AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

The Evolutionary Kaleidoscope of Rhodopsins

Paul-Adrian Bulzu,[a] Vinicius S. Kavagutti,[a,b] Adrian-Stefan Andrei,[c] Rohit Ghai[a]

[a]Biology Centre of the Czech Academy of Sciences (CAS), Institute of Hydrobiology, České Budějovice, Czech Republic
[b]Department of Ecosystem Biology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic
[c]University of Zurich, Limnological Station, Microbial Evogenomics Lab (MiEL), Kilchberg, Switzerland

ABSTRACT Rhodopsins are widely distributed across all domains of life where they perform a plethora of functions through the conversion of electromagnetic radiation into physicochemical signals. As a result of an extensive survey of available genomic and metagenomic sequencing data, we reported the existence of novel clades and exotic sequence motifs scattered throughout the evolutionary radiations of both Type-1 and Type-3 rhodopsins that will likely enlarge the optogenetics toolbox. We expanded the typical rhodopsin blueprint by showing that a highly conserved and functionally important arginine residue (i.e., Arg82) was substituted multiple times during evolution by an extensive amino acid spectrum. We proposed the umbrella term Alt-rhodopsins (AltRs) for all such proteins that departed Arg82 orthodoxy. Some AltRs formed novel clades in the rhodopsin phylogeny and were found in giant viruses. Some newly uncovered AltRs were phylogenetically close to heliorhodopsins, which allowed a closer examination of the phylogenetic border between Type-1 rhodopsins and heliorhodopsins. Comprehensive phylogenetic trees and ancestral sequence reconstructions allowed us to advance the hypothesis that proto-heliorhodopsins were a eukaryotic innovation before their subsequent diversification into the extant Type-3 rhodopsins.

IMPORTANCE The rhodopsin scaffold is remarkably versatile and widespread, coupling light availability to energy production and other light-dependent cellular responses with minor alterations to critical residues. We described an unprecedented spectrum of substitutions at one of the most conserved amino acids in the rhodopsin fold, Arg82. We denoted such phylogenetically diverse rhodopsins with the umbrella name Alt-rhodopsins (AltR) and described a distinct branch of AltRs in giant viruses. Intriguingly, some AltRs were the closest phylogenetic neighbors to Heliorhodopsins (HeRs) whose origins have remained enigmatic. Our analyses of HeR origins in the light of AltRs led us to posit a most unusual evolutionary trajectory that suggested a eukaryotic origin for HeRs before their diversification in prokaryotes.

KEYWORDS rhodopsins, Alt-rhodopsins, AltRs, heliorhodopsins, optogenetics, metagenomics

Rhodopsins are remarkably promising molecules for modulating cell expression with precision (1–3), but for many their biological role in the natural environment remains largely obscure. With increasing sequence data, more rhodopsins are being found (4–9), but it is unclear to what extent the sequence diversity of the rhodopsinverse has been explored. Type-1 (microbial rhodopsins) and Type-2 (animal rhodopsins) share similar, seven-helical topological conformation and membrane orientation with the N terminus in the extracellular space and a Schiff base linkage from a conserved lysine to retinal in the seventh helix (TM7) (10). However, while the overall fold is the same, there is no detectable sequence similarity between these two types. A completely new type of rhodopsin similar to Type-1 rhodopsins was identified recently but with inverse membrane orientation (Heliorhodopsins, HeRs, or Type-3 rhodopsins)

Editor Danielle Tullman-Ercek, Northwestern University

Address correspondence to Paul-Adrian Bulzu, bulzupaul@gmail.com.

The authors declare no conflict of interest.

Received 27 April 2022
Accepted 29 August 2022
Published 19 September 2022

(11). Despite their orientation, HeRs also bind the retina using a conserved lysine in TM7 (transmembrane helix 7). Apart from the lysine in TM7 that is essential for binding retinal, several other functionally important residues have been identified, e.g., several characteristic sequence motifs in TM3 and TM7 that may be predictive of the nature of the ion pump. Proteorhodopsins typically display DTE or DTD motifs in TM3 and a DxxxK motif in TM7. Inward chloride pumps may be recognized by NTQ or TSD motifs in TM3 and a DxxxK motif in TM7 and heliorhodopsins have the ESL motif in TM3 and SxxxK in TM7 (12).

One critical and highly conserved residue is Arg82 (BR numbering) in the third transmembrane helix (TM3). Since the discovery of bacteriorhodopsin (BR) in haloarchaea nearly 5 decades ago (13), no naturally occurring rhodopsins are known that do not have a conserved Arg82 residue (12). Among conserved BR amino acids, Arg82 in TM3 was recognized as an essential player within the photocycle due to its involvement in proton release through interactions with Asp85, Asp212, and the retinal Schiff base (14, 15). Such conclusions are the result of experimental work and observations from multiple mutagenesis studies that describe the effects of targeted Arg82 substitutions on BR photocycle and proton release: R82A (14, 16–18), R82C (16), R82H (15), R82K (18–20), and R82Q (14, 17, 20, 21). In general, these studies indicate that the charge and hydrogen bonding capabilities of the residue in position 82 drastically influence the interactions between the proton acceptor (Asp85) and proton release group, thus altering or even abolishing proton release to the extracellular space under normal physiological conditions (14–16, 18). This changed with the discovery of xenorhodopsins (22) that were shown to have a tryptophan (W) or phenylalanine (F) in this position and to function as unusual inward proton pumps (23). Since then, a few substitutions to Arg82 were reported in (i) anion channels (K instead of R) (4), (ii) a few rhodopsins of unknown function (Q, A, and T instead of R) (6), and (iii) potassium pumps (kalium rhodopsins, W instead of R) (5). However, no such substitutions have ever been described in HeRs.

In this work, we performed an extensive search through hundreds of metagenomes and metatranscriptomes and showed that a surprisingly large number of rhodopsins with peculiar amino acid substitutions and novel motifs had remained out of sight. We utilized this newly unearthed diversity to construct large-scale, highly supported phylogenies and to generate a plausible evolutionary scenario for the origin and evolution of Type-3 rhodopsins (heliorhodopsins; HeRs) (11).

RESULTS AND DISCUSSION

Novel clades of unusual rhodopsins. We scanned large collections of genomic, metagenomic, and metatranscriptomic data sets from various sources (e.g., marine, freshwater, brackish, sediments, prokaryotic/eukaryotic genomes, and eukaryotic transcriptomes) to identify novel rhodopsin sequence variants. Sequences with seven transmembrane helices and a conserved lysine in TM7 were considered bonafide rhodopsins (see Materials and Methods for details). We aligned these sequences with known rhodopsins to identify characteristic motifs in TM3 and TM7. Unexpectedly, these alignments also revealed substantial variation in residue 82 (Arg82) that had not been observed before. We found that the known substitutions for this critical TM3 residue (i.e., Arg82) could be significantly expanded by additional changes at this position in both Type-1 and Type-3 rhodopsins. In sum, we found strong evidence supported by at least 10 sequences in each case that residues H, K, Q, A, P, S, Y, E, and M can replace Arg82. Other substitutions (N, I, F, T, L, G, D, V, and C) were also identified but less than 10 times (see Table S1). We proposed the umbrella name Alt-rhodopsins (AltRs) for all such microbial rhodopsins with a substituted Arg82 (i.e., non-R type rhodopsins as opposed to R-type ones with Arg82). This nomenclature included xenorhodopsins and kalium rhodopsins as subtypes of Alt-rhodopsins. Here, we differentiated among the subtypes of Alt-rhodopsins by indicating the substituent amino acid symbol (e.g., H-type when H replaced the canonical R).

AltRs were present in nearly all phylogenetic clades of rhodopsins in bacteria, archaea, eukaryotes, and viruses (Fig. 1) (insofar as taxonomic origin could be reliably

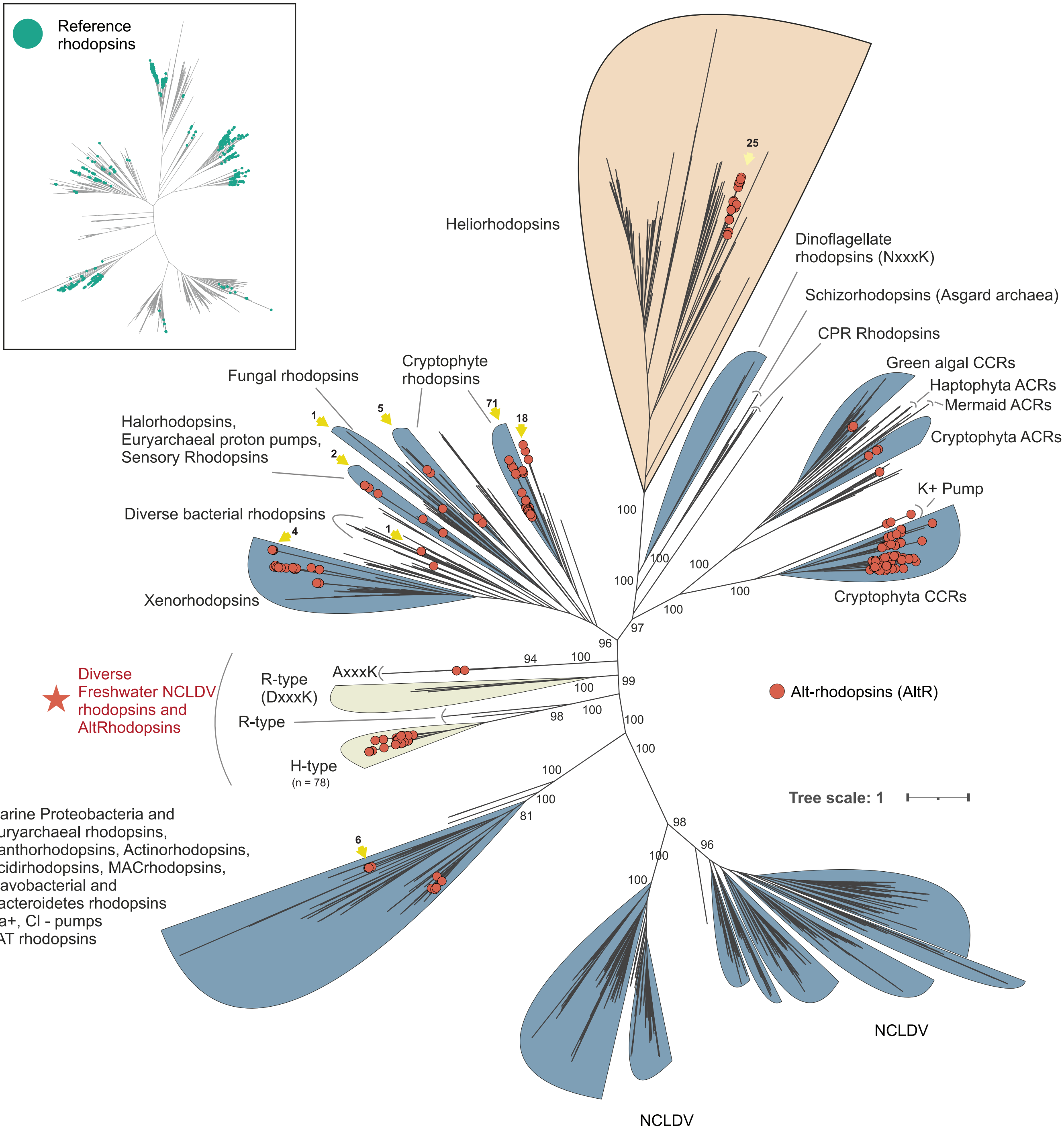
Reference rhodopsins
Heliorhodopsins
25
Dinoflagellate rhodopsins (NxxxK)
Schizorhodopsins (Asgard archaea)
CPR Rhodopsins
Green algal CCRs
Haptophyta ACRs
Mermaid ACRs
Cryptophyta ACRs
K+ Pump
Cryptophyta CCRs
Fungal rhodopsins
Cryptophyte rhodopsins
Halorhodopsins, Euryarchaeal proton pumps, Sensory Rhodopsins
Diverse bacterial rhodopsins
Xenorhodopsins
1
5
71
18
2
1
4
100
100
100
100
100
100
97
96
Diverse Freshwater NCLDV rhodopsins and AltRhodopsins
AxxxK
R-type (DxxxK)
R-type
H-type (n = 78)
94
100
99
100
98
100
100
100
Alt-rhodopsins (AltR)
Tree scale: 1
Marine Proteobacteria and Euryarchaeal rhodopsins, Xanthorhodopsins, Actinorhodopsins, Acidirhodopsins, MACrhodopsins, Flavobacterial and Bacteroidetes rhodopsins Na+, Cl - pumps TAT rhodopsins
6
100
100
81
98
96
100
100
NCLDV
NCLDV

FIG 1 Maximum likelihood phylogenetic tree of rhodopsins. Alt-rhodopsins are indicated by red circles at node tips. Ultrafast bootstrap values are shown at selected nodes. A star indicates the novel clades of rhodopsins and Alt-rhodopsins. Yellow arrows indicate the position of H-type rhodopsins and their counts. The inset at the top left shows a simplified version of the phylogenetic tree marking all reference rhodopsin sequences (in green circles). See also Tables S1 and S2 and FigShare Data at https://figshare.com/s/f2d7b1065930bf350c2f.

ascribed). However, by far the vast majority (*n* = 102) were found in the eukaryotic Cryptophytes. Altogether, we classified 399 sequences as AltRs (see Table S2). At least 121 of these sequences originate from previously described channelrhodopsins found in unicellular algae and giant viruses (4), xenorhodopsins (22), and potassium pumps (5), all classified as Type-1 rhodopsins. The previously undescribed sequences (*n* = 278) encompassed both Type-1 and HeRs with 93 of them being of confident taxonomic origin, with 30 from eukaryotes (26 cryptophytes, 1 fungus, 1 ciliate, and 2 unclassified), 34 of nucleocytoplasmic large DNA viruses (NCLDVs), 25 bacterial (12 proteobacteria, 9 actinobacteria, 3 cyanobacteria, and 1 Verrucomicrobiota) and 4 archaeal (3 Halobacteriota and 1 Thermoplasmatota). Of the sequences with uncertain taxonomy (*n* = 185), 110 were likely eukaryotic, 42 were bacterial, and the remaining 33 were unclassified. This distribution implied that AltRs were universally distributed across all domains of life and appeared to be present in both monoderm and diderm bacteria. This contrasted with heliorhodopsins, which are restricted to monoderms (9, 24).

Giant viruses encoded H-type Alt-rhodopsins. Multiple, phylogenetically distinct lineages of rhodopsins (including HeRs) revealed signs of widespread convergent evolution, which was evident from the dispersed distribution of non-R type rhodopsins. For example, H-type AltRs (*n* = 214), which were the most common novel type, did not form a phylogenetically coherent lineage but seemed to have emerged independently in multiple lineages (indicated with yellow arrows in Fig. 1). Additionally, we observed multiple closely related clades of Type-1 rhodopsins accommodating both classical rhodopsins (with Arg82) and H-type AltRs (Fig. 1, indicated by a star). Clustering of AltR encoding contigs and transcript (wherever available), based on shared gene content, also revealed the same major group (*n* = 47, indicated by a star in Fig. 1). However, their taxonomic origin remained unclear. By analyzing flanking genes near these H-type rhodopsins within these contigs, we could identify typically eukaryotic genes, such as the mRNA capping enzyme (mRNAc) and DNA-dependent RNA polymerase subunits (RNAPL) (See Fig. S1). Such genes, however, are also encoded by giant viruses (25). Further scanning of these contigs using ViralRecall (26) convincingly identified them as belonging to the broad class of nucleocytoplasmic large DNA viruses (NCLDVs) with all contigs showing at least one positive hit to known viral proteins. Among contigs ≥5 kb (*n* = 34) a total of 25 encoded at least one hallmark NCLDV marker gene with the longest contig (L969, ~116 Kbp) harboring 5 distinct markers (see also Tables S1 and S2). All AltRs in this cluster had the Arg82 replaced by histidine (H-Type) and display a DxxxK motif in TM7. A more detailed view of gene context variability in the vicinity (5 kbp upstream and downstream) of selected NCLDV H-type AltRs is provided in Fig. S1.

The phylogenetic border between Type-1 and Type-3 rhodopsins. The phylogenetic tree presented in Fig. 1 provided a tantalizing glimpse into the evolution of rhodopsins at large. We sought to examine the sequences closest related to HeRs in more detail to shed light on their presently mysterious evolutionary history as recent works with large numbers of rhodopsin sequences have either considered them as outgroups (27), had insufficient support for the HeR clade (1) or excluded them altogether (6).

Several clades in the tree display high statistical support suggesting they were well-resolved (Fig. 1, *n* = 2199 sequences). Moreover, it also appeared clear that Type-1 rhodopsin diversity far exceeded HeR diversity because HeRs were restricted to a single, albeit highly coherent clade. The topology also indicated that HeRs were a recent innovation and that Type-1 rhodopsins were ancestral. Additionally, HeRs appear to currently be most closely related to two distinct clades of Type-1 rhodopsins (R-type) originating from eukaryotes, one from Dinoflagellates (TM3 motifs ETK, ETS, ETC, or unusual TM7 motif NxxxK) and the other from Colpodellida, (*Chromera velia*, unusual TM3 motif QTQ, TM7 motif DxxxK), both Alveolates. Notably, several HeRs also shared TM3 motifs similar to the ones present in dinoflagellate rhodopsins, e.g., ESV, ETI, or ESL. The dinoflagellate sequences have been described before (4) and characterized as weak pumps of unknown selectivity (6). Their relatedness to HeRs, however, had not been reported. The next closest clade to these are inward pumping Schizorhodopsins (SzRs) found mainly in Asgard archaea and CPR bacteria (8, 28, 29).

Upon closer examination, some of these unusual dinoflagellate rhodopsins (e.g., NCBI accession no. CAE6957589.1; Type-1, R-type, TM3 motif ETK, and TM7 motif NxxxK) appeared to have an extremely long C-terminal region (~500 aa) that returned no clear hits to any known sequence domain (except the N-terminal rhodopsin domain). We performed multiple iterations of structural modeling for this sequence (both entire and in multiple parts) to identify putative domains using Alphafold2 (30). The modeling results

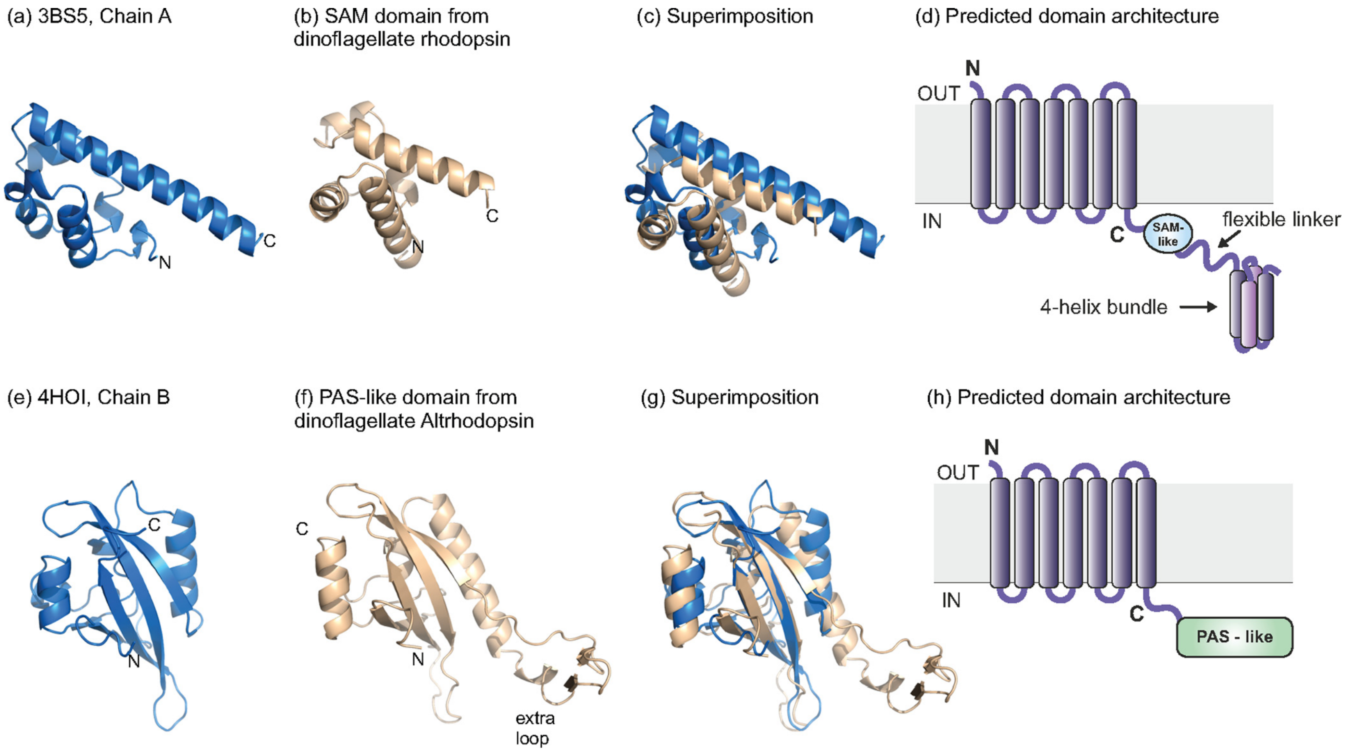
(a) 3BS5, Chain A
N
C
(b) SAM domain from dinoflagellate rhodopsin
N
C
(c) Superimposition
(d) Predicted domain architecture
N
OUT
IN
C
SAM-like
flexible linker
4-helix bundle
(e) 4HOI, Chain B
C
N
(f) PAS-like domain from dinoflagellate Altrhodopsin
C
N
extra loop
(g) Superimposition
(h) Predicted domain architecture
N
OUT
IN
C
PAS - like

FIG 2 Comparison of known and predicted structural domains in C-terminal of selected dinoflagellate Type-1 rhodopsins. (A) Reference structure of a SAM domain, PDB accession no. 3BS5, (B) predicted structure of SAM domain from dinoflagellate rhodopsin sequence NCBI accession no. CAE6957589 (c) superimposition of (A and B), and (D) Predicted domain architecture of the entire protein (E) reference structure of a PAS domain, PDB accession no. 4HOI, (F) predicted structure of the PAS domain from the dinoflagellate Alt-rhodopsin (NCBI accession no. CAE7343182), (G) superimposition of (D and E), and (H) predicted domain architecture of the entire protein.

indicated that the C terminus (intracellular) of the rhodopsin domain was connected to a compact helical domain (modeled with high IDDT scores by AlphaFold2). To identify similar structures, we used this predicted model as a query for structure based-searches with VAST (31). This search identified similar structural motifs in archaeal elongation initiation factor 2 (PDB accession no. 3CW2, domain 2 in alpha subunit) (32), SAM domains (sterile alpha motif) in Yan and Mae proteins that are known to dimerize (33), and SAM domains in proteins CNK and HYP that are known to dimerize as well (PDB accession no. 3BS5) (34). Additionally, distal to the SAM-like domain, there was a flexible linker region (modeled with low support) followed by a second domain composed of multiple helical segments, which also had low modeling support but distant similarities to four-alpha helix bundle domains. The SAM-like domain, the linker, and the four-alpha helix bundle domain were all located in the cytoplasm with no transmembrane helices predicted in this domain (see Fig. 2). The SAM-like domain may likely be useful in the dimerization of such rhodopsins (35) and along with the downstream domains facilitate the transmission of the conformational change in the rhodopsin domain to initiate a signaling cascade. Several dinoflagellate rhodopsins, which showed no or low photocurrents (6) and were coupled to additional domains in the cytoplasm (like the one shown here), were indeed evocative of heliorhodopsins that showed no transport activity and may be coupled to additional domains themselves (9).

To examine the relatedness between heliorhodopsins and dinoflagellate rhodopsins in greater detail, we expanded our search to include additional eukaryotic genomes from dinoflagellates, fungi, etc (from Ensembl and NCBI). Thus, we reconstructed the phylogenetic tree using a subset of the initial sequences while also, including additional sequences closely related to HeRs (Fig. 3). Remarkably, our expanded search led us to identify another unusual Type-1 AltR from the dinoflagellate *Symbiodinium natans* (Q-type, TM3 motif QNL, and unusual TM7 motif TxxxK) that stood out as the phylogenetically closest Type-1 rhodopsin to Type-3 HeRs. The TxxxK

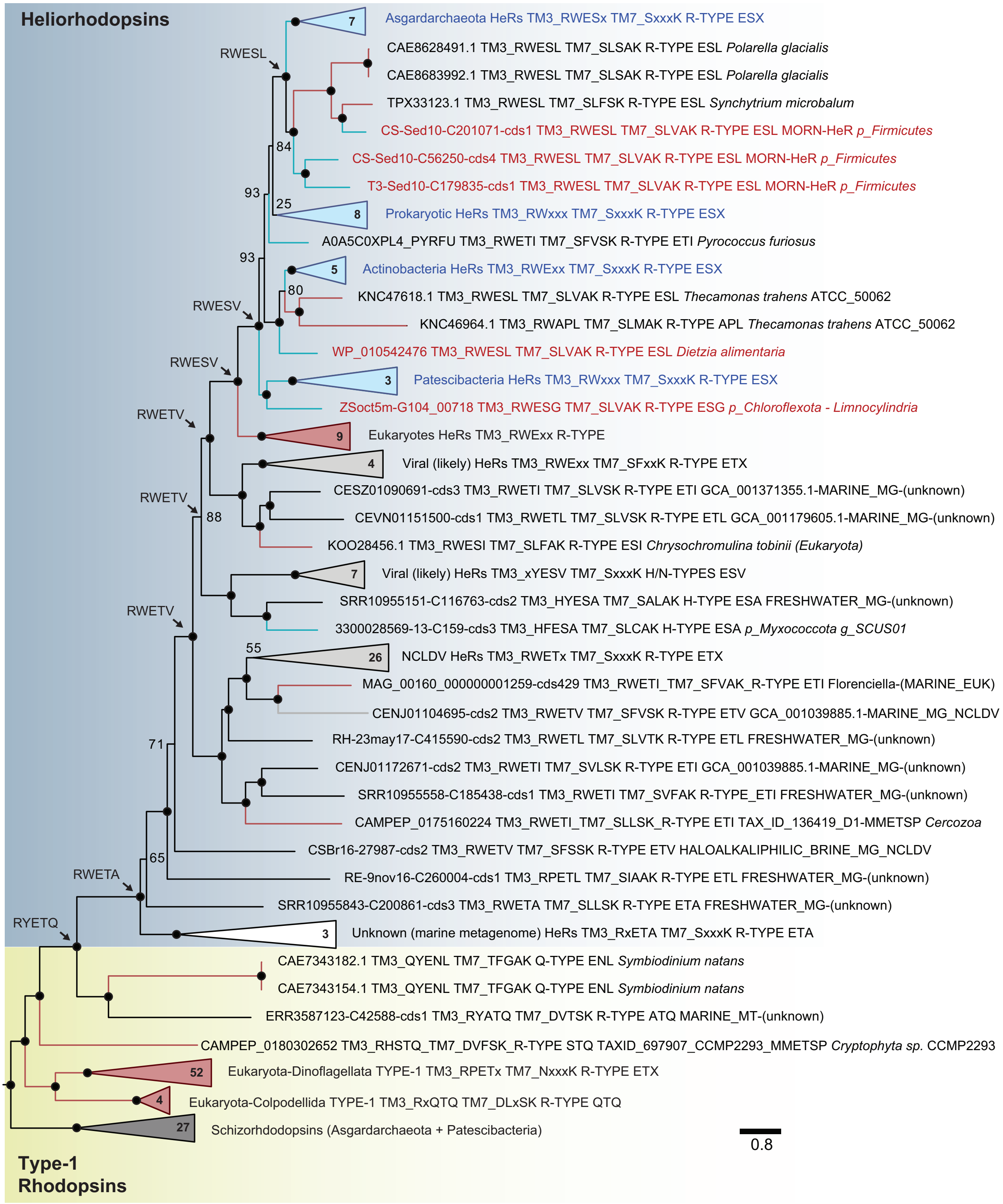
Heliorhodopsins
7 Asgardarchaeota HeRs TM3_RWESx TM7_SxxxK R-TYPE ESX
RWESL
CAE8628491.1 TM3_RWESL TM7_SLSAK R-TYPE ESL Polarella glacialis
CAE8683992.1 TM3_RWESL TM7_SLSAK R-TYPE ESL Polarella glacialis
TPX33123.1 TM3_RWESL TM7_SLFSK R-TYPE ESL Synchytrium microbalum
CS-Sed10-C201071-cds1 TM3_RWESL TM7_SLVAK R-TYPE ESL MORN-HeR p_Firmicutes
84
CS-Sed10-C56250-cds4 TM3_RWESL TM7_SLVAK R-TYPE ESL MORN-HeR p_Firmicutes
93
T3-Sed10-C179835-cds1 TM3_RWESL TM7_SLVAK R-TYPE ESL MORN-HeR p_Firmicutes
25
8 Prokaryotic HeRs TM3_RWxxx TM7_SxxxK R-TYPE ESX
A0A5C0XPL4_PYRFU TM3_RWETI TM7_SFVSK R-TYPE ETI Pyrococcus furiosus
93
5 Actinobacteria HeRs TM3_RWExx TM7_SxxxK R-TYPE ESX
80
KNC47618.1 TM3_RWESL TM7_SLVAK R-TYPE ESL Thecamonas trahens ATCC_50062
RWESV
KNC46964.1 TM3_RWAPL TM7_SLMAK R-TYPE APL Thecamonas trahens ATCC_50062
WP_010542476 TM3_RWESL TM7_SLVAK R-TYPE ESL Dietzia alimentaria
RWESV
3 Patescibacteria HeRs TM3_RWxxx TM7_SxxxK R-TYPE ESX
ZSoct5m-G104_00718 TM3_RWESG TM7_SLVAK R-TYPE ESG p_Chloroflexota - Limnocylindria
RWETV
9 Eukaryotes HeRs TM3_RWExx R-TYPE
4 Viral (likely) HeRs TM3_RWExx TM7_SFxxK R-TYPE ETX
CESZ01090691-cds3 TM3_RWETI TM7_SLVSK R-TYPE ETI GCA_001371355.1-MARINE_MG-(unknown)
RWETV
88
CEVN01151500-cds1 TM3_RWETL TM7_SLVSK R-TYPE ETL GCA_001179605.1-MARINE_MG-(unknown)
KOO28456.1 TM3_RWESI TM7_SLFAK R-TYPE ESI Chrysochromulina tobinii (Eukaryota)
7 Viral (likely) HeRs TM3_xYESV TM7_SxxxK H/N-TYPES ESV
SRR10955151-C116763-cds2 TM3_HYESA TM7_SALAK H-TYPE ESA FRESHWATER_MG-(unknown)
RWETV
3300028569-13-C159-cds3 TM3_HFESA TM7_SLCAK H-TYPE ESA p_Myxococcota g_SCUS01
55
26 NCLDV HeRs TM3_RWETx TM7_SxxxK R-TYPE ETX
MAG_00160_000000001259-cds429 TM3_RWETI_TM7_SFVAK_R-TYPE ETI Florenciella-(MARINE_EUK)
CENJ01104695-cds2 TM3_RWETV TM7_SFVSK R-TYPE ETV GCA_001039885.1-MARINE_MG_NCLDV
71
RH-23may17-C415590-cds2 TM3_RWETL TM7_SLVTK R-TYPE ETL FRESHWATER_MG-(unknown)
CENJ01172671-cds2 TM3_RWETI TM7_SVLSK R-TYPE ETI GCA_001039885.1-MARINE_MG-(unknown)
SRR10955558-C185438-cds1 TM3_RWETI TM7_SVFAK R-TYPE_ETI FRESHWATER_MG-(unknown)
CAMPEP_0175160224 TM3_RWETI_TM7_SLLSK_R-TYPE ETI TAX_ID_136419_D1-MMETSP Cercozoa
CSBr16-27987-cds2 TM3_RWETV TM7_SFSSK R-TYPE ETV HALOALKALIPHILIC_BRINE_MG_NCLDV
65
RWETA
RE-9nov16-C260004-cds1 TM3_RPETL TM7_SIAAK R-TYPE ETL FRESHWATER_MG-(unknown)
SRR10955843-C200861-cds3 TM3_RWETA TM7_SLLSK R-TYPE ETA FRESHWATER_MG-(unknown)
RYETQ
3 Unknown (marine metagenome) HeRs TM3_RxETA TM7_SxxxK R-TYPE ETA
CAE7343182.1 TM3_QYENL TM7_TFGAK Q-TYPE ENL Symbiodinium natans
CAE7343154.1 TM3_QYENL TM7_TFGAK Q-TYPE ENL Symbiodinium natans
ERR3587123-C42588-cds1 TM3_RYATQ TM7_DVTSK R-TYPE ATQ MARINE_MT-(unknown)
CAMPEP_0180302652 TM3_RHSTQ_TM7_DVFSK_R-TYPE STQ TAXID_697907_CCMP2293_MMETSP Cryptophyta sp. CCMP2293
52 Eukaryota-Dinoflagellata TYPE-1 TM3_RPETx TM7_NxxxK R-TYPE ETX
4 Eukaryota-Colpodellida TYPE-1 TM3_RxQTQ TM7_DLxSK R-TYPE QTQ
27 Schizorhdodopsins (Asgardarchaeota + Patescibacteria)
0.8
Type-1
Rhodopsins

FIG 3 Phylogenetic tree depicting Type-1 rhodopsins and Heliorhodopsin evolutionary relationships. Black dots on branches indicate ultrafast bootstrap values between 95 and 100. Numbers of sequences are specified on triangles indicating collapsed clades. Labels of uncollapsed sequences indicate (i) sequence ID, (ii) conserved motifs found within the TM3 region of the rhodopsin, (iii) conserved motifs found within the TM7 region, (iv) type of rhodopsin as defined in Table S1, (v) conserved last 3 amino acids of the TM3 motif, and (vi) taxonomy and/or source of sequence if known. Red, reference sequences, all others are queries; red branches and triangles, eukaryotes; blue, prokaryotes; light gray, sequences retrieved from viruses (or likely viral origin); black lines and empty triangles, sequences of unknown taxonomy; dark-gray triangle, outgroup (Schizorhodopsins). Relevant TM3 motifs generated by ancestral sequences reconstruction are indicated using arrows. MG and MT were used as abbreviations for metagenome and metatranscriptome.

motif in TM7 presented by this sequence was also reminiscent of the SxxxK motif found in HeRs. This sequence (NCBI accession no. CAE7343182.1) was 532 aa long and contained a rhodopsin domain in the first half of the protein. However, the other part of the protein contained no recognizable domains. We further modeled this C-terminal part using AlphaFold2 and obtained a predicted structural model with high iDDT scores (>80). Structure-based searches using VAST suggested close similarities to the PAS domain (See Fig. 2). Additional structure comparisons by TM-align (36) confirmed the same fold (TM-score >0.5; root mean square deviation (rmsd), 3.22). Notably, the PAS domain from this AltR presented an additional loop (Fig. 2). PAS domains are known to be associated with sensory proteins (37), and in several cases, may bind a wide variety of ligands, e.g., heme, hydroxycinnamic acid (38). The PAS domain fold was also found in the LOV domain that was known to bind the flavin mononucleotide (FMN) chromophore acting as a blue light sensor. Remarkably, the VAST search we performed also detected structural similarities to a LOV domain (PDB accession no. 3SW2) (39). The ligand, (if any) that would bind to this PAS domain was unclear but the close similarities to sensory LOV domains reiterate the possible sensory activity of this dinoflagellate rhodopsin. Both rhodopsin sequences that stand as phylogenetic neighbors contained motifs similar to HeRs along with additional domains possibly involved in signaling (or as yet unknown activity), which was consistent with the inference that channeling ions was perhaps not their function.

A eukaryotic origin for heliorhodopsins. We recently argued, based on phylogenetic evidence, that Type-1 rhodopsins were likely the more ancient rhodopsins and HeRs were a comparatively recent innovation (9). Additionally, ancestral reconstruction of Type-1 rhodopsin sequences has suggested that DTE proton pumps most likely represent the ancestral form of Type-1 rhodopsins (27). We performed ancestral reconstruction at multiple nodes in the HeR phylogeny that singled out the ETx motif as ancestral to all HeRs (see Materials and Methods for details). Additional similar ancestral motifs for different clades of HeRs were indicated in Fig. 3. The overall close phylogenetic relatedness and the similarities in the motifs (ESx, ETx) led us to posit that HeRs likely originated from such eukaryotic rhodopsins and were subsequently captured by giant viruses as well. The acquisition of HeRs by prokaryotes and their subsequent diversification in monoderms appeared to have been a later event. This unusual evolutionary trajectory of HeRs (prokaryote to eukaryote being the more common direction) (40), coupled with their fusions/co-occurrences with multiple sensory domains also suggested that HeR was cast in an enabling role as a flexible scaffold allowing innovation in cellular signaling in response to light.

With the diverse array of new rhodopsin sequences and more intermediate sequences sampled, it has become possible to tease out real evolutionary relationships. The diversity of AltRs reiterated the enormous plasticity in the rhodopsin scaffold that continued to surprise us even nearly 5 decades after its discovery (13). Specific activities of such new sequences must be examined individually as even with many structures the entire set of axioms that govern activity remain out of bounds. Even if the activity is understood, the functional role in the organism is unclear. The recent advent of improved methods for protein structure prediction is expected to boost hypothesis generation and structure-aided design. However, given the sheer diversity of both rhodopsins and the organisms that express them, the general lack of molecular tools outside classical model organisms, an enormous effort will still be required in the development of specific assays to finally cut the gordian knot of function for most rhodopsins.

MATERIALS AND METHODS

Sequence data and initial analyses. We used a comprehensive collection of publicly available sequences, including the entire Genome Taxonomy Database (GTDB) (Release 95) (41), Uniprot (42), and ca. 50K Genomes from Earth's Microbiomes (GEM) catalog (43). We also used publicly available metagenomic data from diverse data sets from all over the world, e.g., freshwater metagenomes and metatranscriptomes from multiple European freshwater sites like Rimov Reservoir, Jiricka Pond, Lake Zurich, Lake Constance, Lake Thun (9, 44–47), Lake Tanganyika in Tanzania (48), Lake Baikal in Russia (49), Amadorio and Tous Reservoirs in Spain (50, 51), Lake Mendota in the USA (52), Amazon River (53) in Brazil, the brackish Caspian Sea (54), marine metagenomic data from GEOTRACES (55), metagenomes and metatranscriptomes from TARA Oceans Expeditions (56), metagenomes from brackish sediments (8), metagenomes and metatranscriptomes from haloalkaliphilic brine and sediments (57–59), and eukaryotic culture transcriptomes from the MMETSP database (60). Metagenomic/metatranscriptomic sequences were downloaded and processed with BBMap tools available from https://github.com/BioInfoTools/BBMap/. Briefly, the bbduk.sh script from the BBmap project was used to remove low-quality reads (qtrim = rl trimq = 18), phiX and p-Fosil2 control reads as well as Illumina adapters (k= 21 ref=adapterfile ordered cardinality). Cleaned reads were assembled *de novo* with MEGAHIT v1.2.9 (61) using default parameters with a custom k-mer list: 29, 49, 69, 89, 109, 119, 129, and 149. All sequences in this work were named or retained existing names that allowed tracing them to their original data sets. We also collected reference rhodopsin sequences from a wide variety of previously published sources (1, 4, 5, 8, 22, 62–65).

Rhodopsin identification. Gene prediction on assembled data sets was performed using Prodigal v2.6.3 (66). Candidate rhodopsin sequences were scanned using hmmsearch (67) against existing PFAM models for Type-1 rhodopsins (PF01036), heliorhodopsins (PF18761), and a new HMM built from an alignment of known Type-1 and Type-3 rhodopsins (see FigShare Data at https://figshare.com/s/f2d7b1065930bf350c2f). All sequences were compared to a database of reference rhodopsins to collect homologous sequences using MMseqs2 (68) and multiple alignments were built for each candidate sequence with mafft (–localpair) (69). These alignments were used as input to Polyphobius for the prediction of putative transmembrane helices (70). Only sequences with seven transmembrane helices and a lysine (K) residue in TM7 were retained.

Gene context analysis. Protein coding genes from all collected contigs harboring Alt-rhodopsins (n = 349) were predicted *de novo* by Prodigal v2.6.3 (66) in metagenomic mode (–p meta). Inferred protein sequences were annotated using a local installation of Interproscan (71) and by scanning them against the Protein Families (PFAM v.31) database with the Perl script pfam_scan.pl (available from ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools). Annotation was also performed by scanning proteins with hmmsearch (67) against the COGs (clusters of orthologous groups) (72) and TIGRFAMs (73) HMM databases (E value ≤ 1e-3). BlastKOALA (74) was used to assign KO numbers to predicted orthologous proteins. To facilitate gene-context analysis for Alt-rhodopsins, the collection of contigs was clustered based on shared homologous proteins (i.e., requiring a minimum of 2 shared genes between any 2 members). For this purpose, all predicted proteins were clustered together using MMseqs2 (68) in easy-cluster mode (–cluster-mode 1 –c 0.5 –s 7.5) to identify homologs. Protein clustering information was further used to group contigs based on shared gene content. Contig clusters were plotted using Gcluster (75) with default parameters and showing consensus gene annotation generated as previously described. Manual curation of plotted clusters involved removal of contigs with less than 5 genes and collapsing of nearly identical ones while recording their number.

Taxonomic classification of contigs. Taxonomy was assigned to protein-coding genes we previously predicted within rhodopsin-encoding contigs by screening them with MMseqs2 (68) ("search" option, default parameters) against the annotated proteomes from the Genome Taxonomy Database (GTDB; release 95) (41). We followed a very conservative approach to pinpoint the taxonomic origin of these contigs, considering only those of at least 5 kb in length, and a minimum of 60% of genes giving best hits to the same phylum. Shorter contigs were retained as unclassified.

Phylogenetic trees of rhodopsins. The tree shown in Fig. 1 contains 2199 sequences of which 694 were reference rhodopsin sequences and the remaining 1505 were identified from our scan. These 1505 sequences were chosen as representatives of the total 6478 rhodopsin sequences identified in our scan. We chose to cluster all identified sequences at a 90% level of identity using MMseqs2 (easy-cluster) to include only representative ones in the tree. Additionally, all rhodopsin sequences >350 aa were excluded. Multiple alignments were performed using mafft (69) and a maximum likelihood tree was made using iqtree2 (76) with automatic model selection performed by ModelFinder (77), and 1000 iterations of ultrafast bootstrapping with 1000 rounds of SH-aLRT testing (-alrt 1000 -B 1000) (78).

The tree shown in Fig. 3 contains 834 sequences of which 226 were reference rhodopsins and 608 were identified from our scan. This reduced set of rhodopsins was obtained from the initial collection of 2201 sequences following clustering at 70% identity using MMseqs2. Sequences were aligned using PASTA (79) (default parameters) and tree construction performed by iqtree2 (v 2.1.2) with the same parameters used for the original tree. Phylogenetic tree pruning meant to highlight the Type-1/HeR split and manual annotations of the subtree were carried out in FigTree v 1.4.3 (https://github.com/rambaut/figtree/).

Domain predictions and structural analyses. Sequence-based domain predictions were carried out using Pfam (80), the Conserved Domain Database (81), HMMER (82), and HHPred (83). Structure prediction and domain definitions for selected sequences were performed using Alphafold2 (30) provided via ColabFold (84). Protein structures were visualized using ChimeraX (85). Structure-based searches were carried out using VAST (31). Transmembrane helix predictions were performed using Polyphobius (70) and supplemented wherever necessary with additional predictions from TOPCONS (86) and Phobius (87).

Identification of NCLDV contigs. Rhodopsin-encoding contigs of at least 5 kb (n = 485) were scanned with ViralRecall (26) to identify signatures of putative nucleocytoplasmic large DNA viruses (NCLDVs). The minimum number of viral hits to be reported by ViralRecall was reduced to 1 (-g 1) from the default value of 4 hits. We further classified contigs according to ViralRecall scores into four categories: NCLDV_high (≥5), NCLDV_medium (<5, ≥2), NCLDV_low (<2, >0), and non_NCLDV (<0) (results shown in Tables S1 and S2).

Ancestral sequence reconstruction. The evolutionary history of TM3 amino acid motifs from collected rhodopsins was inferred by ancestral sequence reconstruction (ASR) (88). In brief, the rhodopsin alignment generated for the reduced rhodopsin tree (n = 834 sequences; available at https://figshare

.com/s/f2d7b1065930bf350c2f) was used as input in iqtree2 (v 2.1.2) with the –asr option specified and best model previously chosen according to the Bayesian information criterion (BIC) (–perturb 0.2 –nstop 500 -B 1000 -m LG+I+G4 –alrt 1000 -asr). Results were filtered to keep only ancestral sequence positions with a probability cutoff ≥0.4. TM3 motifs were identified in final ASR sequences by comparison to references and were indicated for 8 nodes in the phylogenetic tree shown in Fig. 3.

Data availability. All sequences used in this work, including reference sequences and derived data such as alignments and phylogenetic trees, have been deposited at FigShare at https://figshare.com/s/f2d7b1065930bf350c2f and are publicly available for download as of the date of publication.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, EPS file, 2.3 MB.

TABLE S1, XLSX file, 0.4 MB.

TABLE S2, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

P.-A.B. and R.G. were supported by the research grant 20-12496X (Grant Agency of the Czech Republic). V.S.K. was supported by the research grant 116/2019/P (Grant Agency of the University of South Bohemia in České Budějovice, 2019 to 2021), 19-23261S (Grant Agency of the Czech Republic), and 21-21990S (Grant Agency of the Czech Republic). A.-S.A. was supported by the Ambizione grant PZ00P3_193240 (Swiss National Science Foundation). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

R.G. conceived the work, P.-A.B., V.S.K., A.-S.A., and R.G. performed the data analyses, P.-A.B., R.G., and A.-S.A. wrote the manuscript. All authors read and approved the final manuscript.

We declare no competing interests.

REFERENCES

1. Rozenberg A, Inoue K, Kandori H, Béjà O. 2021. Microbial Rhodopsins: the last two decades. Annu Rev Microbiol 75:427–447. https://doi.org/10.1146/annurev-micro-031721-020452.
2. Gushchin I, Gordeliy V. 2018. Microbial Rhodopsins. Subcell Biochem 87:19–56. https://doi.org/10.1007/978-981-10-7757-9_2.
3. Kandori H. 2021. History and perspectives of ion-transporting rhodopsins. Adv Exp Med Biol 1293:3–19. https://doi.org/10.1007/978-981-15-8763-4_1.
4. Rozenberg A, Oppermann J, Wietek J, Fernandez Lahore RG, Sandaa R-A, Bratbak G, Hegemann P, Béjà O. 2020. Lateral gene transfer of anion-conducting channelrhodopsins between green algae and giant viruses. Curr Biol 30:4910–4920.e5. https://doi.org/10.1016/j.cub.2020.09.056.
5. Govorunova EG, Gou Y, Sineshchekov OA, Li H, Wang Y, Brown LS, Xue M, Spudich JL. 2021. Kalium rhodopsins: natural light-gated potassium channels. bioRxiv. https://doi.org/10.1101/2021.09.17.460684.
6. Govorunova EG, Sineshchekov OA, Li H, Wang Y, Brown LS, Palmateer A, Melkonian M, Cheng S, Carpenter E, Patterson J, Wong GK-S, Spudich JL. 2021. Cation and anion channelrhodopsins: sequence motifs and taxonomic distribution. mBio 12:e0165621. https://doi.org/10.1128/mBio.01656-21.
7. Govorunova EG, Sineshchekov OA, Li H, Wang Y, Brown LS, Spudich JL. 2020. RubyACRs, nonalgal anion channelrhodopsins with highly red-shifted absorption. Proc Natl Acad Sci U S A 117:22833–22840. https://doi.org/10.1073/pnas.2005981117.
8. Bulzu P-A, Andrei A-Ş, Salcher MM, Mehrshad M, Inoue K, Kandori H, Beja O, Ghai R, Banciu HL. 2019. Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. Nat Microbiol 4:1129–1137. https://doi.org/10.1038/s41564-019-0404-y.
9. Bulzu P-A, Kavagutti VS, Chiriac M-C, Vavourakis CD, Inoue K, Kandori H, Andrei A-S, Ghai R. 2021. Heliorhodopsin evolution is driven by photosensory promiscuity in monoderms. mSphere 6:e0066121. https://doi.org/10.1128/mSphere.00661-21.
10. Ernst OP, Lodowski DT, Elstner M, Hegemann P, Brown LS, Kandori H. 2014. Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. Chem Rev 114:126–163. https://doi.org/10.1021/cr4003769.
11. Pushkarev A, Inoue K, Larom S, Flores-Uribe J, Singh M, Konno M, Tomida S, Ito S, Nakamura R, Tsunoda SP, Philosof A, Sharon I, Yutin N, Koonin EV, Kandori H, Béjà O. 2018. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. Nature 558:595–599. https://doi.org/10.1038/s41586-018-0225-9.
12. Nagata T, Inoue K. 2021. Rhodopsins at a glance. J Cell Sci 134:jcs258989. https://doi.org/10.1242/jcs.258989.
13. Oesterhelt D, Stoeckenius W. 1971. Rhodopsin-like protein from the purple membrane of Halobacterium halobium. Nat New Biol 233:149–152. https://doi.org/10.1038/newbio233149a0.
14. Otto H, Marti T, Holz M, Mogi T, Stern LJ, Engel F, Khorana HG, Heyn MP. 1990. Substitution of amino acids Asp-85, Asp-212, and Arg-82 in bacteriorhodopsin affects the proton release phase of the pump and the pK of the Schiff base. Proc Natl Acad Sci U S A 87:1018–1022. https://doi.org/10.1073/pnas.87.3.1018.
15. Imasheva ES, Balashov SP, Ebrey TG, Chen N, Crouch RK, Menick DR. 1999. Two groups control light-induced schiff base deprotonation and the proton affinity of Asp85 in the Arg82His mutant of bacteriorhodopsin. Biophys J 77:2750–2763. https://doi.org/10.1016/S0006-3495(99)77108-0.
16. Hutson MS, Alexiev U, Shilov SV, Wise KJ, Braiman MS. 2000. Evidence for a perturbation of arginine-82 in the bacteriorhodopsin photocycle from time-resolved infrared spectra. Biochemistry 39:13189–13200. https://doi.org/10.1021/bi000426q.
17. Brown LS, Bonet L, Needleman R, Lanyi JK. 1993. Estimated acid dissociation constants of the Schiff base, Asp-85, and Arg-82 during the bacteriorhodopsin photocycle. Biophys J 65:124–130. https://doi.org/10.1016/S0006-3495(93)81064-6.
18. Balashov SP, Govindjee R, Imasheva ES, Misra S, Ebrey TG, Feng Y, Crouch RK, Menick DR. 1995. The two pKa's of aspartate-85 and control of thermal isomerization and proton release in the arginine-82 to lysine mutant of bacteriorhodopsin. Biochemistry 34:8820–8834. https://doi.org/10.1021/bi00027a034.
19. Balashov SP, Imasheva ES, Govindjee R, Ebrey TG. 1996. Titration of aspartate-85 in bacteriorhodopsin: what it says about chromophore isomerization and proton release. Biophys J 70:473–481. https://doi.org/10.1016/S0006-3495(96)79591-7.
20. Govindjee R, Misra S, Balashov SP, Ebrey TG, Crouch RK, Menick DR. 1996. Arginine-82 regulates the pKa of the group responsible for the light-

driven proton release in bacteriorhodopsin. Biophys J 71:1011–1023. https://doi.org/10.1016/S0006-3495(96)79302-5.
21. Stern LJ, Khorana HG. 1989. Structure-function studies on bacteriorhodopsin. X. Individual substitutions of arginine residues by glutamine affect chromophore formation, photocycle, and proton translocation. J Biol Chem 264:14202–14208. https://doi.org/10.1016/S0021-9258(18)71663-3.
22. Ugalde JA, Podell S, Narasingarao P, Allen EE. 2011. Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria. Biol Direct 6:52. https://doi.org/10.1186/1745-6150-6-52.
23. Shevchenko V, Mager T, Kovalev K, Polovinkin V, Alekseev A, Juettner J, Chizhov I, Bamann C, Vavourakis C, Ghai R, Gushchin I, Borshchevskiy V, Rogachev A, Melnikov I, Popov A, Balandin T, Rodriguez-Valera F, Manstein DJ, Bueldt G, Bamberg E, Gordeliy V. 2017. Inward H+ pump xenorhodopsin: mechanism and alternative optogenetic approach. Sci Adv 3:e1603187. https://doi.org/10.1126/sciadv.1603187.
24. Flores-Uribe J, Hevroni G, Ghai R. 2019. Heliorhodopsins are absent in diderm (Gram-negative) bacteria: some thoughts and possible implications for activity. Environmental Microbiology Reports.
25. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. PLoS Biol 19:e3001430. https://doi.org/10.1371/journal.pbio.3001430.
26. Aylward FO, Moniruzzaman M. 2021. ViralRecall-a flexible command-line tool for the detection of giant virus signatures in 'omic data. Viruses 13:150. https://doi.org/10.3390/v13020150.
27. Sephus CD, Fer E, Garcia AK, Adam ZR, Schwieterman EW, Kaçar B. 2021. Functional divergence and spectral tuning of microbial rhodopsins from an ancestral proton pump. Mol Bio Evol 39:msac100. https://doi.org/10.1093/molbev/msac100.
28. Inoue K, Tsunoda SP, Singh M, Tomida S, Hososhima S, Konno M, Nakamura R, Watanabe H, Bulzu P-A, Banciu HL, Andrei A-Ş, Uchihashi T, Ghai R, Béjà O, Kandori H. 2020. Schizorhodopsins: a family of rhodopsins from Asgard archaea that function as light-driven inward H+ pumps. Sci Adv 6:eaaz2441. https://doi.org/10.1126/sciadv.aaz2441.
29. Wong HL, MacLeod FI, White RA, 3rd, Visscher PT, Burns BP. 2020. Microbial dark matter filling the niche in hypersaline microbial mats. Microbiome 8:135. https://doi.org/10.1186/s40168-020-00910-0.
30. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2.
31. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. 2014. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res 42:D297–D303. https://doi.org/10.1093/nar/gkt1208.
32. Stolboushkina E, Nikonov S, Nikulin A, Bläsi U, Manstein DJ, Fedorov R, Garber M, Nikonov O. 2008. Crystal structure of the intact archaeal translation initiation factor 2 demonstrates very high conformational flexibility in the alpha- and beta-subunits. J Mol Biol 382:680–691. https://doi.org/10.1016/j.jmb.2008.07.039.
33. Qiao F, Song H, Kim CA, Sawaya MR, Hunter JB, Gingery M, Rebay I, Courey AJ, Bowie JU. 2004. Derepression by depolymerization; structural insights into the regulation of Yan by Mae. Cell 118:163–173. https://doi.org/10.1016/j.cell.2004.07.010.
34. Rajakulendran T, Sahmi M, Kurinov I, Tyers M, Therrien M, Sicheri F. 2008. CNK and HYP form a discrete dimer by their SAM domains to mediate RAF kinase signaling. Proc Natl Acad Sci U S A 105:2836–2841. https://doi.org/10.1073/pnas.0709705105.
35. Peterson AJ, Kyba M, Bornemann D, Morgan K, Brock HW, Simon J. 1997. A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. Mol Cell Biol 17:6683–6692. https://doi.org/10.1128/MCB.17.11.6683.
36. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309. https://doi.org/10.1093/nar/gki524.
37. Aravind L, Iyer LM, Anantharaman V. 2010. Natural history of sensor domains in bacterial signaling systems, p 1–38. *In* Stephen Spiro RD (ed), Sensory Mechanisms in Bacteria: molecular Aspects of Signal Recognition. Caister Academic Press Norfolk, UK.
38. Möglich A, Ayers RA, Moffat K. 2009. Structure and signaling mechanism of Per-ARNT-Sim domains. Structure 17:1282–1294. https://doi.org/10.1016/j.str.2009.08.011.
39. Circolone F, Granzin J, Jentzsch K, Drepper T, Jaeger K-E, Willbold D, Krauss U, Batra-Safferling R. 2012. Structural basis for the slow dark recovery of a full-length LOV protein from Pseudomonas putida. J Mol Biol 417:362–374. https://doi.org/10.1016/j.jmb.2012.01.056.
40. Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–618. https://doi.org/10.1038/nrg2386.
41. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38:1079–1086. https://doi.org/10.1038/s41587-020-0501-8.
42. UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100.
43. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe-Fadrosh EA, IMG/M Data Consortium. 2021. A genomic catalog of Earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-00769-4.
44. Kavagutti VS, Andrei A-Ş, Mehrshad M, Salcher MM, Ghai R. 2019. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome 7:135. https://doi.org/10.1186/s40168-019-0752-0.
45. Mujakić I, Andrei A-Ş, Shabarova T, Fecskeová LK, Salcher MM, Piwosz K, Ghai R, Koblížek M. 2021. Common presence of phototrophic gemmatimonadota in temperate freshwater lakes. mSystems 6:e01241-20. https://doi.org/10.1128/mSystems.01241-20.
46. Andrei A-Ş, Salcher MM, Mehrshad M, Rychtecký P, Znachor P, Ghai R. 2019. Niche-directed evolution modulates genome architecture in freshwater Planctomycetes. ISME J 13:1056–1071. https://doi.org/10.1038/s41396-018-0332-5.
47. Mehrshad M, Salcher MM, Okazaki Y, Nakano S-I, Šimek K, Andrei A-S, Ghai R. 2018. Hidden in plain sight-highly abundant and diverse planktonic freshwater Chloroflexi. Microbiome 6:176. https://doi.org/10.1186/s40168-018-0563-8.
48. Tran PQ, Bachand SC, McIntyre PB, Kraemer BM, Vadeboncoeur Y, Kimirei IA, Tamatamah R, McMahon KD, Anantharaman K. 2021. Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. ISME J 15:1971–1986. https://doi.org/10.1038/s41396-021-00898-x.
49. Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko MV, Ivanov VG, Ghai R, Rodriguez-Valera F. 2019. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. Limnol Oceanogr 94:fiy163. https://doi.org/10.1002/lno.11401.
50. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. 2014. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. Mol Ecol 23:6073–6090. https://doi.org/10.1111/mec.12985.
51. Cabello-Yeves PJ, Haro-Moreno JM, Martin-Cuadrado A-B, Ghai R, Picazo A, Camacho A, Rodriguez-Valera F. 2017. Novel Synechococcus genomes reconstructed from freshwater reservoirs. Front Microbiol 8:1151. https://doi.org/10.3389/fmicb.2017.01151.
52. Linz AM, He S, Stevens SLR, Anantharaman K, Rohwer RR, Malmstrom RR, Bertilsson S, McMahon KD. 2018. Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. PeerJ 6:e6075. https://doi.org/10.7717/peerj.6075.
53. Satinsky BM, Fortunato CS, Doherty M, Smith CB, Sharma S, Ward ND, Krusche AV, Yager PL, Richey JE, Moran MA, Crump BC. 2015. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. Microbiome 3:39. https://doi.org/10.1186/s40168-015-0099-0.
54. Mehrshad M, Amoozegar MA, Ghai R, Shahzadeh Fazeli SA, Rodriguez-Valera F. 2016. Genome reconstruction from metagenomic data sets reveals novel microbes in the brackish waters of the Caspian Sea. Appl Environ Microbiol 82:1599–1612. https://doi.org/10.1128/AEM.03381-15.
55. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. 2018. Marine microbial metagenomes sampled across space and time. Sci Data 5:180176. https://doi.org/10.1038/sdata.2018.176.

56. Sunagawa S, Coelho LP, Chaffron S, Kultima JR. 2015. Structure and function of the global ocean microbiome. Science 348:6237. https://doi.org/10.1126/science.1261359.
57. Vavourakis CD, Ghai R, Rodriguez-Valera F, Sorokin DY, Tringe SG, Hugenholtz P, Muyzer G. 2016. Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline Soda Lake brines. Front Microbiol 7:211. https://doi.org/10.3389/fmicb.2016.00211.
58. Vavourakis CD, Andrei A-S, Mehrshad M, Ghai R, Sorokin DY, Muyzer G. 2018. A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. Microbiome 6:168. https://doi.org/10.1186/s40168-018-0548-7.
59. Vavourakis CD, Mehrshad M, Balkema C, van Hall R, Andrei A-Ş, Ghai R, Sorokin DY, Muyzer G. 2019. Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. BMC Biol 17:69. https://doi.org/10.1186/s12915-019-0688-7.
60. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:e1001889. https://doi.org/10.1371/journal.pbio.1001889.
61. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.
62. Oppermann J, Fischer P, Silapetere A, Liepe B, Rodriguez-Rozada S, Flores-Uribe J, Peter E, Keidel A, Vierock J, Kaufmann J, Broser M, Luck M, Bartl F, Hildebrandt P, Wiegert JS, Béjà O, Hegemann P, Wietek J. 2019. MerMAIDs: a family of metagenomically discovered marine anion-conducting and intensely desensitizing channelrhodopsins. Nat Commun 10:3315. https://doi.org/10.1038/s41467-019-11322-6.
63. Inoue K, Ono H, Abe-Yoshizumi R, Yoshizawa S, Ito H, Kogure K, Kandori H. 2013. A light-driven sodium ion pump in marine bacteria. Nat Commun 4:1–10. https://doi.org/10.1038/ncomms2689.
64. Kataoka C, Sugimoto T, Shigemura S, Katayama K, Tsunoda SP, Inoue K, Béjà O, Kandori H. 2021. TAT Rhodopsin is an ultraviolet-dependent environmental pH sensor. Biochemistry 60:899–907. https://doi.org/10.1021/acs.biochem.0c00951.
65. Yutin N, Koonin EV. 2012. Proteorhodopsin genes in giant viruses. Biol Direct 7:34. https://doi.org/10.1186/1745-6150-7-34.
66. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
67. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
68. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988.
69. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
70. Käll L, Krogh A, Sonnhammer ELL. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21 Suppl 1:i251–i257. https://doi.org/10.1093/bioinformatics/bti1014.
71. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49:D344–D354. https://doi.org/10.1093/nar/gkaa977.
72. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–D269. https://doi.org/10.1093/nar/gku1223.
73. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31:371–373. https://doi.org/10.1093/nar/gkg128.
74. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726–731. https://doi.org/10.1016/j.jmb.2015.11.006.
75. Li X, Chen F, Chen Y. 2020. Gcluster: a simple-to-use tool for visualizing and comparing genome contexts for numerous genomes. Bioinformatics 36:3871–3873. https://doi.org/10.1093/bioinformatics/btaa212.
76. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015.
77. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589. https://doi.org/10.1038/nmeth.4285.
78. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518–522. https://doi.org/10.1093/molbev/msx281.
79. Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J Comput Biol 22:377–386. https://doi.org/10.1089/cmb.2014.0156.
80. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995.
81. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2020. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265–D268. https://doi.org/10.1093/nar/gkz991.
82. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448.
83. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.
84. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2021. ColabFold - Making protein folding accessible to all. Nat Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1.
85. Pettersen EF, Goddard TD, Huang CC. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084.
86. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res 43:W401–W407. https://doi.org/10.1093/nar/gkv485.
87. Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016.
88. Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet 5:366–375. https://doi.org/10.1038/nrg1324.