doi.org/10.1126/science.1142490) PMID: [17761848](https://pubmed.ncbi.nlm.nih.gov/17761848/)
57. Tillman T, Cheng MH, Chen Q, Tang P, Xu Y. Reversal of ion-charge selectivity renders the pentameric ligand-gated ion channel GLIC insensitive to anaesthetics. *Biochem J*. 2013; 449(1):61–68. doi: [10.1042/BJ20121072](https://doi.org/10.1042/BJ20121072) PMID: [22978431](https://pubmed.ncbi.nlm.nih.gov/22978431/)
58. Dent JA. Evidence for a diverse Cys-loop ligand-gated ion channel superfamily in early bilateria. *J Mol Evol*. 2006; 62(5):523–535. doi: [10.1007/s00239-005-0018-2](https://doi.org/10.1007/s00239-005-0018-2) PMID: [16586016](https://pubmed.ncbi.nlm.nih.gov/16586016/)
59. Lynagh T, Beech RN, Lalande MJ, Keller K, Cromer BA, Wolstenholme AJ, et al. Molecular basis for convergent evolution of glutamate recognition by pentameric ligand-gated ion channels. *Sci Rep*. 2015; 5:8558. doi: [10.1038/srep08558](https://doi.org/10.1038/srep08558) PMID: [25708000](https://pubmed.ncbi.nlm.nih.gov/25708000/)
60. Forterre P, Philippe H. Where is the root of the universal tree of life? *Bioessays*. 1999; 21(10):871–879. doi: [10.1002/\(SICI\)1521-1878\(199910\)21:10%3C871::AID-BIES10%3E3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1521-1878(199910)21:10%3C871::AID-BIES10%3E3.0.CO;2-Q) PMID: [10497338](https://pubmed.ncbi.nlm.nih.gov/10497338/)
61. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25(17):3389–3402. doi: [10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389) PMID: [9254694](https://pubmed.ncbi.nlm.nih.gov/9254694/)
62. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011; 39(Web Server issue):W29–37. doi: [10.1093/nar/gkr367](https://doi.org/10.1093/nar/gkr367) PMID: [21593126](https://pubmed.ncbi.nlm.nih.gov/21593126/)
63. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res*. 2015; 43(Database issue): D213–D221. doi: [10.1093/nar/gku1243](https://doi.org/10.1093/nar/gku1243) PMID: [25428371](https://pubmed.ncbi.nlm.nih.gov/25428371/)
64. Hunter S, Jones P, Mitchell A, Yong SY. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*. 2012; 40(Database issue):D306–12. doi: [10.1093/nar/gkr948](https://doi.org/10.1093/nar/gkr948) PMID: [22096229](https://pubmed.ncbi.nlm.nih.gov/22096229/)
65. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998; 14(9):755–763. doi: [10.1093/bioinformatics/14.9.755](https://doi.org/10.1093/bioinformatics/14.9.755) PMID: [9918945](https://pubmed.ncbi.nlm.nih.gov/9918945/)
66. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001; 305(3):567–580. doi: [10.1006/jmbi.2000.4315](https://doi.org/10.1006/jmbi.2000.4315) PMID: [11152613](https://pubmed.ncbi.nlm.nih.gov/11152613/)
67. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 2006; 34(6):1692–1699. doi: [10.1093/nar/gkl091](https://doi.org/10.1093/nar/gkl091) PMID: [16556910](https://pubmed.ncbi.nlm.nih.gov/16556910/)

68. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013 Apr; 30(4):772–780. doi: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010) PMID: [23329690](https://pubmed.ncbi.nlm.nih.gov/23329690/)
69. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004; 32(5):1792–1797. doi: [10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340) PMID: [15034147](https://pubmed.ncbi.nlm.nih.gov/15034147/)
70. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* 2008; 36(7):2295–2300. doi: [10.1093/nar/gkn072](https://doi.org/10.1093/nar/gkn072) PMID: [18287115](https://pubmed.ncbi.nlm.nih.gov/18287115/)
71. Baussand J, Carbone A. Inconsistent distances in substitution matrices can be avoided by properly handling hydrophobic residues. *Evol Bioinform Online.* 2008; 4:255–261. PMID: [19204823](https://pubmed.ncbi.nlm.nih.gov/19204823/)
72. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 2009; 25(9):1189–1191. doi: [10.1093/bioinformatics/btp033](https://doi.org/10.1093/bioinformatics/btp033) PMID: [19151095](https://pubmed.ncbi.nlm.nih.gov/19151095/)
73. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009; 25(15):1972–1973. doi: [10.1093/bioinformatics/btp348](https://doi.org/10.1093/bioinformatics/btp348) PMID: [19505945](https://pubmed.ncbi.nlm.nih.gov/19505945/)
74. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One.* 2010; 5(3):e9490. doi: [10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490) PMID: [20224823](https://pubmed.ncbi.nlm.nih.gov/20224823/)
75. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000; 17(4):540–552. doi: [10.1093/oxfordjournals.molbev.a026334](https://doi.org/10.1093/oxfordjournals.molbev.a026334) PMID: [10742046](https://pubmed.ncbi.nlm.nih.gov/10742046/)
76. Dress AW, Flamm C, Fritzsche G, Grünewald S, Kruspe M, Prohaska SJ, et al. Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology.* 2008; 3(1):1–10. doi: [10.1186/1748-7188-3-7](https://doi.org/10.1186/1748-7188-3-7)
77. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics.* 2005; 21(9):2104–2105. doi: [10.1093/bioinformatics/bti263](https://doi.org/10.1093/bioinformatics/bti263) PMID: [15647292](https://pubmed.ncbi.nlm.nih.gov/15647292/)
78. Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol.* 2008; 25(7):1307–1320. doi: [10.1093/molbev/msn067](https://doi.org/10.1093/molbev/msn067) PMID: [18367465](https://pubmed.ncbi.nlm.nih.gov/18367465/)
79. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology.* 2011; 60(5):685–699. doi: [10.1093/sysbio/syr041](https://doi.org/10.1093/sysbio/syr041) PMID: [21540409](https://pubmed.ncbi.nlm.nih.gov/21540409/)
80. Vernet B, Stolzer M, Goldman A, Durand D. Reconciliation with non-binary species trees. *Comput Syst Bioinformatics Conf.* 2007; 6:441–452. doi: [10.1142/9781860948732\\_0044](https://doi.org/10.1142/9781860948732_0044) PMID: [17951846](https://pubmed.ncbi.nlm.nih.gov/17951846/)
81. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 2011; 39(Web Server issue):W475–W478. doi: [10.1093/nar/gkr201](https://doi.org/10.1093/nar/gkr201) PMID: [21470960](https://pubmed.ncbi.nlm.nih.gov/21470960/)
82. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics.* 2009; 10:356. doi: [10.1186/1471-2105-10-356](https://doi.org/10.1186/1471-2105-10-356) PMID: [19860910](https://pubmed.ncbi.nlm.nih.gov/19860910/)