AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE

Heliorhodopsin Evolution Is Driven by Photosensory Promiscuity in Monoderms

Paul-Adrian Bulzu,[a] Vinicius Silva Kavagutti,[a,b] Maria-Cecilia Chiriac,[a] Charlotte D. Vavourakis,[c,d] Keiichi Inoue,[e] Hideki Kandori,[f,g] Adrian-Stefan Andrei,[h] Rohit Ghai[a]

[a]Department of Aquatic Microbial Ecology, Institute of Hydrobiology, Biology Centre of the Academy of Sciences of the Czech Republic, České Budějovice, Czech Republic
[b]Department of Ecosystem Biology, Faculty of Science, University of South Bohemia, Branišovská, České Budějovice, Czech Republic
[c]Research Institute for Biomedical Aging Research, University of Innsbruck, Innsbruck, Austria
[d]European Translational Oncology Prevention and Screening (EUTOPS) Institute, Tirol, Austria
[e]The Institute for Solid State Physics, The University of Tokyo, Kashiwa, Japan
[f]Department of Life Science and Applied Chemistry, Nagoya Institute of Technology, Showa, Nagoya, Japan
[g]OptoBioTechnology Research Center, Nagoya Institute of Technology, Showa, Nagoya, Japan
[h]Limnological Station, Department of Plant and Microbial Biology, University of Zurich, Kilchberg, Switzerland

ABSTRACT Rhodopsins are light-activated proteins displaying an enormous versatility of function as cation/anion pumps or sensing environmental stimuli and are widely distributed across all domains of life. Even with wide sequence divergence and uncertain evolutionary linkages between microbial (type 1) and animal (type 2) rhodopsins, the membrane orientation of the core structural scaffold of both was presumed universal. This was recently amended through the discovery of heliorhodopsins (HeRs; type 3), that, in contrast to known rhodopsins, display an inverted membrane topology and yet retain similarities in sequence, structure, and the light-activated response. While no ion-pumping activity has been demonstrated for HeRs and multiple crystal structures are available, fundamental questions regarding their cellular and ecological function or even their taxonomic distribution remain unresolved. Here, we investigated HeR function and distribution using genomic/metagenomic data with protein domain fusions, contextual genomic information, and gene coexpression analysis with strand-specific metatranscriptomics. We bring to resolution the debated monoderm/diderm occurrence patterns and show that HeRs are restricted to monoderms. Moreover, we provide compelling evidence that HeRs are a novel type of sensory rhodopsins linked to histidine kinases and other two-component system genes across phyla. In addition, we also describe two novel putative signal-transducing domains fused to some HeRs. We posit that HeRs likely function as generalized light-dependent switches involved in the mitigation of light-induced oxidative stress and metabolic circuitry regulation. Their role as sensory rhodopsins is corroborated by their photocycle dynamics and their presence/function in monoderms is likely connected to the higher sensitivity of these organisms to light-induced damage.

IMPORTANCE Heliorhodopsins are enigmatic, novel rhodopsins with a membrane orientation that is opposite to all known rhodopsins. However, their cellular and ecological functions are unknown, and even their taxonomic distribution remains a subject of debate. We provide evidence that HeRs are a novel type of sensory rhodopsins linked to histidine kinases and other two-component system genes across phyla boundaries. In support of this, we also identify two novel putative signal transducing domains in HeRs that are fused with them. We also observe linkages of HeRs to genes involved in mitigation of light-induced oxidative stress and increased carbon and nitrogen metabolism. Finally, we synthesize these findings into a framework that connects HeRs with the cellular response to light in monoderms,

Editor Steven J. Hallam, University of British Columbia

Address correspondence to Rohit Ghai, ghai.rohit@gmail.com.

The authors declare no conflict of interest.

Received 26 July 2021
Accepted 9 November 2021
Published 24 November 2021

activating light-induced oxidative stress defenses along with carbon/nitrogen metabolic circuitries. These findings are consistent with the evolutionary, taxonomic, structural, and genomic data available so far.

KEYWORDS heliorhodopsin, rhodopsins, metagenomics, oxidative stress

The ability to harness the Sun's electromagnetic radiation by channeling it into high-energy phosphate bonds empowered microorganisms to tap into an inexpensive and inexhaustible source of energy. Life's billion-year history of metabolic innovations led to the emergence of only two biological complexes capable of harvesting light: one based on rhodopsins and the other on (bacterio)chlorophyll. Rhodopsins encompass the most diverse and abundant photoactive proteins on Earth and were until recently canonically split between type 1 (microbial rhodopsins) and type 2 (animal rhodopsins) families. Type 1 and 2 rhodopsins families share a similar topological conformation and little or no sequence similarity among each other. Despite dissimilarities in function, structure, and phylogeny, type 1 and 2 rhodopsins have a similar membrane orientation, with their N terminus being situated in the extracellular space. Recently identified during a functional metagenomics screen and characterized by low sequence similarity compared to type 1 rhodopsins, heliorhodopsins (HeRs) have attracted increasing research interest due to their peculiar membrane orientation (i.e., the N terminus in the cytoplasm and the C terminus in the extracellular space) (1), unusual protein structure (2), and controversial taxonomic distribution (3). While electrophysiological (1), physicochemical (4), and structural (2, 5) studies have achieved great progress in elucidating a series of characteristics ranging from photocycle length and lack of ion-pumping activity to detailed protein organization, they provide little information regarding the biological function of HeRs. Moreover, polarized opinions regarding the putative ecological role and taxonomic distribution of HeR-encoding organisms (2, 3) call for the use of novel approaches in establishing HeR functionality. The present study draws its essence from the tenet that functionally linked genes within prokaryotes are coregulated and thus occur close to each other (6, 7). Within this framework, the functions of uncharacterized genes (i.e., HeRs) can be inferred from their genomic surroundings. Here, we couple HeR distributional patterns with contextual genomic information involving protein domain fusions and operon organization and gene expression data to shed light on HeR functionality.

RESULTS

In order to shed light on the distribution and functional role of HeRs in nature, we conducted a comprehensive survey of genomes and metagenomes enabling us to phylogenetically constrain HeR distribution patterns. Once these patterns were constrained, we evaluated genomic and metagenomic sequences for domain fusions and gene context information in order to identify potential effectors of HeR signaling, identifying potential effector domains and several operons with the potential to couple light sensing to metabolic responses. Finally, to better evaluate the potential for cotranscriptional responses identified *in silico*, we conducted strand-specific metatranscriptomics in a freshwater ecosystem to identify expressed HeRs linked to functional genes.

Taxonomic distribution. Previous assessments of taxonomic distribution of HeRs reported conflicting data regarding their presence in monoderm (3) and diderm (2) prokaryotes. In order to accurately map HeR taxonomic distribution, we used the GTDB database (release 89), since it contains a wide-range of high-quality genomes derived from isolated strains and environmental metagenome-assembled genomes, classified within a robust phylogenomic framework (8). By scanning 24,706 genomes, we identified 469 *bona fide* HeR sequences (topology: C-terminal inside and N-terminal outside, seven transmembrane helices and a SxxxK motif in helix 7; see Table S1 [https://doi.org/10.6084/m9.figshare.13286486]) spanned across 17 phyla (out of 151; see Table S2 [https://doi.org/10.6084/m9.figshare.13286486]). In order to assign HeR-containing genomes to either monoderm or diderm categories, we employed a set of 27 manually curated protein

domain markers (see Table S13 [https://doi.org/10.6084/m9.figshare.13286486]) that are expected to be restricted to organisms possessing double-membrane cellular envelopes (i.e., diderms) (9). While most analyses were expected to be influenced by various levels of genome completeness, we found that a conservative criterion of presence of at least 10 marker domains singled out all diderm lineages (i.e., *Negativicutes*, *Halanaerobiales*, and *Limnochondria*) (9, 10) within the larger monoderm phylum Firmicutes, apart from correctly identifying other well-known diderms. Except for three genomes (one each belonging to *Myxococcota*, *Spirochaetota*, and *Dictyoglomota* phyla), all other HeR occurrences were restricted to monoderms (see Table S2 [https://doi.org/10.6084/m9.figshare.13286486]). Examination of the HeR-encoding *Myxococcota* contig by querying its predicted proteins against the RefSeq and GTDB databases revealed it to be an actinobacterial contaminant. The *Spirochaeta* genome was incomplete (60% estimated completeness) and only encoded two outer membrane marker genes, making any inferences regarding its affiliation to monoderm or diderm bacteria impossible. However, we could not rule out that this genome could belong to a member lacking lipopolysaccharides (LPS) (9). The *Dictyoglomota* genome belongs to an isolate, and despite its high completeness, it encodes only five markers. Combined with the notion that *Dictyoglomota* are known to have atypical membrane architectures (11), the presence of only five marker points toward the absence of a classical diderm cell envelope. Apart from these exceptions, all other HeR-encoding genomes are monoderm and, at least within this collection, we found no strong evidence of HeRs being present in any organism that is conclusively diderm. We also identified HeRs in several assembled metagenomes and metatranscriptomes (see Materials and Methods). For improved resolution of taxonomic origin, we considered only contigs of at least 5 kb in length (n = 1,340 from metagenomes and n = 4 from metatranscriptomes). Following a strict approach to taxonomy assignment (i.e., at least 60% genes giving best-hits to the same phylum and not just majority-rule), we could designate a phylum for most HeRs. Without any exception, we found that all the contigs that received robust taxonomic classification (n = 1,319) belonged to known monoderm phyla (see Table S3 [https://doi.org/10.6084/m9.figshare.13286486]).

Domain fusions. Domain fusions with rhodopsins are recently providing novel insights into the diverse functional couplings that enhance the utility of a light sensor, e.g., the case of a phosphodiesterase domain fused with a type 1 rhodopsin (12). As far as we are aware, no domain fusions have yet been described for HeRs. In our search for such domain fusions that may shed light on HeR functionality, the MORN repeat (Membrane Occupation and Recognition Nexus; PF02493) was found in multiple copies (typically 3) at the cytoplasmic N terminus of some HeRs (n = 36). A tentative three-dimensional (3D) model for a representative MORN-HeR could be generated and is shown in Fig. 1A.

These MORN-HeR sequences were phylogenetically restricted to two environmental branches of metagenome-assembled genomes (MAGs) recovered from haloalkaline sediments that affiliate to the family *Syntrophomonadaceae* (phylum *Firmicutes*) (13–15) (see Fig. S1 [https://doi.org/10.6084/m9.figshare.13286486]). The prototypic MORN repeat, consisting of 14 amino acids with the consensus sequence YEGEWxNGKxHGYG, was first described in 2000 (16) from junctophilins present in skeletal muscle and later recognized to be ubiquitous in both eukaryotes and prokaryotes (17). This conserved signature can be seen in the alignment of MORN-repeats fused to HeRs (see Fig. S2 [https://doi.org/10.6084/m9.figshare.13286486]). MORN-repeats have been shown to bind to phospholipids (18, 19), promoting stable interactions with plasma membranes (16) and also function as protein-protein interaction modules involved in di- and oligomerization (20). They are expected to be intracellular and provide a large putative interaction surface (either with other MORN-HeRs or other proteins). A widespread adaptation of bacteria to alkaline environmental conditions is the increased fluidity of their plasma membranes achieved by the incorporation of branched-chain and unsaturated fatty acids, which ultimately influences the configuration and activity of membrane integral proteins such as ATP

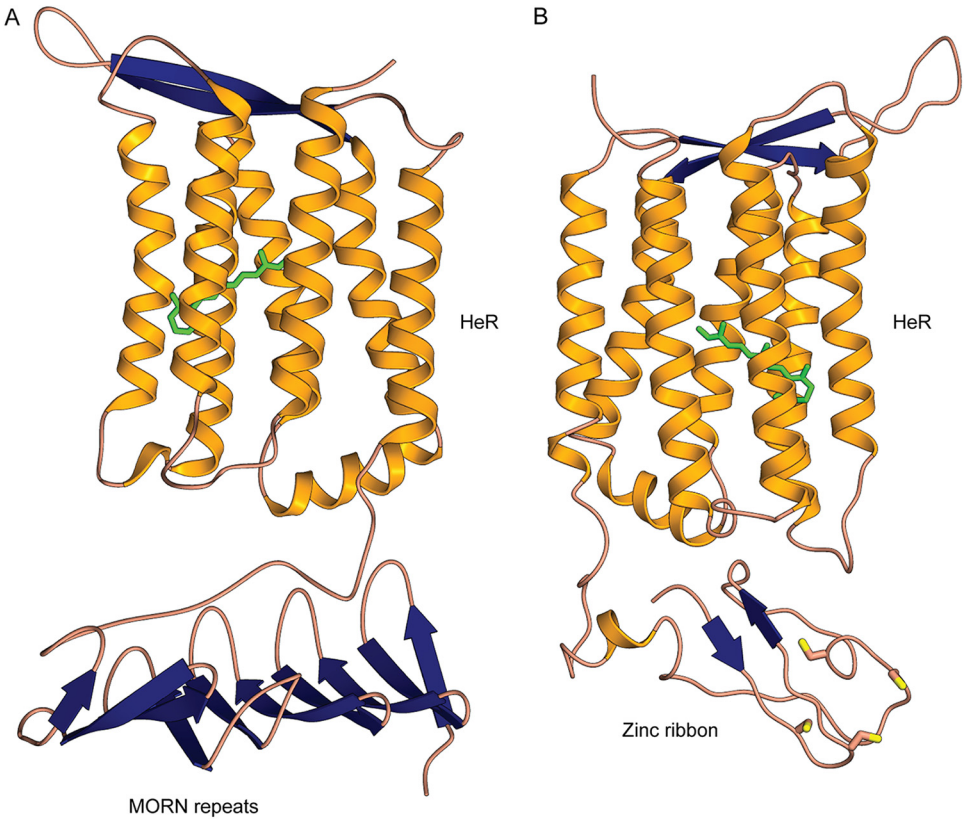

FIG 1 Modeled 3D structures of MORN-HeR and Znf-HeR protein domain fusions. (A) 3D model of a heliorhdodopsin (HeR) containing three N-terminal MORN domain repeats. (B) 3D model of a HeR containing an N-terminal Zn ribbon motif. Both models are oriented with the extracellular side up and the intracellular side down. Retinal is colored green, and cysteine residues are depicted with yellow-topped orange sticks.

synthases and various transporters (21). Microbial rhodopsins typically associate as oligomers *in vivo*, which is also the case with heliorhodopsins that are known to form dimers (5, 22). Indeed, it has been demonstrated that lipid composition of the membrane can directly affect proteorhodopsin dimerization (23). The presence of MORN-repeats in HeRs exclusively within extreme haloalkaliphilic bacteria (class *Dethiobacteria*) may be accounted for via their potential role in stabilizing HeR dimers in conditions of increased membrane fluidity (see Fig. S4 [https://doi.org/10.6084/m9.figshare.13286486]). Another possibility would be the interaction of MORN-repeats with other MORN-repeat containing proteins encoded in these MAGs. We could indeed identify multiple MORN-protein domain fusions co-occurring in genomes of analyzed *Dethiobacteria* (see Fig. S1 and S3 and Table S15 [https://doi.org/10.6084/m9.figshare.13286486]). Even though the nature of interactions among these proteins with intracellular MORN-repeats is unclear, they raise the possibility that MORN-repeats act as downstream transducers of conformational changes occurring in HeRs. Such tandem repeat structures may function as versatile target recognition sites capable of binding not only small molecules like nucleotides but also peptides and larger proteins (24). If true, this would render HeRs as sensory rhodopsins. In support of this, we found several genes in close proximity to MORN-HeRs encoding signature protein domains (e.g., PAS, HisKA, and HATPase_c) that are known to be involved in histidine kinase signaling (25) (Fig. 2A).

Since no other obvious domains were found to be fused with HeRs using standard profile searches, we examined all N- and C-terminal extensions, as well as loops longer than 50 amino acids, by performing more sensitive profile-profile searches using HHpred (26). We found at least 10 N-terminal extensions of HeRs (ntv1 to ntv10), 22 variants of ECL1 (extracellular loop 1), a single type of loop extension for

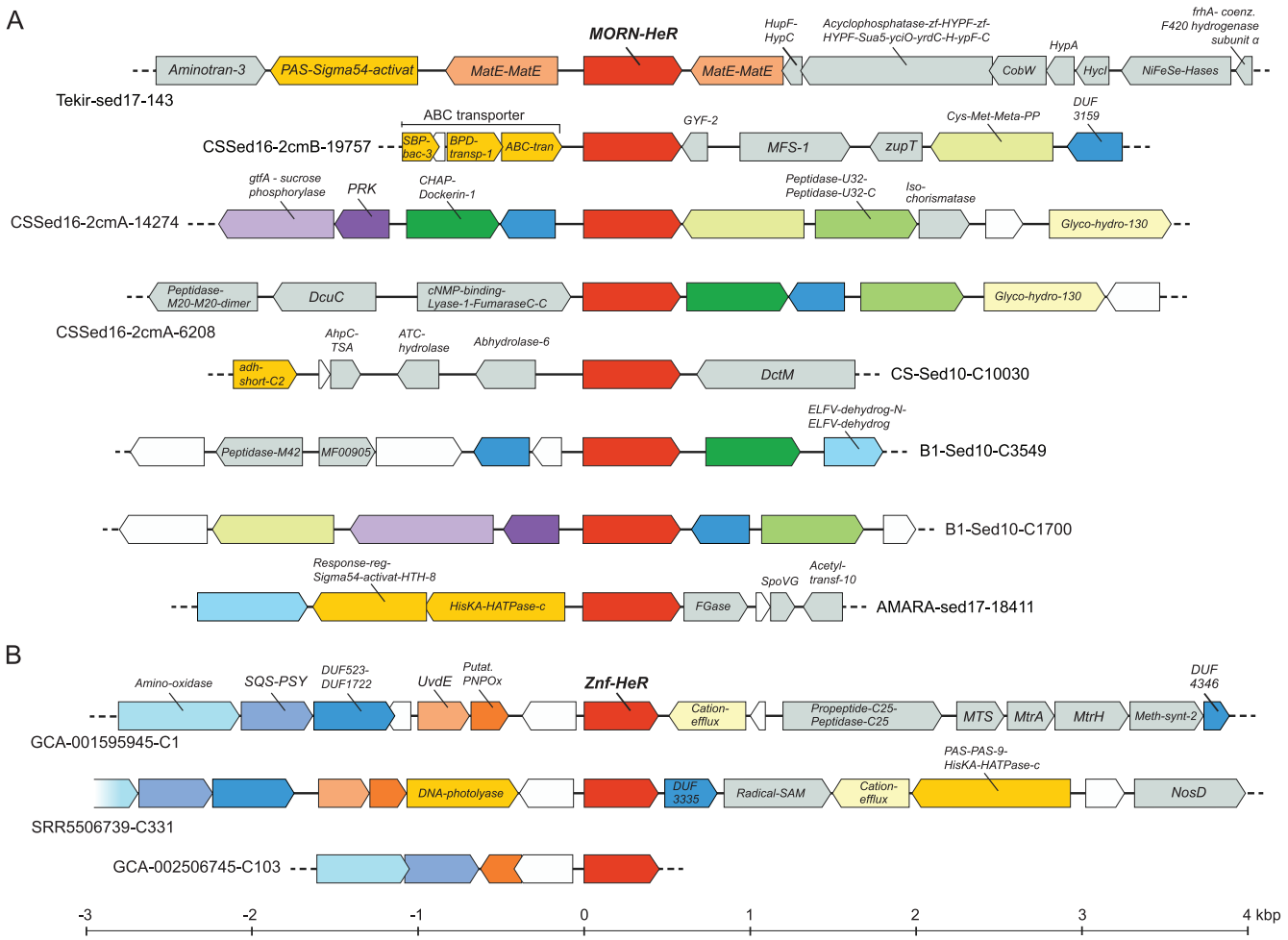

FIG 2 Genomic context of HeR-protein domain fusion genes. (A) Representative MORN-HeR encoding contigs identified in strictly anaerobic *Firmicutes*. (B) Contigs encoding Znf-HeR domain fusions. Neighboring genes were depicted within an interval spanning ~7 kb, centered on HeR. Genes occurring only once within the considered intervals are colored gray; genes encoding HisKA, PAS, and regulatory domains, as well as other discussed HeR neighbors, are depicted in bright yellow. Homologous genes occurring multiple times found within each category of HeR-protein fusion contigs are depicted using matching colors. Hypothetical genes are white.

ICL2 (intracellular loop 2), and three variants of ICL3 (intracellular loop 3). A complete listing of all alignments and summary results of HHpred can be found in Table S8 (https://doi.org/10.6084/m9.figshare.13286486). Remarkably, we found significant matches in a set of six sequences (all originating from *Thermoplasmatales* archaea) to zinc ribbon proteins (Pfam domain zinc_ribbon_4) at the N terminus of some heliorhodopsins (these extensions are termed N-terminal variant 1 or ntv1; see Table S8 [https://doi.org/10.6084/m9.figshare.13286486]). Zinc ribbons belong to the larger family of zinc-finger domains (27). A CxxC-17x-CxxC was found in this region that likely coordinates a metal (e.g., zinc or iron). These CxxC_CxC-type motifs are common to a wider family of zinc-finger-like proteins that were initially found to bind to DNA and later shown to be capable of binding to RNAs, proteins, and small molecules (27). Similar motifs are also seen in rubredoxins and Cys_rich_KTR domains. We term these fused ntv1 protein variants as Znf-HeRs (zinc-finger heliorhodopsins). A modeled structure for a representative Znf-HeR is shown in Fig. 1. In one contig encoding a Znf-HeR we identified a histidine kinase that could be functionally linked (Fig. 2B). Notably, most identified Znf-HeRs are flanked by genes known to be triggered by light exposure and play key roles in photoprotection (i.e., the carotenoid biosynthesis genes, e.g., lycopene cyclase, phytoene desaturase, amino oxidase, and squalene/phytoene synthase [SQS-PSY]) and UV-induced DNA damage repair (DNA

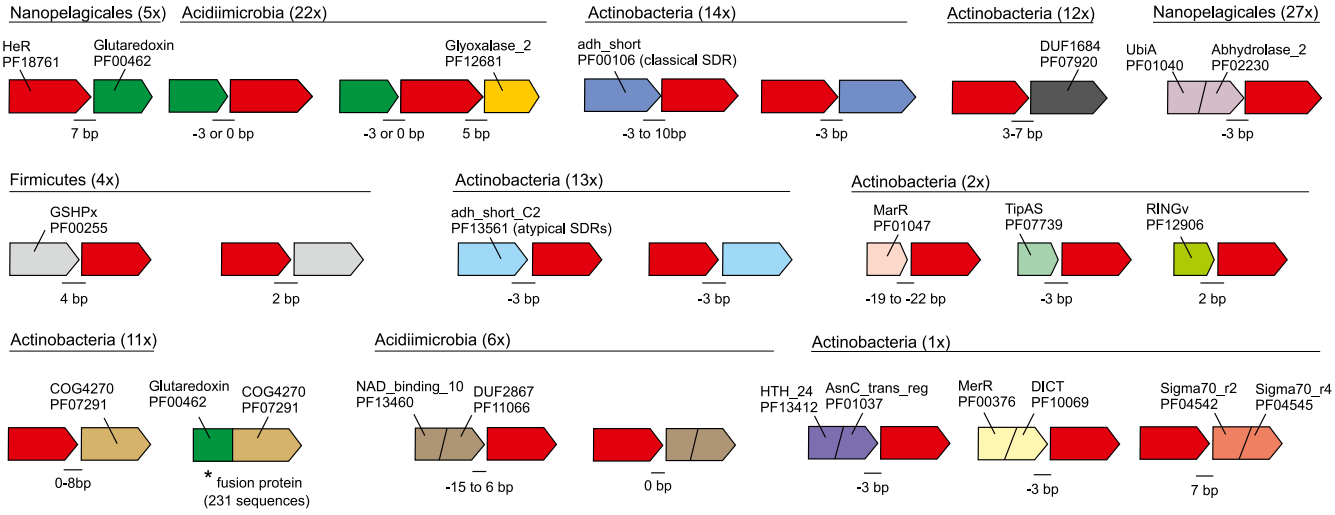

FIG 3 Schematic representation of genes that may be transcriptionally linked to HeRs. Taxonomic categories and number of occurrences are shown at the top of each putative operon. Intergenic distances (in bp) are indicated at gene junctions. Negative distance values indicate overlapping genes. Pfam or COG identifiers are used to represent domain architectures. An asterisk (*) indicates a fused gene (two domains: glutaredoxin and COG4270) found in at least 473 genomes from GTDB and 231 unique sequences in UniProt, suggesting a functional linkage of COG4270 with glutaredoxin.

photolyases and UV-DNA damage endonucleases [UvdE]) (28, 29). Recent research showed that HeRs from *Thermoplasmatales* archaea (*Ta*HeR) and uncultured freshwater *Actinobacteria* (48C12) (for which the structure is resolved and lacks the ntv1 extension) might bind zinc (30). Since the zinc binding site could not be precisely identified, it was suggested that it could be located in the cytoplasmic part and responsible for modifying the function of HeR. Our discovery of Znf-HeRs offers additional, more direct indications of the role of zinc in the possible downstream signaling by HeRs.

Gene context analysis. We reasoned that aside from domain fusions that represent a more direct functional association, gene context analyses, i.e., the repeated presence of specific genes/domains in close proximity to HeRs, may also provide additional clues toward linkage with specific functions. Such linkage may take the form of potential operons or overrepresented genes in the HeR neighborhood. Given the large number of long contigs encoding HeRs (from genomes and metagenomes), we sought to identify candidate genes that could be transcribed together with HeRs (in the same operon). We used the following strict criteria for obtaining such genes: (i) the intergenic distance between such a gene and the HeR must be $<$10 bp, and (ii) the gene must be located on the same strand. A number of interesting candidates emerged in this analysis with the most frequent ones being summarized in Fig. 3 (for a complete table, see Table S9 [https://doi.org/10.6084/m9.figshare.13286486]).

We identified multiple instances in which genes with glutaredoxin and GSHPx PFAM domains were found adjacent to HeRs (n = 31). Glutaredoxins are small redox proteins with active disulfide bonds that utilize reduced glutathione as an electron donor to catalyze thiol-disulfide exchange reactions. They are involved in a wide variety of critical cellular processes such as the maintenance of cellular redox state, iron and redox sensing, and the biosynthesis of iron-sulfur clusters (31, 32). Glutathione is also used by glutathione peroxidase (GSHPx) to reduce hydrogen peroxide and peroxide radicals, i.e., as an antioxidative stress protection system (33). In addition, there are also instances where glutaredoxin and genes containing glyoxalase_2 domains may be cotranscribed with HeRs. Glyoxalases, in concert with glutaredoxins, are critical for the detoxification of methylglyoxal, a toxic by-product of glycolysis (34). Moreover, adjacent to HeRs we find at least three instances where a catalase gene is also present (in *Actinobacteria*; see Fig. S10 and S11 [https://doi.org/10.6084/m9.figshare.13286486]). Collectively, these observations suggest a role for HeRs in oxidative stress mitigation. In one case, we found a gene encoding the DICT domain (Fig. 3), which is frequently

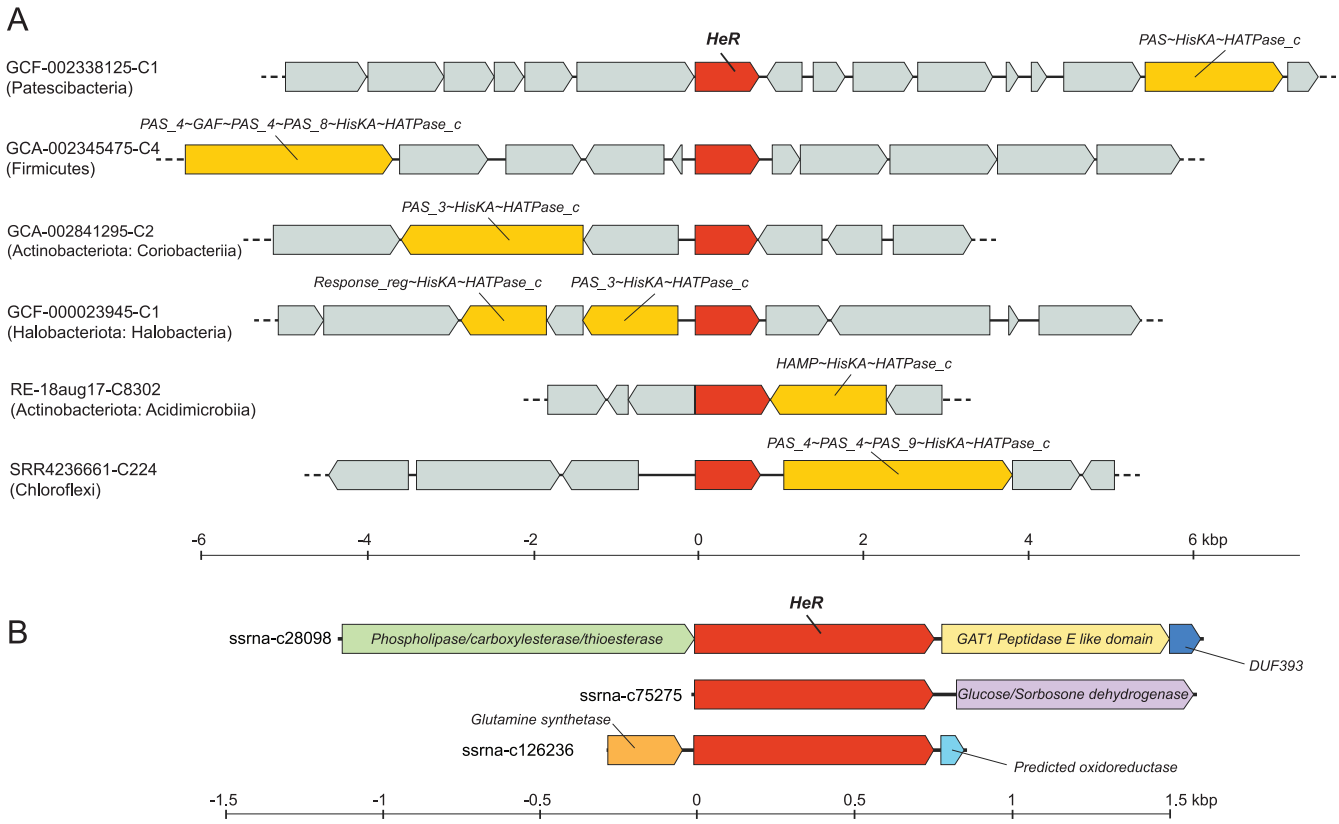

FIG 4 Selected HeR gene contexts. (A) Genes encoding HisKA domain signaling proteins identified in the proximity of HeR genes from diverse phyla. All genes containing HisKA domains are colored bright yellow, HeRs are shown in red, and all other genes are indicated in gray. (B) Transcripts obtained by strand-specific metatranscriptomics from freshwater encoding genes coexpressed with HeR.

associated with GGDEF, EAL, HD-GYP, STAS, and two-component system histidine kinases. Notably, it has been predicted to have a role in light response (25).

Strand-specific metatranscriptomics. Although we assembled contigs encoding HeRs from previously published metatranscriptomes, the lack of strand-specific transcriptomes hampered any clear conclusions on whether or not genes adjacent to HeRs are indeed cotranscribed, leaving open the possibility that they might simply be artifacts of assembly (35). In order to gather more definitive evidence for cotranscription of HeRs with neighboring genes, we performed strand-specific metatranscriptome sequencing for a freshwater sample (see Materials and Methods). The freshwater habitat was chosen because HeRs are widely distributed in these habitats and in particular in freshwater *Actinobacteria* (from which they were originally described) (1). In addition, *Actinobacteria* being among the most abundant microbes in these habitats (36, 37) would increase the chances for recovery of such polycistronic transcripts.

We recovered six HeR-encoding transcripts that were >1 kb in length. All these transcripts are predicted to originate from highly abundant freshwater *Actinobacteria* with streamlined genomes (four transcripts from "*Ca.* Planktophila" and two from "*Ca.* Nanopelagicus") (see Table S12 [https://doi.org/10.6084/m9.figshare.13286486]) (37). Overall, there are three types of transcripts based upon gene content: (i) class1, encoding glutamine synthetase catalytic subunit and NAD^+ synthetase; (ii) class 2, encoding a hydrolase, a peptidase, and a DUF393 domain containing protein; and (iii) class 3, encoding glucose/sorbosone dehydrogenase (GSDH) (Fig. 4B; see also Table S12 [https://doi.org/10.6084/m9.figshare.13286486]). A common theme for glutamine synthetase and NAD^+ synthetase is that both utilize ammonia and ATP to produce glutamine and NAD^+, respectively. Moreover, some NAD^+ synthetases may be glutamine dependent (38). Glutamine synthetase in particular is a key enzyme for nitrogen metabolism in prokaryotes at large (39). For hydrolases and peptidases, the function

prediction is somewhat broad. Glucose/sorbosone dehydrogenase catalyzes the production of gluconolactone from glucose (40). Therefore, it appears that all six HeRs are generally cotranscribed with genes involved in nitrogen assimilation and degradation/assimilation of sugars and peptides. This would suggest that these processes are also influenced by light, with such a link between light-dependent increase in sugar uptake and metabolic activity being recently proposed in nonphototrophic *Actinobacteria* (41). Light also triggers photosynthetic activity, increasing the availability of sugars and other nutrients (e.g., glutamine and ammonia) for heterotrophs. In this vein, a link between a light sensing mechanism, e.g., via heliorhodopsins, may lead to elevated metabolic activity.

In a previous study, histidine kinases were deemed absent in the vicinity of HeRs (2). Given that our initial analyses predicted a sensory function, we examined genomic regions spanning 10 kb up- and downstream of HeRs. Already in the case of MORN-HeRs and Znf-HeRs, we observed histidine kinase signaling components in close proximity to them (Fig. 2). In our search we detected multiple instances of histidine kinases (HisKA) fused with PAS, GAF, MCP_Signal, HAMP, or HATPase_c domains in the gene neighborhoods of HeRs in distinct phyla (e.g., *Actinobacteria*, *Chloroflexi*, *Patescibacteria*, *Firmicutes*, *Dictyoglomota*, and *Thermoplasmatota*) (Fig. 4B; for more details, see Fig. S4 to S15 [https://doi.org/10.6084/m9.figshare.13286486]). Moreover, in many cases multiple response regulator genes were present in the same regions (Pfam domains Response_reg and Trans_reg_C). Less frequently, GGDEF and EAL domains, usually associated with bacterial signaling proteins, were also present. Using overrepresentation analysis (42), we found that the occurrence of two-component system protein domains in the vicinity of HeRs is statistically significant (see Materials and Methods and Table S11 [https://doi.org/10.6084/m9.figshare.13286486]). In addition to these two-component system proteins, the same regions also appear enriched in redox proteins (e.g., thioredoxin, peroxidase, and catalase). The close association of two-component systems, genes involved in oxidative stress mitigation and HeRs points toward a functional interaction.

DISCUSSION

Contextual genomic information shows that monoderm prokaryotes use HeRs in multiple mechanisms for the activation of downstream metabolic pathways after light sensing. These observations offer tantalizing clues regarding the involvement of HeRs in multiple cellular processes and add new lines of inquiry for the primary role of HeRs in light-activated signal transduction. Additional support for the role of HeRs in light sensing is inferred from the frequent association of HeRs with classical histidine kinases and associated protein domains in multiple phyla. Furthermore, multiple types of N-terminal domain fusions found in specific subfamilies of HeRs (i.e., MORN domains in haloalkaliphilic *Firmicutes* and zinc-ribbon-type domains in *Thermoplasmatales* archaea) point to possible downstream signaling which may be effected by the recruitment of additional, as-yet-unknown, partner proteins.

We further propose a critical role for HeRs in protecting monoderm cells from light-induced oxidative damage. In this sense, we observed a close association and probable transcriptional linkage of HeRs to glyoxylases and glutaredoxins (sometimes seen as overlapping genes). Given that light can induce the uptake and metabolism of sugars, as previously discussed for certain *Actinobacteria* (41), it is expected that increased sugar availability resulting from photosynthesis leads to increased glycolytic activity in heterotrophic bacteria. Glycolysis also produces small amounts of toxic methylglyoxal that can be neutralized by the combined action of glyoxylases and glutaredoxins. In this sense, it appears that at least in some *Actinobacteria* glyoxylases and glutaredoxins may be transcribed together with HeRs, but how the transcription is controlled remains unclear. Additional evidence of transcriptional linkages of HeRs to proteins like peroxiredoxin and catalase also imply a light-dependent activation, boosting the cellular response to light induced oxidative damage which may be critical for both aerobes

and anaerobes. Evidence from strand-specific HeR transcripts originating from freshwater *Actinobacteria* suggests the further involvement of HeRs in nitrogen and sugar metabolism via glutamate synthase, NAD^+ synthases, and glucose/sorbosone dehydrogenases in these organisms. However, direct experimental evidence of interactions of HeRs with the genes proposed here could take multiple forms, e.g., strand-specific transcriptomics data from cultured microbes that encode HeRs supplemented with similar data from diverse environments, and the use of HeR knockouts combined with transcriptomic data under conditions of light and dark.

Overall, the picture that emerges (at least for some organisms) is one of HeR's roles in responding to light and transmitting the signal via histidine kinases. Downstream processes that are ultimately regulated are diverse, including possible roles for HeRs in the mitigation of light-induced oxidative damage and in the regulation of nitrogen assimilation and carbohydrate metabolism, processes that may benefit from a light-dependent activation through more efficient utilization of available resources.

Recent work has shown more support for the diderm-first ancestor (43) and, given the far broader distribution of type 1 rhodopsins in both mono- and diderm organisms, it appears likely that type 1 rhodopsins emerged prior to HeRs. The very restricted distribution of HeRs to monoderms would support this view as well. Even so, HeRs are not universally present in monoderms and, when present, appear to be associated with diverse genes involved in signal transduction, oxidative stress mitigation, and nitrogen and glucose metabolism. This suggests they have been exapted as generalized sensory switches that may allow light-dependent control of metabolic activity in multiple lineages, somewhat similar to type 1 rhodopsins where minor modifications have led to emergence of a wide variety of ion pumps (44). The frequent distribution of HeRs in aquatic environments (habitats characterized by increased light penetration), where they commonly occur within phylum *Actinobacteriota*, helps us to explain their monoderm-restricted presence. Abundant freshwater actinobacterial lineages are generally typified by lower GC content (45) and increased vulnerability to oxidative stress damage (46). This susceptibility is also illustrated by actinobacterial phages that exhibit positive selection toward reactive oxygen species defense mechanisms (36). This suggests that oxidative stress is a considerable influence in environment at large, and it has indeed been identified as such before (47, 48). Light-induced, oxygen-dependent inactivation has also been demonstrated in other bacterial species as well (49). Such inactivation is understood to be the direct result of the production of reactive oxygen species by endogenous porphyrins in the presence of light (50, 51). Moreover, reactive oxygen species are also released by other community members and generated by UV-induced photochemical reactions (47). Given the fact that monoderms are generally more sensitive to light-induced damage (52) and taken together with the above-mentioned metabolic implications, we consider that HeRs evolved as sensory switches capable of triggering a fast response against photo-oxidative stress in prokaryotic lineages more sensitive to light.

MATERIALS AND METHODS

Metagenomes and metatranscriptomes. We used previously published metagenomics and metatranscriptomics data from freshwaters (36, 53, 54), haloalkaline brine and sediment (14, 15), brackish sediments (55), a GEOTRACES cruise (56) and TARA expeditions (57). In addition, we downloaded multiple environmental metagenomes (sludge, marine, pond, estuary, etc.) from EBI MGnify (https://www.ebi.ac.uk/metagenomics/) (58) and assembled them using Megahit v1.2.9 (59). All contigs in this work are named or retain existing names that allow tracing them to their original data sets.

Sequence search for *bona fide* rhodopsins. Genes were predicted in metagenomics contigs using Prodigal (60). Candidate rhodopsin sequences were scanned with hmmsearch (61) using PFAM models (PF18761, heliorhodopsin; PF01036, bac_rhodopsin), and only hits with significant E values ($<$1E–3) were retained. Homologs for these sequences were identified by comparison to a known set of rhodopsin sequences (55) using MMSeqs2 (62), and alignments were made using MAFFT-linsi (63). These alignments were used as input to Polyphobius (64) for transmembrane helix prediction. Only those sequences that had seven transmembrane helices and either a SxxxK motif (for heliorhodopsins) or DxxxK motif (for proteorhodopsins) in TM7 were retained. In addition, we also screened the entire UniProtKB for HeRs. In total, we accumulated at least 4,108 (3,606 + 502) *bona fide* HeR sequences.

Taxonomic classification of assembled contigs. Contigs were dereplicated using cd-hit (65) (95% sequence identity and 95% coverage). Only contigs $\geq$5 kb were retained for this analysis. A custom

protein database was created by predicting and translating genes in all GTDB genomes (release 89) (8) using Prodigal (60). These sequences were supplemented with viral and eukaryotic proteins from UniProtKB (66). Best hits against predicted proteins in contigs were obtained using MMSeqs2 (62). Taxonomy was assigned to a contig (minimum length, 5 kb) only if ≥60% of genes in the contig gave best hits to the same phylum. All contigs that appeared to originate from diderms were cross-checked against NCBI RefSeq (accessed online on 15 December 2020).

Outer-envelope detection. A set of protein domains found in genes encoding the outer-envelope (9) was further reduced to include only those domains that were found mostly in known diderms. These domains were searched against the predicted proteins in all genomes in GTDB using hmmsearch (E value < 1E–3). The results are shown in Table S13 (https://doi.org/10.6084/m9.figshare.13286486).

Protein function-structure predictions. Predicted proteins were annotated using TIGRFAMs (67) and COGs (68). Domain predictions were carried out using the pfam_scan.pl script against the PFAM database (release 32) (17). Profile-profile searches were carried out online using the HHPred server (26). Additional annotations were added using Interproscan (69). Protein structure predictions were carried out using the Phyre2 server (70), and structures were visualized with CueMol (http://www.cuemol.org/en/).

Domains overrepresentation near heliorhodopsin. A subset of high-quality MAGs (n = 240) containing HeR-encoding genes flanked both up- and downstream by a minimum of 10 genes were selected from GTDB (release 89) (8). For each genome, the probability of finding any particular domain by chance in a random subset of 20 genes was calculated using the hypergeometric distribution (without replacement) in R with the function *phyper* (Stats package) (71). In order to account for type I errors arising from multiple comparisons, hypergeometric test P values were adjusted using the Benjamini-Hochberg procedure (72). Further, we selected domains located in the proximity of HeR in at least 10% of genomes with low probability (false discovery rate corrected P value < 0.05). This procedure that was initially employed for the whole GTDB genome collection was repeated for individual phyla containing HeR-encoding genes within at least five genomes.

Strand-specific freshwater transcriptome sequencing and assembly. Sampling was performed on the 16th of August 2020 at 9:00 in Římov Reservoir, Czech Republic, (48°50′54.4″N, 14°29′16.7″E) using a hand-held vertical Friedinger (2 L) sampler. A total of 20 L of water were collected from a depth of 0.5 m and immediately transported to the laboratory. Serial filtration was carried out by passing water sample through a 20-μm-pore-size prefilter mesh, followed by a 5-μm-pore-size PES filter (Sterlitech) and a 0.22-μm-pore-size PES (polyethersulfone) filter (Sterlitech, USA) using a Masterflex peristaltic pump (Cole-Palmer, USA). Filtration was done at maximum speed for 15 min to limit cell lysis and RNA damage as much as possible. A total volume of 3.7 L was filtered during this time. PES filters (5-μm and 0.22-μm pore sizes) were loaded into cryo-vials prefilled with 500 μl of DNA/RNA Shield (Zymo Research, USA) and stored at −80°C. RNA was extracted from filters using the Direct-zol RNA MicroPrep (Zymo Research) after they had been previously thawed, partitioned, and subjected to mechanical lysis by bead beating in ZR BashingBead lysis tubes (with 0.1- and 0.5-mm spheres). DNase treatment was performed to remove genomic DNA during RNA extraction as an "in-column" step described in the Direct-zol protocol and was repeated after RNA elution, by using the Ambion Turbo DNA-free kit (Life Technologies, USA). RNA was quantified using a NanoDrop ND-1000 UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), and integrity was verified by agarose gel (1%) electrophoresis. A total of 4.6 μg of RNA from the 0.22-μm-pore-size filter and 2.6 μg from the 5-μm-pore-size filter were sent for dUTP-marking based strand-specific metatranscriptomic sequencing at Novogene. Following quality control at Novogene, the samples were mixed into a single reaction, subjected to rRNA depletion, and used for stranded library preparation. Strand specificity was achieved by the incorporation of dUTPs instead of dTTPs in the second-strand cDNA, followed by digestion of dUTPs by uracil-DNA glycosylase to prevent PCR amplification of this strand. Paired-end (PE 150 bp) sequencing was carried out using a Novaseq 6000 platform. A total of 166,213,184 raw sequencing reads, amounting to 24.9 Gb, were produced. *De novo* assembly of metatranscriptomic data was performed using rnaSPAdes v3.14.1 (73) in reverse-forward strand-specific mode (–ss rf) with the custom k-mers list 29, 39, 49, 59, 69, 79, 89, 99, 109, 119, and 127. A total of 156,235 hard-filtered transcripts of a minimum length of 1 kb were assembled. Protein coding sequences were predicted *de novo* using Prodigal (60) in metagenomic mode (-p meta). Protein domains were annotated by scanning with InterProScan (69), while PFAM (Protein Families) (17) domains were identified using the publicly available Perl script pfam_scan.pl (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/). Proteins were scanned locally using HMMER3 (61) against the COGs (Clusters of Orthologous Groups) (68) HMM database (E value < 1E–5) and the TIGRFAMs (TIGR Families) (67) HMM collection with trusted score cutoffs. BlastKOALA (74) was used to assign KO identifiers (KO numbers). Annotations for representative transcripts encoding HeR are summarized in Table S12 (https://doi.org/10.6084/m9.figshare.13286486).

Data availability. Sequence data generated in this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under project accession number PRJEB35770 (run ERR5100021). The derived data that support the findings of this paper are available in FigShare (https://doi.org/10.6084/m9.figshare.13286486). All other relevant data supporting the findings of this study are available within the paper and its supplementary information files. The R code used for statistical analyses is available in FigShare (https://doi.org/10.6084/m9.figshare.13286486).

ACKNOWLEDGMENTS

We thank Petr Znachor and Pavel Rychtecký for help with the Římov Reservoir sampling.

We declare no competing interests.

P.-A.B., V.S.K., and R.G. were supported by the research grant 20-12496X (Grant Agency of the Czech Republic). V.S.K. was additionally supported by the research grant 116/2019/P (Grant Agency of the University of South Bohemia in České Budějovice, 2019-2021). A.-S.A. was supported by Ambizione grant PZ00P3_193240 (Swiss National Science Foundation). M.-C.C. was supported by the Program for the Support of Perspective Human Resources (PPLZ), Czech Academy of Sciences (grant L200961953). K.I. was supported by Grants-in-Aid from the Japan Society for the Promotion of Science (JSPS) for Scientific Research (KAKENHI grants 20K21383 and 20H05758). H.K. was supported by a research grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology (18H03986) and a grant from CREST, Japan Science and Technology Agency (JPMJCR1753). The funders had no role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

R.G. and P.-A.B. designed the study. P.-A.B., A.-S.A., and R.G. wrote the manuscript. P.-A.B., R.G., V.S.K., M.-C.C., C.D.V., and A.-S.A. analyzed and interpreted the data. K.I. and H.K. performed rhodopsin structural analyses. All authors commented on and approved the manuscript.

REFERENCES

1. 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.
2. Kovalev K, Volkov D, Astashkin R, Alekseev A, Gushchin I, Haro-Moreno JM, Chizhov I, Siletsky S, Mamedov M, Rogachev A, Balandin T, Borshchevskiy V, Popov A, Bourenkov G, Bamberg E, Rodriguez-Valera F, Büldt G, Gordeliy V. 2020. High-resolution structural insights into the heliorhodopsin family. Proc Natl Acad Sci U S A 117:4131–4141. https://doi.org/10.1073/pnas.1915888117.
3. Flores-Uribe J, Hevroni G, Ghai R. 2019. Heliorhodopsins are absent in diderm (Gram-negative) bacteria: some thoughts and possible implications for activity. Environ Microbiol Rep 11:419–424.
4. Tanaka T, Singh M, Shihoya W, Yamashita K, Kandori H, Nureki O. 2020. Structural basis for unique color tuning mechanism in heliorhodopsin. Biochem Biophys Res Commun 533:262–267. https://doi.org/10.1016/j.bbrc.2020.06.124.
5. Shihoya W, Inoue K, Singh M, Konno M, Hososhima S, Yamashita K, Ikeda K, Higuchi A, Izume T, Okazaki S, Hashimoto M, Mizutori R, Tomida S, Yamauchi Y, Abe-Yoshizumi R, Katayama K, Tsunoda SP, Shibata M, Furutani Y, Pushkarev A, Béjà O, Uchihashi T, Kandori H, Nureki O. 2019. Crystal structure of heliorhodopsin. Nature 574:132–136. https://doi.org/10.1038/s41586-019-1604-6.
6. Aravind L. 2000. Guilt by association: contextual information in genome analysis. Genome Res 10:1074–1077. https://doi.org/10.1101/gr.10.8.1074.
7. Huynen M, Snel B, Lathe W, III, Bork P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10:1204–1210. https://doi.org/10.1101/gr.10.8.1204.
8. 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.
9. Taib N, Megrian D, Witwinowski J, Adam P, Poppleton D, Borrel G, Beloin C, Gribaldo S. 2020. Genome-wide analysis of the *Firmicutes* illuminates the diderm/monoderm transition. Nat Ecol Evol 4:1661–1672. https://doi.org/10.1038/s41559-020-01299-7.
10. Megrian D, Taib N, Witwinowski J, Beloin C, Gribaldo S. 2020. One or two membranes? Diderm *Firmicutes* challenge the Gram-positive/Gram-negative divide. Mol Microbiol 113:659–671. https://doi.org/10.1111/mmi.14469.
11. Saiki T, Kobayashi Y, Kawagoe K, Beppu T. 1985. *Dictyoglomus thermophilum* gen. nov., sp. nov., a chemoorganotrophic, anaerobic, thermophilic bacterium. Int J Syst Evol Microbiol Microbiol Soc 35:253–259.
12. Ikuta T, Shihoya W, Sugiura M, Yoshida K, Watari M, Tokano T, Yamashita K, Katayama K, Tsunoda SP, Uchihashi T, Kandori H, Nureki O. 2020. Structural insights into the mechanism of rhodopsin phosphodiesterase. Nat Commun 11:5605. https://doi.org/10.1038/s41467-020-19376-7.
13. Timmers PHA, Vavourakis CD, Kleerebezem R, Damsté JSS, Muyzer G, Stams AJM, Sorokin DY, Plugge CM. 2018. Metabolism and occurrence of methanogenic and sulfate-reducing syntrophic acetate oxidizing communities in haloalkaline environments. Front Microbiol 9:3039. https://doi.org/10.3389/fmicb.2018.03039.
14. 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.
15. 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.
16. Takeshima H, Komazaki S, Nishi M, Iino M, Kangawa K. 2000. Junctophilins: a novel family of junctional membrane complex proteins. Mol Cell 6:11–22.
17. 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.
18. Im YJ, Davis AJ, Perera IY, Johannes E, Allen NS, Boss WF. 2007. The N-terminal membrane occupation and recognition nexus domain of *Arabidopsis* phosphatidylinositol phosphate kinase 1 regulates enzyme activity. J Biol Chem 282:5443–5452. https://doi.org/10.1074/jbc.M611342200.
19. Ma H, Lou Y, Lin WH, Xue HW. 2006. MORN motifs in plant PIPKs are involved in the regulation of subcellular localization and phospholipid binding. Cell Res 16:466–478. https://doi.org/10.1038/sj.cr.7310058.
20. Sajko S, Grishkovskaya I, Kostan J, Graewert M. 2020. Structures of three MORN repeat proteins and a re-evaluation of the proposed lipid-binding properties of MORN repeats. bioRxiv https://www.biorxiv.org/content/10.1101/826180v2.abstract.
21. Kanno M, Tamaki H, Mitani Y, Kimura N, Hanada S, Kamagata Y. 2015. pH-induced change in cell susceptibility to butanol in a high butanol-tolerant bacterium, *Enterococcus faecalis* strain CM4A. Biotechnol Biofuels 8:69. https://doi.org/10.1186/s13068-015-0251-x.
22. Shibata M, Inoue K, Ikeda K, Konno M, Singh M, Kataoka C, Abe-Yoshizumi R, Kandori H, Uchihashi T. 2018. Oligomeric states of microbial rhodopsins determined by high-speed atomic force microscopy and circular dichroic spectroscopy. Sci Rep 8:8262. https://doi.org/10.1038/s41598-018-26606-y.
23. Sefah E, Mertz B. 2021. Bacterial analogs to cholesterol affect dimerization of proteorhodopsin and modulates preferred dimer interface. J Chem Theory Comput 17:2502–2512. https://doi.org/10.1021/acs.jctc.0c01174.
24. Kajava AV. 2012. Tandem repeats in proteins: from sequence to structure. J Struct Biol 179:279–288. https://doi.org/10.1016/j.jsb.2011.08.009.
25. Aravind L, Iyer LM, Anantharaman V. 2010. Natural history of sensor domains in bacterial signaling systems, p 1–38. *In* Stephen RD (ed),

Sensory mechanisms in bacteria: molecular aspects of signal recognition. Caister Academic Press, Norfolk, UK.
26. 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.
27. Krishna SS, Majumdar I, Grishin NV. 2003. Structural classification of zinc fingers: survey and summary. Nucleic Acids Res 31:532–550. https://doi.org/10.1093/nar/gkg161.
28. Rastogi RP, Richa Kumar A, Tyagi MB, Sinha RP. 2010. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. J Nucleic Acids 2010:592980. https://doi.org/10.4061/2010/592980.
29. Yatsunami R, Ando A, Yang Y, Takaichi S, Kohno M, Matsumura Y, et al. 2014. Identification of carotenoids from the extremely halophilic archaeon *Haloarcula japonica*. Front Microbiol 5:100.
30. Hashimoto M, Katayama K, Furutani Y, Kandori H. 2020. Zinc binding to heliorhodopsin. J Phys Chem Lett 11:8604–8609. https://doi.org/10.1021/acs.jpclett.0c02383.
31. Lillig CH, Berndt C, Holmgren A. 2008. Glutaredoxin systems. Biochim Biophys Acta 1780:1304–1317. https://doi.org/10.1016/j.bbagen.2008.06.003.
32. Rouhier N, Couturier J, Johnson MK, Jacquot J-P. 2010. Glutaredoxins: roles in iron homeostasis. Trends Biochem Sci 35:43–52. https://doi.org/10.1016/j.tibs.2009.08.005.
33. Bhabak KP, Mugesh G. 2010. Functional mimics of glutathione peroxidase: bioinspired synthetic antioxidants. Acc Chem Res 43:1408–1419. https://doi.org/10.1021/ar100059g.
34. Ferguson GP, Tötemeyer S, MacLean MJ, Booth IR. 1998. Methylglyoxal production in bacteria: suicide or survival? Arch Microbiol 170:209–218. https://doi.org/10.1007/s002030050635.
35. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, von Schack D, Zhang B. 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics 16:675. https://doi.org/10.1186/s12864-015-1876-7.
36. 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.
37. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. 2018. Microdiversification in genome-streamlined ubiquitous freshwater *Actinobacteria*. ISME J 12:185–198. https://doi.org/10.1038/ismej.2017.156.
38. Resto M, Yaffe J, Gerratana B. 2009. An ancestral glutamine-dependent NAD(+) synthetase revealed by poor kinetic synergism. Biochim Biophys Acta 1794:1648–1653. https://doi.org/10.1016/j.bbapap.2009.07.014.
39. García-Domínguez M, Reyes JC, Florencio FJ. 1999. Glutamine synthetase inactivation by protein-protein interaction. Proc Natl Acad Sci U S A 96:7161–7166. https://doi.org/10.1073/pnas.96.13.7161.
40. Oubrie A, Rozeboom HJ, Kalk KH, Olsthoorn AJ, Duine JA, Dijkstra BW. 1999. Structure and mechanism of soluble quinoprotein glucose dehydrogenase. EMBO J 18:5187–5194. https://doi.org/10.1093/emboj/18.19.5187.
41. Maresca JA, Keffer JL, Hempel PP, Polson SW, Shevchenko O, Bhavsar J, Powell D, Miller KJ, Singh A, Hahn MW. 2019. Light modulates the physiology of nonphototrophic actinobacteria. J Bacteriol 201 https://doi.org/10.1128/JB.00740-18.
42. Shmakov SA, Makarova KS, Wolf YI, Severinov KV, Koonin EV. 2018. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc Natl Acad Sci U S A 115:E5307–E5316. https://doi.org/10.1073/pnas.1803440115.
43. Coleman GA, Davín AA, Mahendrarajah T, Spang A, Hugenholtz P, Szöllősi GJ, et al. 2020. A rooted phylogeny resolves early bacterial evolution. Cold Spring Harb Lab https://www.biorxiv.org/content/10.1101/2020.07.15.205187v1.
44. Kandori H. 2020. Biophysics of rhodopsins and optogenetics. Biophys Rev 12:355–361. https://doi.org/10.1007/s12551-020-00645-0.
45. Ghai R, McMahon KD, Rodriguez-Valera F. 2012. Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. Environ Microbiol Rep 4:29–35. https://doi.org/10.1111/j.1758-2229.2011.00274.x.
46. Kim S, Kang I, Seo J-H, Cho J-C. 2019. Culturing the ubiquitous freshwater actinobacterial acI lineage by supplying a biochemical "helper" catalase. ISME J 13:2252–2263. https://doi.org/10.1038/s41396-019-0432-x.
47. Imlay JA. 2013. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. Nat Rev Microbiol 11:443–454. https://doi.org/10.1038/nrmicro3032.
48. Ezraty B, Gennaris A, Barras F, Collet J-F. 2017. Oxidative stress, protein damage and repair in bacteria. Nat Rev Microbiol 15:385–396. https://doi.org/10.1038/nrmicro.2017.26.
49. Feuerstein O, Ginsburg I, Dayan E, Veler D, Weiss EI. 2005. Mechanism of visible light phototoxicity on *Porphyromonas gingivalis* and *Fusobacterium nucleatum*. Photochem Photobiol 81:1186–1189. https://doi.org/10.1562/2005-04-06-RA-477.
50. Hamblin MR, Hasan T. 2004. Photodynamic therapy: a new antimicrobial approach to infectious disease? Photochem Photobiol Sci 3:436–450. https://doi.org/10.1039/b311900a.
51. Wainwright M. 1998. Photodynamic antimicrobial chemotherapy (PACT). J Antimicrob Chemother 42:13–28. https://doi.org/10.1093/jac/42.1.13.
52. Maclean M, MacGregor SJ, Anderson JG, Woolsey G. 2009. Inactivation of bacterial pathogens following exposure to light from a 405-nanometer light-emitting diode array. Appl Environ Microbiol 75:1932–1937. https://doi.org/10.1128/AEM.01892-08.
53. Andrei A-S, 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.
54. 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.
55. 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.
56. 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.
57. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, Field CM, Coelho LP, Cruaud C, Engelen S, Gregory AC, Labadie K, Marec C, Pelletier E, Royo-Llonch M, Roux S, Sánchez P, Uehara H, Zayed AA, Zeller G, Carmichael M, Dimier C, Ferland J, Kandels S, Picheral M, Pisarev S, Poulain J, Acinas SG, Babin M, Bork P, Bowler C, de Vargas C, Guidi L, Hingamp P, Iudicone D, Karp-Boss L, Karsenti E, Ogata H, Pesant S, Speich S, Sullivan MB, Wincker P, Sunagawa S, Tara Oceans Coordinators. 2019. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell 179:1068–1083. https://doi.org/10.1016/j.cell.2019.10.014.
58. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 48:D570–D578.
59. 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.
60. 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.
61. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
62. Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 32:1323–1330. https://doi.org/10.1093/bioinformatics/btw006.
63. 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.
64. 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.
65. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
66. UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049.
67. 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.

68. 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.

69. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong S-Y, Finn RD. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. https://doi.org/10.1093/nar/gky1100.

70. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858. https://doi.org/10.1038/nprot.2015.053.

71. Johnson NL, Kemp AW, Kotz S. 2005. Univariate discrete distributions. John Wiley & Sons, New York, NY.

72. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x

73. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8:giz100. https://doi.org/10.1093/gigascience/giz100.

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.