# Flotillin-Associated rhodopsin (FArhodopsin), a widespread paralog of proteorhodopsin in aquatic bacteria with streamlined genomes

Jose M. Haro-Moreno[a,b], Mario López-Pérez[a], Alexey Alekseev[c], Elizaveta Podoliak[d], Kirill Kovalev[e], Valentin Gordeliy[b,f], Ramunas Stepanauskas[g], and Francisco Rodriguez-Valera[a,*]

[a]Evolutionary Genomics Group, Departamento Producción Vegetal y Microbiología, Universidad Miguel Hernández, Apartado 18, San Juan 03550, Alicante, Spain. [b]Institut de Biologie Structurale J.-P. Ebel, Université Grenoble Alpes-Commission for Atomic Energy (CEA)-CNRS, 38000 Grenoble, France. [c]Institute for Auditory Neuroscience and InnerEarLab, University Medical Center Göttingen, Göttingen, Germany. [d]Department of Ophthalmology, Universitäts-Augenklinik Bonn, Germany. [e]EMBL Hamburg, Hamburg, Germany. [f]Institute of Biological Information Processing (Institute of Biological Information Processing: Structural Biochemistry), Forschungszentrum Jülich, 52428 Jülich, Germany. [g]Bigelow Laboratory for Ocean Sciences. East Boothbay, USA.

*Corresponding author: frvalera@umh.es

Evolutionary Genomics Group, Departamento Producción Vegetal y Microbiología, Universidad Miguel Hernández, Apartado 18, San Juan de Alicante, 03550 Alicante, Spain.

Phone +34-965919313, Fax +34-965 919457

Running title: Rhodopsin diversity in streamlined genomes

Keywords: Rhodopsin / Pelagibacterales / Single-amplified genomes / Metagenomics / rhodopsin paralogs / streamlined genomes /

## ABSTRACT

Microbial rhodopsins are often found more than once in a single genome (paralogs) that often have different functions. We screened a large dataset of open ocean single-amplified genomes (SAGs) for co-occurrences of multiple rhodopsin genes. Many such cases were found among Pelagibacterales (SAR11), HIMB59 and the Gammaproteobacteria *Pseudothioglobus* SAGs. These genomes always had a *bona fide* proteorhodopsin and a separate cluster of genes containing a second rhodopsin associated with a predicted flotillin coding gene and have thus been named flotillin-associated rhodopsins (FArhodopsins). They are quite divergent from the other proteorhodopsin paralog and contain either DTT, DTL or DNI motives in their key functional amino acids. FArhodopsins are mainly associated with the lower layers of the epipelagic zone. All marine FArhodopsins had the retinal binding lysine, but we found their relatives in freshwater metagenomes that lack this key amino acid. Alfa-fold predictions of marine FArhodopsins indicate that their retinal pocket might be very reduced or absent, hinting that they are retinal-less (blind). Freshwater FArhodopsins were more diverse than marine FArhodopsins, but we could not determine if they are present as paralogs of other rhodopsins, due to the lack of SAGs or isolates. Although the function of FArhodopsins could not be established, their conserved genomic context indicated involvement in the formation of membrane microdomains. The conservation of FArhodopsins in diverse and globally abundant microorganisms suggests that they may be important in the adaptation to the twilight zone of aquatic environments.

## IMPORTANCE

Rhodopsins have been shown to play a key role in the ecology of aquatic microbes. Here we describe a group of widespread rhodopsins in aquatic microbes associated with dim light conditions. Their characteristic genomic context found in both marine and freshwater environments indicates a novel potential involvement in membrane microstructure that could be important for the function of the co-existing proteorhodopsin proton pumps. The absence or reduction of the retinal binding pocket points to drastically different physiology. In addition to their ecological importance, novel rhodopsins have biotechnological potential in the nascent field of optogenetics.

## INTRODUCTION

Rhodopsins are probably the most universal biological light-energy transducers. They are found in all the domains of life (eukaryotes, bacteria, and archaea) and also in viruses (1). It is now well established that microbial rhodopsins (like microbes themselves) are barely represented by culture collections. Most of their diversity is found in the vast reservoir of uncultivated microbes present in the ocean and other aquatic environments exposed to light (2–6). Microbial rhodopsins were initially discovered in the halophilic archaeon *Halobacterium salinarum* (7), which contains four paralogous rhodopsin genes coding for different functions: $H^+$ or $Cl^-$ pumping, or light sensing with two different light ranges (8). In haloarchaea, the most common case is that more than one rhodopsin gene is found per genome reaching up to six described for *Haloarcula marismortui* (9). In other aquatic bacteria, the presence of multiple

rhodopsin genes (paralogs) in the genome is much rarer, and only a few marine bacteria, such as *Dokdonia eikasta, Gillisia limnaea,* or *Nonlabens marinus* encode for, in addition to the regular proton-pumping rhodopsin, a sodium-pump and a chloride-pump (10–12). One potential reason for such scarcity is that many pelagic aquatic microbes have streamlined genomes that are depleted in paralog genes. Besides, they tend to be hard to grow and are largely known by metagenomic assemblies that are often incomplete. However, the presence of paralog copies of rhodopsins is interesting, particularly when the sequence is widely divergent since they may provide different functions to the microbe. Hence paralog rhodopsins are a potential source of novel functions of these versatile proteins that have also biotechnological potential for their use in optogenetics (13, 14).

Here we evaluate the genetic repertoire of rhodopsins in a large collection of single-amplified genomes (SAGs), which were generated from a global set of seawater samples from the tropical and subtropical, epipelagic ocean (15). The screening for rhodopsins resulted in approximately 3% of SAGs containing two type-I rhodopsin paralogs within their genomes. While one of them was predicted to be a standard proton pump, the other paralog belonged to a formerly described cluster of DTT-T rhodopsins discovered in contigs assembled from deep-water metagenomes in station ALOHA (16). The presence of this second rhodopsin was found restricted to three taxa known to have small and streamlined genomes- Pelagibacterales, HIMB59, and the Gammaproteobacteria *Pseudothioglobus*. This rhodopsin gene was always found next to a flotillin gene. Metagenomic screening in marine and freshwater samples revealed a broad distribution of these sequences in both types of environments, always associated

with low light intensities. Our results on the protein structure, metagenomic recruitment, and genomic context can help to understand its function in these streamlined microbes.

## RESULTS AND DISCUSSION

### Paralog rhodopsin recovery and diversity

To identify rhodopsin paralogs, we first tried to locate them in the RefSeq database (April 2022) which includes cultivated marine microbes, but none was found, with the known exceptions mentioned above. This was not surprising considering the narrow range of diversity covered by cultivated marine bacteria, particularly of streamlined and abundant microbes. Fortunately, an alternative to monoclonal genomes of marine bacteria is single-amplified genomes (SAGs). Although many of them are incomplete they offer a significant guarantee that they start from a single cell and that the redundant genes found in them are real paralogs. Thus, we used a large collection of marine SAGs, composed of 12,715 partial genomes, which were produced using a random cell selection strategy from the epipelagic seawater samples from tropical and subtropical latitudes of the Atlantic and Pacific Oceans (15). The taxonomic composition of marine prokaryoplankton in this database is consistent with previous 16S rRNA amplicon and shotgun metagenomic studies of marine off-shore waters, indicating that this dataset is representative of the in situ microbial communities (15). We removed SAG assemblies with an estimated completeness lower than 50 %, which is the threshold for classifying them as medium quality (17). Finally, a total of 4,751 SAGs were screened for the presence of type-1 rhodopsin genes in their genomes (see methods). Among them, 63

113 % coded for type-I rhodopsins (**Figure 1A**). As previously described from metagenomic
114 assemblies, most of these sequences were found in microbes from the order
115 Pelagibacterales (#1,619), SAR86 (#386), Flavobacteriales (#258), HIMB59 clade
116 (#191), and *Ca.* Actinomarinales (#139), as well as other less abundant taxa (**Table**
117 **S1**). A small set of SAGs (~3 %) coded for two type-I rhodopsins within their genomes.
118 A close inspection revealed that they all belonged to three taxonomic groups:
119 Pelagibacterales (SAR11), HIMB59 (formerly AEGEAN-169), and the
120 gammaproteobacterial *Pseudothioglobus* **(Figure 1A)**, all typical marine streamlined
121 microbes. To analyze the phylogenetic diversity of the recovered paralogs of type-I
122 microbial rhodopsins, a maximum likelihood phylogenetic tree was constructed,
123 including several reference sequences covering all types of rhodopsins. The
124 phylogenetic tree showed that, although one of the rhodopsin sequences clustered
125 according to their taxonomic classification and into the proton-pump proteorhodopsin
126 family (**Figure 1B**), the sequences of the second rhodopsin paralog, which we named
127 FArhodopsin, clustered on an independent branch (**Figure 1B**).
128

129 An analysis of the protein alignment of the two paralog rhodopsins (**Figure 1C)**
130 indicated that all seemed to be potentially functional. Both paralogs code for the lysine
131 (K) in helix G responsible for the binding of the retinal to the protein via a Schiff base
132 reaction although structural and experimental predictions make it unlikely that one of
133 them bind retinal (see below). In addition to many major differences, including a deletion
134 of 23 aa at the N terminus of the second rhodopsin (FArhodopsin), which might include
135 a leader peptide **(Figure S1)**, we observed differences in the ion-pumping motif. While

one of the paralogs showed the DTE motif (positions 85, 89, and 96 according to the amino acid sequence of bacteriorhodopsin from *Halobacterium salinarum*) identical to the green and blue light proton-pumping proteorhodopsins (**Figure 1C**), in the FArhodopsin we detected three variations. In the case of Pelagibacterales and the *Pseudothioglobus*, the FArhodopsin differed in the last amino acid of the triad, being threonine (DTT) as has been described previously for metagenomic assembled rhodopsins from ALOHA (16) and leucine (DTL), respectively, while in HIMB59, the DTE motif was replaced by aspartate-asparagine-isoleucine (DNI). In addition, the amino acid residue suggested being critical to the spectral tuning of rhodopsins (position 93 in bacteriorhodopsin) also changed. Leucine, methionine, and isoleucine residues are detected commonly in proteorhodopsins absorbing in the green light spectrum, whereas a rhodopsin containing glutamine absorbs in the blue light range. The FArhodopsin sequences of HIMB59 and *Pseudothioglobus* had either the glutamine or leucine, while the Pelagibacterales FArhodopsin coded for threonine at that residue.

## The genomic context of FArhodopsins

The genomic comparison of Pelagibacterales SAGs containing the second rhodopsin against the genomes of isolates HIMB083 (genomospecies Ia.3/V), HTCC7011 (genomospecies Ia.3/I), and HTCC1062 (genomospecies Ia.1/I) (18) indicated that FArhodopsins were always located on a gene cluster of approx. 26 Kb that was absent from the isolate genomes (**Figure 2A**). Although SAGs containing this cluster were phylogenetically distant from each other, as they belonged to different subclades (listed in **Figure 2B**), most of the cluster genes were conserved and syntenic (**Figure 2B**).

159 Despite some genomic rearrangements, the rhodopsin in the cluster always appeared
160 co-localized on the same strand next (16 nucleotides intergenic spacer) to a large gene
161 (1,977 bp) predicted to code for a flotillin. Flotillins are members of a widely conserved
162 superfamily of proteins termed Stomatin, Prohibitin, Flotillin, and HfK/C (SPFH) domain
163 proteins (19). Eukaryotic flotillins have been extensively studied (20) and only later they
164 were found in prokaryotes (21, 22). They are classified as scaffold proteins associated
165 with lipid rafts, lipid microdomains in membranes, which enhance the assembly of raft-
166 associated proteins (23, 24). The gene clusters containing the rhodopsin in HIMB59 and
167 *Pseudothioglobus* were delimited by comparing the genomes against closely related
168 (>90% average nucleotide identity [ANI]) SAGs (the pure culture HIMB59 available was
169 too distant). Regardless of the gain or loss of some genes, synteny was maintained
170 among HIMB59 relatives, despite a wide phylogenetic distance. Even *Pseudothioglobus*
171 (a gammaproteobacterium) genomes conserved the flotillin and the three last genes of
172 the cluster located on the same strand and potentially co-transcribed (**Figure 2B and**
173 **Table S2**). We evaluated the presence of flotillin proteins as well as any other "SPFH"
174 domain-containing proteins in the initial set of marine SAGs (see methods). Among
175 them, nearly 85% were positive for at least one "SPFH" domain-containing hit. A manual
176 inspection resulted in a broad classification of these proteins as HflC and HflK. These
177 proteins form the complex HflKC, involved in the regulation of the protease FtsH not
178 related to the eukaryotic flotillins (25). Otherwise, the flotillin gene was detected mainly
179 in genomes coding for the FArhodopsin (only twelve bona-fide flotillin proteins were
180 detected not belonging to the rhodopsin gene cluster) (**Figure S2**). Phylogenetic
181 classification of detected flotillins against well-characterized eukaryotic and prokaryotic

182 sequences resulted in flotillin homologs from Pelagibacterales, HIMB59, and
183 *Pseudothioglobus* forming a separate cluster (**Figure S2**).

184 The analysis of the genes found in the FArhodopsin cluster (FAR cluster) of the
185 Pelagibacterales SAGs (**Figure 2B**) shows very conserved synteny despite the
186 divergence (ANI) of the overall genomes. The gene cluster is always bordered by long
187 spacers (236 and 219 nucleotides in the AG-390-L03) that are also well conserved.
188 After the spacer, there is a set of ten genes on the same strand and likely also co-
189 transcribed (see below). The next gene is annotated as a sterol-binding-domain of a
190 sterol-carrier protein, likely involved in lipid transport. There is also an inner membrane
191 translocator and a membrane sodium symporter. Thus, a correlation between
192 membrane structure and function is strongly suggested. Only Pelagibacterales and
193 HIMB59 encoded for a cluster of three genes annotated as biotin-dependent
194 carboxyltransferases and carboxylases, that might be involved in the degradation of
195 urea. Other metabolic genes detected included an N,N-dimethylformamidase, involved
196 in the glyoxylate and dicarboxylate metabolism (26), the PxpA-like subunit, involved in
197 the conversion of 5-oxoproline to L-glutamate (27), as well as some uncharacterized
198 alpha/beta hydrolases, iron-dependent alcohol dehydrogenases, and aldolases. In
199 addition, the *Pseudothioglobus* FAR cluster encoded for the transcriptional regulator
200 LuxR and a hybrid sensor histidine kinase/response regulator (genes annotated in the
201 FAR gene cluster are listed in **Table S2**).

202

203 **FArhodopsins in aquatic metagenomes.**

204 The presence of homologs to FArhodopsin in the ALOHA station depth profile has been
205 extensively studied (16). The authors concluded that the DTT-T rhodopsin was present
206 mainly in samples from 200 to 1000 m depth. Thus, to recover more genetic diversity of
207 FArhodopsins we analyzed the assemblies of marine metagenomes, including western
208 Mediterranean Sea depth profiles (28, 29), BATs (30), and global ocean surveys such
209 as Tara Oceans (31) and GEOTRACES (30). A western Mediterranean sample has also
210 been analyzed recently with long-read HiFi sequencing (32) that does not need
211 assembly to recover complete gene clusters. Considering that microbes of the order
212 Pelagibacterales, which are the main contributors of FArhodopsins, can be also found in
213 freshwater ecosystems (e.g. *Fonsibacter*), we included in the analysis samples taken
214 from lakes in North America (33), Europe, and Asia (34–36), including Lake Baikal (37,
215 38), where representatives of the *Pelagibacter* Ia.5 (18, 37) clade were recently found.

216 **Table S3** summarizes the number of rhodopsins type-I, type-III, FArhodopsins, and
217 flotillins recovered in each metagenomic sample. In agreement with what we found in
218 the dataset of marine SAGs, microbial type-I rhodopsins represented the majority of
219 rhodopsin sequences in marine and freshwater assembled contigs, followed by
220 heliorhodopsins. We could recover hundreds of FArhodopsin sequences, including from
221 freshwater datasets (**Table S3**). This analysis showed the uneven recovery of
222 FArhodopsin and flotillin sequences, which could call into question the strong
223 association described previously between FArhodopsin and flotillin. However, a
224 comprehensive study of the assembled contigs containing the rhodopsin revealed that
225 most of them had FArhodopsins at the contig ends, indicating a breakpoint in the
226 assembly, a common occurrence in regions of high microdiversity. Only in the

227 metagenomic sample from the Mediterranean Sea sequenced with PacBio CCS (32),
228 FArhodopsin and flotillin numbers were almost at par (407 and 455 sequences,
229 respectively). All FArhodopsins well covered by the long reads contained the tandem
230 confirming a consistent association. This result illustrates one of the major advantages
231 of long-read metagenomics, allowing to recover complete genes and even operons from
232 the bulk of CCS reads, without relying on metagenomic assembly.

233

234 The phylogeny of all of the recovered FArhodopsin proteins (#912) resulted in two major
235 phylogenetic branches, separating the marine from the majority of the freshwater
236 sequences (**Figure 3A**). All marine metagenomic FArhodopsin-containing assemblies
237 were classified within Pelagibacterales, HIMB59, and *Pseudothioglobus* clades,
238 regardless of depth, or geographical origin (e.g. latitude), indicating that apparently, no
239 other microbes encode FArhodopsin globally. Interestingly, one cluster composed of
240 sequences from deep samples of Lake Baikal (390 m, Russia) and Lake Constance
241 (200 m, Switzerland) grouped within the Pelagibacterales FArhodopsin. A close
242 inspection revealed a well-conserved synteny between these contigs and the marine
243 SAG AG-390-L03 (**Figure 3B**). Based on gene annotation of the surrounding regions,
244 these sequences belonged to the freshwater Ia.5 *Pelagibacter* clade, which was
245 reported for the first time in freshwaters in Lake Baikal (37). Other freshwater
246 metagenomes, on the other hand, showed a less restricted taxonomic distribution of
247 FArhodopsins, with several distinct subclades indicating a higher sequence diversity
248 (**Figure 3A**). Interestingly, protein alignment of these novel freshwater metagenomic
249 sequences revealed that most did not encode for the lysine (K) in helix G responsible

250 for the covalent binding of retinal to the protein, but they showed the amino acid variants
251 glutamine, threonine, leucine, or arginine at this site (**Figure 3A and Table S4**). **Figure**
252 **3B** also shows that despite the divergence, sequences coming from freshwater
253 preserve some genes of the FAR cluster found in the marine FARs. Besides, although
254 the contigs harboring this protein were short hampering their taxonomic classification
255 (as is typical of complex communities), those that were long enough always encoded for
256 a flotillin next to the rhodopsin gene, highlighting the key role the flotillin gene plays in
257 the cluster. Lysine-less rhodopsins (Rh-noK) have been already described for a group
258 of xanthorhodopsins related sequences in marine betaproteobacteria (39) and were
259 shown to be retinal binding and functional proton pumps when the lysine residue was
260 reconstituted by genetic manipulation. Interestingly, Rh-noK was found located in
261 tandem with a functional proteorhodopsin gene, and the authors speculated that both
262 gene products could form a multimer in which Rh-noK could regulate the proton
263 pumping activity of the neighbor (39).

264

265 **Recruitment of metagenomic and metatranscriptomic fragments on the FAR gene**
266 **cluster**

267 To assess the abundance and distribution of this novel group of rhodopsins, we
268 performed metagenomic fragment recruitment analysis by recruiting all the FAR gene
269 clusters in the SAGs of Pelagibacterales (#24), HIMB59 (#69) and *Pseudothioglobus*
270 (#10) taxa to 140 metagenomes from a depth profile in the Mediterranean Sea (29) and
271 the *Tara* Oceans datasets (31). As expected, the three taxa showed differential depth
272 recruitment in the Mediterranean Sea (**Figure 4A**) and the same was found for the *Tara*

273 Oceans datasets containing deeper samples (**Figure S3**). Pelagibacterales are the
274 most abundant microbes in surface layers of the water column, although it is known that
275 this diverse order has an eurybathic distribution, with some clades being detected in the
276 deep chlorophyll maximum (DCM), lower photic (LP) metagenomes, and in meso- and
277 bathypelagic samples (29). However, the Pelagibacterales FAR cluster recruited more
278 in deeper samples such as the DCM, lower photic and bathypelagic samples,
279 represented by 45, 60, 75, 90, 1000, and 2000m depths in the available Mediterranean
280 metagenomes (30) (**Figure 4A**). A similar preference for deeper waters was found
281 globally with the *Tara* Ocean samples (**Figure S3**). Thus, confirming the trend
282 discovered for the DTT-T rhodopsins found by Olson et al at ALOHA (16). In the
283 Mediterranean metagenomes, the maximum recruitment of the cluster was found at the
284 LP (75-90 m), much shallower than the 500 m detected at the more oligotrophic,
285 permanently stratified waters at ALOHA (16). A similar trend was found for the
286 Gammaproteobacteria *Pseudothioglobus*. Contrastingly, HIMB59 FAR (with the DNI
287 motif in the FArhodopsin) recruited the most at the upper photic (15-30 m deep),
288 following closely the recruitment of the full SAGs (data not shown) that were detected
289 almost exclusively in the upper photic. In addition, we could not find a clear association
290 of freshwater FArhodopsins with deeper waters, in Lake Baikal they appeared near the
291 surface and down to 390 m deep but not in a bathypelagic sample (data not shown). In
292 freshwaters, the photic zone is harder to delimit given the diversity of trophic status that
293 characterize lakes and their overall shallower depths.

294

295 Metatranscriptomes were available from the same seawater samples (Western
296 Mediterranean Sea, 15 and 60 m) (40). Their recruitment on SAGs confirmed the
297 transcription of the FAR cluster of Pelagibacterales AG-390-L03 at 60 m (**Figure 4B**
298 **and 4C**) but, as expected, no matching transcripts were found at 15m. The tandem
299 rhodopsin-flotillin was the second most recruited region in the cluster (**Figure 4B**),
300 indicating that these genes are abundantly expressed. The average recruitment value
301 (in reads per kilobase gene per million transcripts – RPKM) of the FAR (0.32 RPKM)
302 was substantially above the average number of transcripts aligning to the genome (0.2
303 RPKM, with the rRNA ribosomal operon excluded). The transcriptomic read coverage
304 indicates clearly that both FArhodopsin and flotillin (at least) are cotranscribed in a
305 single polycistronic mRNA despite the 16 nucleotides separating both genes (**Figure**
306 **4C**). Several mRNA reads containing the spacer could be detected (data not shown).

307 **Predicted structure of the retinal binding pocket**

308 The most notable characteristic of microbial rhodopsins is their capability to respond to
309 light through the use of the retinal cofactor. However, when (16) expressed in
310 *Escherichia coli* a rhodopsin gene containing the DTT motif, they detect no binding of
311 retinal, judging from the pellet pigmentation of centrifuged cells nor change in pH upon
312 illumination. The lack of function or pigmentation in *E. coli* can be associated with a lack
313 of proper membrane insertion into a lipid raft due to the lack of the flotillin-associated
314 protein. Interestingly, Olson *et. al.* (16) modified a functional DTE-Q sequence to DTT-T,
315 which resulted in the loss of function.

316 We used AlphaFold (41) to predict the structure of marine FArhodopsin to examine its
317 retinal binding pocket. The predicted structure showed high similarity to that of the

318 closest homolog with an experimentally determined high-resolution structure (PDB:
319 4jq6, blue proteorhodopsin from Med12, a predicted protein from a metagenomic
320 sample from the Mediterranean Sea at 12 m deep (42)), which allowed us to adapt the
321 retinal position to the predicted structure. However, when analyzing amino acids that
322 are highly conserved among Pelagibacterales FArhodopsins (present in more than 99%
323 of cases), we observed that four of these residues would intersect the retinal in its all-
324 trans conformation (**Figure 5A**). A similar observation was made in other groups of
325 marine FArhodopsins (**Figure 5B and 5C**). In addition, as mentioned above, most of the
326 freshwater FArhodopsins lack the lysine that is used to bind retinal. These findings
327 suggest that all FArhodopsins may be retinal-less like those previously described in the
328 Halobacteriaceae (43) or betaproteobacteria Rh-noK (39). Their seven transmembrane
329 helices may have evolved to perceive other stimuli, similar to G protein-coupled
330 receptors (G-PCRs), which have adapted for responding to both light and chemical
331 activation or they could be involved in either regulation or membrane microdomain
332 organization as shown for some G-PCRs in the animal retina (44). However, it is
333 important to note that there is no direct evidence that the AlphaFold prediction is
334 precisely correct. Recently, an unwound helix in a structure of a channelrhodopsin
335 called ChRmine was not predicted by AlphaFold (45). It is possible that another
336 significant distortion from the classic rhodopsin fold could still allow for binding of the
337 retinal, or that FArhodopsins may bind other carotenoid derivatives to absorb light.

338

339 **SYNTHESIS**

340 The discovery, two decades ago (2), of rhodopsin sequences in marine
341 prokaryoplankton communities resulted in a significant leap forward in the
342 understanding of the biology of these microorganisms and their interaction with the
343 ecosystem. In the photic zone, the proton-pumping activity of rhodopsins may provide
344 "cheap" solar energy (46). However, rhodopsins can perform a wide variety of other
345 functions (47, 48). Furthermore, the presence of paralogs with very divergent activities
346 and even sequences has been known since the first microbial rhodopsins were
347 described in halophilic archaea, and they are intensively looked after since they reveal
348 novel physio-ecological roles and can be of potential application for optogenetics. In this
349 work, we have analyzed the distribution and the presence of rhodopsin paralogs in a
350 large dataset of marine SAGs (15) that provide strong evidence that the detected genes
351 come from the same genome (15). We tried with the meager representation of cultured
352 microbes but, unsurprisingly, could not find any examples of genomes with two
353 rhodopsins. MAGs, on the other hand, are a composite of genomic fragments, and the
354 presence of contigs belonging to different organisms (chimerism) may interfere with the
355 analysis.

356

357 Unsurprisingly, only a small subset of SAGs coded for a rhodopsin paralog, most of
358 which belonged to well-known streamlined lineages, such as Pelagibacterales (49–52),
359 *Ca*. Actinomarinales (53, 54), and SAR86 (55, 56). These microbes are characterized
360 by having a small genome size, a higher coding density, and a reduced set of paralogs,
361 to cope with limited nutrient and energy availability in the oligotrophic open ocean (57).
362 Nevertheless, we detected two copies of the rhodopsin gene in the genomes of three

363 widespread streamlined taxa, Pelagibacterales, HIMB59 and the Gammaproteobacteria
364 *Pseudothioglobus*. One paralog was a typical proteorhodopsin proton pump, while the
365 other affiliated with a novel clade described among assembled contigs in ALOHA
366 metagenomes at depths of 200 to 1000 m (16). The access to the genomic context,
367 which was not possible previously, showed a consistent association with a conspicuous
368 flotillin gene forming a single mRNA transcript. Flotillins are members of a superfamily
369 of proteins termed "SPFH" (Stomatin, Prohibitin, Flotillin, and HfK/C) (19), detected for
370 the first time in the cytoplasmic membrane of eukaryotes (20). These proteins are
371 involved in the formation of microdomains (also known as lipid rafts) in the plasma
372 membrane of all mammalian cells (20). Its biological role is not well understood, but it is
373 believed that they act as scaffold proteins, enhancing the recruitment of raft-associated
374 proteins to these microdomains to be active, as well as to facilitate their interaction and
375 oligomerization with other proteins (23). Flotillin sequences have been also detected in
376 bacteria. Specifically in *Bacillus subtilis*, which encodes for two subunits, YuaG and
377 YqfA (renamed as FloT and FloA) (23). Mutants lacking both genes showed reduced
378 sporulation efficiency and changes in the biofilm formation (21, 22, 58). Remarkably, our
379 results indicated that the flotillin encoded in the FAR cluster was the only copy present
380 in the genome, and therefore, its presence might be essential for the formation of
381 specific microdomains in these organisms. In *B. subtilis* (and other bacteria) flotillin
382 appears in lipid rafts that concentrate the phospholipid cardiolipin and cyclic (hopanoids)
383 or non-cyclic isoprenoids, such as those derived from carotenoids (23). The genomes of
384 Pelagibacterales, HIMB59, and *Pseudothioglobus* encode for a cardiolipin synthase. For
385 that reason, the presence of lipid rafts, despite not being demonstrated "*in vivo*", seems

386 plausible. We would like to hypothesize that the flotillin detected in the genomic island
387 might be critical for the function of the FAR cluster genes. Many FAR genes annotated
388 were classified as membrane transporters or signal transduction/regulators, proteins
389 typically enriched in flotillin-associated microdomains (21, 24).

390 In the Mediterranean, FArhodopsin genes were abundant in metagenomes coming from
391 the lower photic zone, with a maximum at the deep-chlorophyll-maximum (45-60 m
392 deep). In contrast, FArhodopsin genes peaked in deeper waters (500 m) at ALOHA
393 station in the Pacific Ocean. This fact, together with the lack of *in vitro* activity and even
394 retinal binding of some ALOHA representatives cloned in *E. coli* suggests that these
395 rhodopsins might be retinal-less or even blind rhodopsins (not using light). Furthermore,
396 most freshwater FArhodopsins do not have the lysine residue needed for retinal binding.
397 However, a more in-depth screening of several metagenomic datasets indicates that
398 they tend to associate with the photic (or twilight) zone, but not with bathypelagic
399 waters. In addition, we have always detected a standard retinal-binding proton pumping
400 proteorhodopsin present elsewhere in the genome, which indicates a clear light-
401 interactive lifestyle (at least periodically). A possible explanation would be that flotillins
402 enhance the formation of a heterocomplex of DTE and FArhodopsins in lipid rafts. This
403 kind of rhodopsin heterocomplexes has been described in fungi (59, 60), on which a
404 classic blue or green-absorbing rhodopsin cyclase (a kind of rhodopsin linked to a type
405 III guanylyl cyclase) interacts with another rhodopsin cyclase, but with an absorption in
406 the near-infrared spectrum. Thus, there is a chance that the FArhodopsin interacts with
407 the proton-pump proteorhodopsin, present in the same genome, in a heterocomplex,
408 modifying the spectral tunning to accommodate ion transport to deeper waters.

409 The presence of FArhodopsins in freshwater metagenomes came as a surprise. They
410 are much more diverse at the sequence level than marine FArhodopsins and are
411 present in a much wider taxonomic array of prokaryotes. The association with flotillin
412 was also detected in all cases in which contigs were long enough suggesting similar
413 functional properties. In one case, we found a FAR cluster syntenic with a marine
414 *Pelagibacter* genome in a distant relative from Lake Baikal. The high similarity of the
415 FAR cluster between marine and freshwater Pelagibacterales might indicate marine-to-
416 freshwater transitions of the FAR gene cluster (37).

417

418 There is always the possibility that FArhodopsins act as sensors of other stimuli (as
419 hypothesized for retinal-less rhodopsins of haloarchaea (43) or Rh-noK (39), but the
420 lack of genes coding for transducers in the FAR cluster and the association with the
421 photic zone in many metagenomes indicates otherwise. We hypothesize that
422 FArhodopsins may act as complements to the *bona fide* proton pump, by increasing its
423 performance under the dimmer light conditions prevalent in the lower photic zone. The
424 streamlined nature of genomes containing FArhodopsin and the high conservation
425 detected in the FAR gene cluster, even across marine-freshwater transitions, point to an
426 important role in the physiology and ecology of some of the most abundant aquatic
427 microbes. Perhaps extending the “solar energy” strategy to deeper waters where a large
428 part of the water column microbiome activity takes place.

429

430

431

432

433 **METHODS**

434 **Screening of rhodopsin proteins from single-amplified genomes and**
435 **metagenomic datasets.** To evaluate the presence and the number of rhodopsin
436 sequences in the euphotic marine microbiome, a compendium of 12,715 single-
437 amplified genomes (SAGs) were downloaded from NCBI under BioProject PRJEB33281
438 (15) and subjected to study. Prior to the analysis, the degree of completeness and
439 contamination of SAGs were estimated using CheckM v1.1.2 (61), and SAGs with > 50
440 % completeness and < 5 % contamination were kept. Taxonomic classification of SAGs
441 was performed using the GTDB-Tk v 2.1.0 tool (62) using the Genome Taxonomy
442 Database (GTDB) release R207 (63).

443 Metagenomes from different marine and freshwater datasets were used to retrieve
444 similar protein sequences to FArhodopsins, as well as the contigs containing them.
445 Marine datasets included metagenomic assemblies and PacBio CCS reads from a local
446 time series in the Mediterranean Sea (28, 29), Hawaii Ocean Time-series (HOTs) and
447 Bermuda Atlantic Time-series Study (BATS) (30), as well as from large oceanic
448 expeditions, such as Tara Oceans (31) and GEOTRACES (30). Freshwater datasets
449 represented metagenomic assemblies from Lake Baikal (37, 38), Spanish reservoirs
450 (34), and a compendium of 17 different freshwater lakes located in Europe and Asia
451 (36). Assembled contigs from marine and freshwater metagenomes longer or equal to
452 5Kb were assigned family- or genus-level classification if at least 50 % of the genes

453 shared the same best-hit taxonomy. Contigs failing this threshold were classified at the
454 level of phylum.

455 Recovery of rhodopsin proteins was achieved using a custom and curated database of
456 rhodopsin sequences, comprising all the families described so far from both type-I and
457 type-III rhodopsins. Hidden Markov models (HMMs) of these families were built using
458 hmmbuild (64) after a previous alignment by muscle (65). Putative protein sequences
459 were then screened using hmmscan (64). Only sequence hits >= 200 aa, with an evalue
460 < 1e-10 and a bitscore >= 75 were kept.

461 **Phylogenetic classification of microbial rhodopsins.** A maximum likelihood
462 phylogenetic tree was built with all the type-I microbial rhodopsins recovered from the
463 pool of marine SAGs, as well as from the metagenomic sequences retrieved from
464 marine and freshwater datasets. As references, we included well-characterized
465 rhodopsin proteins with several functions described so far ($H^+$, $Na^+$, $Cl^-$ and sensory
466 rhodopsins). Sequences were first aligned with muscle (65), and the resulting alignment
467 trimmed with trimal (66) to remove regions with gaps in more than 80% of the
468 sequences. The phylogenetic tree was built with iqtree (67) with 5000 ultrafast
469 bootstraps and the -m MFP option to find the best model that fitted to our data.

470 **Determination of gene clusters carrying FArhodopsins**. In order to determine the
471 location of the FArhodopsin and the absence of it in other genomes from the same
472 taxonomic group, Pelagibacterales SAGs containing the second rhodopsin were
473 compared using blastp (68) against the isolate genomes HIMB083 (genomospecies
474 Ia.3/V), HTCC7011 (genomospecies Ia.3/I), and HTCC1062 (genomospecies Ia.1/I).
475 Genomic regions were then extracted and genes predicted for further analyses. In a

476 similar approach, the genomic islands containing the rhodopsin in HIMB59 and the
477 gammaproteobacterial *Pseudothioglobus* were determined and extracted by comparing
478 the genomes against closely related (>90% average nucleotide identity [ANI]) SAGs.

479 **Taxonomic and functional annotation of rhodopsin-carrying contigs.** Prodigal
480 v2.6.3 (69) was used to predict genes from contigs retrieved from the individual SAGs,
481 as well as from the metagenomic assemblies. Predicted protein-encoded genes were
482 taxonomically and functionally annotated against the NCBI NR database using
483 DIAMOND 0.9.15 (70) and against COG (71) and TIGRFAM (72) using HMMscan v3.3
484 (64). Domains within proteins were predicted using InterPro (73). In order to determine
485 the location of predicted proteins within the gene cluster, transmembrane domains and
486 signal peptides were identified with DeepTMHMM (74) and SignalP v6.0 (75),
487 respectively.

488 **Retrieval of SPFH proteins and Phylogenetic tree of flotillin sequences.** Domain
489 sequences of the SPFH family of proteins were downloaded from the Pfam database
490 (76) and then used to search for candidates within SAGs with hmmscan. The resulting
491 flotillin sequences were then phylogenetically classified. Reference sequences from
492 eukaryotic flotillins Flot-1 and Flot-2 (*Bos Taurus*, *Homo sapiens*, and *Rattus*
493 *norvergicus*) together with well-characterized sequences from *Bacillus, Staphylococcus*
494 and *E. coli* were downloaded from the Uniprot database. We also included the closest
495 hits determined by a blastp search against the NCBI protein database. Sequences were
496 first aligned with muscle, and the resulting alignment trimmed with trimal to remove
497 regions with gaps in more than 80% of the sequences. A maximum-likelihood

498 phylogenetic tree was built with iqtree with 5000 ultrafast bootstraps and the -m MFP
499 option to find the best model that fitted to our data.

500 **Metagenomic and metatranscriptomic read recruitment.** Illumina raw reads from
501 Tara Oceans metagenomes and metatranscriptomes were downloaded from the
502 European Nucleotide Archive (ENA) following the accession numbers PRJEB1787 and
503 PRJEB6608, respectively. Eight additional samples covering a metagenomic depth
504 profile from the Mediterranean Sea during thermal stratification (summer -- 15, 30, 45,
505 60, 75, 90, 1000 and 2000m) were downloaded from NCBI BioProjects PRJNA352798
506 and PRJNA257723. Raw reads were trimmed with Trimmomatic v0.39 (77). Coverage
507 values were calculated by read alignment (in subsets of 20 million reads) against the
508 genomic islands containing the FArhodopsin using BLASTN v2.9.0 (90 % identity, > 50
509 bp alignment). At least 70% of the length was required to be covered by reads to be
510 considered a recruit. Reads were normalized by the size of the contig in Kb and by the
511 size of the metagenome in Gb (RPKGs). Transcripts from metatranscriptomic samples
512 were aligned using BLASTN (98 % identity, > 50 bp alignment) and normalized as
513 transcripts per kilobase of gene per million reads (RPKM).

514 **In-silico determination of protein structures.** We used AlphaFold (41), a protein
515 structure prediction server hosted on Google Colab (78), to predict the structures of
516 three FArhodopsins (AG-390-L03 of Pelagibacterales, AG-337-N20 of HIMB59, and
517 AG-313-D08 of *Pseudothioglobus*) using standard parameters. We then used PyMOL
518 (https://pymol.org/2/) to align these structures to the structure of a blue-light absorbing
519 proteorhodopsin from Med12 (PDB ID 4JQ6). Since the FArhodopsins and
520 proteorhodopsin shared high overall similarity and the positions of the lysine residues,

521 which are putative binding residues for retinal coincided, we transferred the all-trans
522 retinal molecule from proteorhodopsin (PDB ID 4JQ6) to the FArhodopsins for further
523 analysis of the retinal binding pocket.

524

525 **ACKNOWLEDGEMENTS**

526 This work was supported by grant “FLEX3GEN” PID2020-118052GB-I00 (cofounded
527 with FEDER funds) from the Spanish Ministerio de Economía, Industria y
528 Competitividad to ML-P and FR-V. JMH-M was supported with a PhD fellowship from
529 Margarita Salas program, cofounded by the Spanish Ministerio de Universidades and
530 the European Union -- Next Generation EU (2021/PER/00020). RS was supported by
531 the Simons Foundation grants 510023 and 827839. The work of AA was supported by
532 funding of the German Research Foundation to Dr. Tobias Moser via the Multiscale
533 Bioimaging - Cluster of Excellence (EXC 2067/1-390729940).

534

535 **AUTHORS’ CONTRIBUTIONS**

536 JMH-M conceived the study. JMH-M, ML-P, AA, EP, KK and FR-V analysed the data.
537 JMH-M, ML-P, AA, VG, RS and FR-V contributed to write the manuscript.

538

539 **COMPETING INTERESTS**

540 The authors declare that they have no competing interests.

541

542

543

544 **REFERENCES**


545 1. Béjà O, Lanyi JK. 2014. Nature's toolkit for microbial rhodopsin ion pumps. Proc
546 Natl Acad Sci 111:6538–6539.

547 2. Béjà O, Suzuki MT, Hadd A, Nguyen LP, Spudich JL, Spudich EN, Delong EF.
548 2000. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea.
549 Science (80- ) 1902:1902–1907.

550 3. Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin
551 phototrophy in the ocean. Nature 411:786–789.

552 4. Béjà O, Pinhassi J, Spudich JL. 2013. Proteorhodopsins: Widespread Microbial
553 Light-Driven Proton Pumps. Encycl Biodivers Second Ed 280–285.

554 5. Inoue K, Kato Y, Kandori H. 2015. Light-driven ion-translocating rhodopsins in
555 marine bacteria. Trends Microbiol 23:91–98.

556 6. Gómez-Consarnau L, Raven JA, Levine NM, Cutter LS, Wang D, Seegers B,
557 Arístegui J, Fuhrman JA, Gasol JM, Sañudo-Wilhelmy SA. 2019. Microbial
558 rhodopsins are major contributors to the solar energy captured in the sea. Sci Adv
559 5:8855–8862.

560 7. Oesterhelt D, Stoeckenius W. 1971. Rhodopsin-like Protein from the Purple
561 Membrane of Halobacterium halobium. Nat New Biol 233:149–152.

562 8. Spudich JL, Bogomolni RA. 1984. Mechanism of colour discrimination by a

563 bacterial sensory rhodopsin. Nat 1984 3125994 312:509–513.

564 9. Fu HY, Lin YC, Chang YN, Tseng H, Huang CC, Liu KC, Huang CS, Su CW,
565 Weng RR, Lee YY, Ng WV, Yang CS. 2010. A novel six-rhodopsin system in a
566 single archaeon. J Bacteriol 192:5866–5873.

567 10. Inoue K, Ono H, Abe-Yoshizumi R, Yoshizawa S, Ito H, Kogure K, Kandori H.
568 2013. A light-driven sodium ion pump in marine bacteria. Nat Commun 4:1678.

569 11. Riedel T, Held B, Nolan M, Lucas S, Lapidus A, Tice H, Del Rio TG, Cheng JF,
570 Han C, Tapia R, Goodwin LA, Pitluck S, Liolios K, Mavromatis K, Pagani I,
571 Ivanova N, Mikhailova N, Pati A, Chen A, Palaniappan K, Land M, Rohde M,
572 Tindall BJ, Detter JC, Göker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P,
573 Kyrpides NC, Klenk HP, Woyke T. 2012. Genome sequence of the Antarctic
574 rhodopsins-containing flavobacterium Gillisia limnaea type strain (R-8282T).
575 Stand Genomic Sci 7:107–119.

576 12. Yoshizawa S, Kumagai Y, Kim H, Ogura Y, Hayashi T, Iwasaki W, DeLong EF,
577 Kogure K. 2014. Functional characterization of flavobacteria rhodopsins reveals a
578 unique class of light-driven chloride pump in bacteria. Proc Natl Acad Sci U S A
579 111:6732–6737.

580 13. Govorunova EG, Sineshchekov OA, Li H, Spudich JL. 2017. Microbial
581 Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications. Annu Rev
582 Biochem 86:845–872.

583 14. Shevchenko V, Mager T, Kovalev K, Polovinkin V, Alekseev A, Juettner J,
584 Chizhov I, Bamann C, Vavourakis C, Ghai R, Gushchin I, Borshchevskiy V,

585 Rogachev A, Melnikov I, Popov A, Balandin T, Rodriguez-Valera F, Manstein
586 Dietmar J, Bueldt G, Bamberg E, Gordeliy V. 2017. Inward H+ pump
587 xenorhodopsin: Mechanism and alternative optogenetic approach. Sci Adv
588 3:e1603187.

589 15. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton
590 NJ, Burkart MD, La Clair JJ, Chisholm SW, Stepanauskas R. 2019. Charting the
591 Complexity of the Marine Microbiome through Single-Cell Genomics. Cell
592 179:1623-1635.e11.

593 16. Olson DK, Yoshizawa S, Boeuf D, Iwasaki W, DeLong EF. 2018. Proteorhodopsin
594 variability and distribution in the North Pacific Subtropical Gyre. ISME J
595 https://doi.org/10.1038/s41396-018-0074-4.

596 17. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy
597 TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN,
598 Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P,
599 Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO,
600 Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S,
601 Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon
602 KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke
603 C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF,
604 Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified
605 genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria
606 and archaea. Nat Biotechnol. Nature Publishing Group.

607 18. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-
608 Garcia JJ, Gomez ML, Fornas O, Martinez-Garcia M, López-Pérez M. 2020.
609 Ecogenomics of the SAR11 clade. Environ Microbiol 22:1748–1763.

610 19. Browman DT, Hoegg MB, Robbins SM. 2007. The SPFH domain-containing
611 proteins: more than lipid raft markers. Trends Cell Biol 17:394–402.

612 20. Otto GP, Nichols BJ. 2011. The roles of flotillin microdomains – endocytosis and
613 beyond. J Cell Sci 124:3933–3940.

614 21. Donovan C, Bramkamp M. 2009. Characterization and subcellular localization of a
615 bacterial flotillin homologue. Microbiology 155:1786–1799.

616 22. López D, Kolter R. 2010. Functional microdomains in bacterial membranes.
617 Genes Dev 24:1893–1902.

618 23. Bramkamp M, Lopez D. 2015. Exploring the Existence of Lipid Rafts in Bacteria.
619 Microbiol Mol Biol Rev 79:81–100.

620 24. Bach JN, Bramkamp M. 2013. Flotillins functionally organize the bacterial
621 membrane. Mol Microbiol 88:1205–1217.

622 25. Bandyopadhyay K, Parua PK, Datta AB, Parrack P. 2010. Escherichia coli HflK
623 and HflC can individually inhibit the HflB (FtsH)-mediated proteolysis of λCII in
624 vitro. Arch Biochem Biophys 501:239–243.

625 26. SCHÄR H -P, HOLZMANN W, RAMOS TOMBO GM, GHISALBA O. 1986.
626 Purification and characterization of N,N-dimethylformamidase from Pseudomonas
627 DMF 3/3. Eur J Biochem 158:469–475.

628 27. Oke M, Oni O, Bello R, Samuel-Omoyajowo K, Senbadejo T. 2019. Structure–
629 function relationships of the 5-oxoprolinase subunit A: Guiding biological sciences
630 students down the path less traveled. Biochem Mol Biol Educ 47:620–631.

631 28. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M,
632 Rodriguez-Valera F. 2017. Genome diversity of marine phages recovered from
633 Mediterranean metagenomes: Size matters. PLoS Genet 13:e1007018.

634 29. Haro-Moreno JM, López-Pérez M, de la Torre JR, Picazo A, Camacho A,
635 Rodriguez-Valera F. 2018. Fine metagenomic profile of the Mediterranean
636 stratified and mixed water columns revealed by assembly and recruitment.
637 Microbiome 6:128.

638 30. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL,
639 Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D,
640 Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Chisholm SW. 2018.
641 Data descriptor: Marine microbial metagenomes sampled across space and time.
642 Sci Data 5:180176.

643 31. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G,
644 Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI,
645 Cruaud C, D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F,
646 Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M,
647 Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S,
648 Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O,
649 Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J,

650 Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton.
651 Structure and function of the global ocean microbiome. Science 348:1261359.

652 32. Haro-Moreno JM, López-Pérez M, Rodriguez-Valera F. 2021. Enhanced
653 Recovery of Microbial Genes and Genomes From a Marine Water Column Using
654 Long-Read Metagenomics. Front Microbiol 0:2410.

655 33. Garner RE, Kraemer SA, Onana VE, Huot Y, Walsh DA. 2022. Genome-resolved
656 community structure and function of freshwater bacteria at a continental scale.
657 bioRxiv 2022.08.12.503676.

658 34. Cabello-Yeves PJ, Haro-Moreno JM, Martin-Cuadrado AB, Ghai R, Picazo A,
659 Camacho A, Rodriguez-Valera F. 2017. Novel Synechococcus genomes
660 reconstructed from freshwater reservoirs. Front Microbiol 8.

661 35. Cabello-Yeves PJ, Callieri C, Picazo A, Mehrshad M, Haro-Moreno JM, Roda-
662 Garcia JJ, Dzhembekova N, Slabakova V, Slabakova N, Moncheva S, Rodriguez-
663 Valera F. 2021. The microbiome of the Black Sea water column analyzed by
664 shotgun and genome centric metagenomics. Environ Microbiome 16:5.

665 36. Chiriac M-C, Bulzu P-A, Andrei A-S, Okazaki Y, Nakano S, Haber M, Kavagutti
666 VS, Layoun P, Ghai R, Salcher MM. 2022. Ecogenomics sheds light on diverse
667 lifestyle strategies in freshwater CPR. Microbiome 10:84.

668 37. Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov
669 V V, Rodriguez-Valera F. 2018. Genomes of novel microbial lineages assembled
670 from the sub-ice waters of Lake Baikal. Appl Environ Microbiol 84.

671 38. Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko M V, Ivanov VG, Ghai
672 R, Rodriguez-Valera F. 2020. Microbiome of the deep Lake Baikal, a unique oxic
673 bathypelagic habitat. Limnol Ocean 65:1471–1488.

674 39. Yamauchi Y, Konno M, Yamada D, Yura K, Inoue K, Béjà O, Kandori H. 2019.
675 Engineered Functional Recovery of Microbial Rhodopsin Without Retinal-Binding
676 Lysine. Photochem Photobiol 95:1116–1121.

677 40. López-Pérez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. 2018.
678 Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled
679 from metagenomes. Environ Microbiol https://doi.org/10.1111/1462-2920.14462.

680 41. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O,
681 Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl
682 SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T,
683 Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M,
684 Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K,
685 Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with
686 AlphaFold. Nature 596:583–589.

687 42. Ran T, Ozorowski G, Gao Y, Sineshchekov OA, Wang W, Spudich JL, Luecke H.
688 2013. Cross-protomer interaction with the photoactive site in oligomeric
689 proteorhodopsin complexes. Acta Crystallogr Sect D Biol Crystallogr 69:1965–
690 1980.

691 43. Becker EA, Yao AI, Seitzer PM, Kind T, Wang T, Eigenheer R, Shao KSY, Yarov-
692 Yarovoy V, Facciotti MT. 2016. A Large and Phylogenetically Diverse Class of

693 Type 1 Opsins Lacking a Canonical Retinal Binding Site. PLoS One 11:e0156543.

694 44. Goren MA, Morizumi T, Menon I, Joseph JS, Dittman JS, Cherezov V, Stevens
695 RC, Ernst OP, Menon AK. 2014. Constitutive phospholipid scramblase activity of
696 a G protein-coupled receptor. Nat Commun 2014 51 5:1–10.

697 45. Kishi KE, Kim YS, Fukuda M, Inoue M, Kusakizako T, Wang PY, Ramakrishnan
698 C, Byrne EFX, Thadhani E, Paggi JM, Matsui TE, Yamashita K, Nagata T, Konno
699 M, Quirin S, Lo M, Benster T, Uemura T, Liu K, Shibata M, Nomura N, Iwata S,
700 Nureki O, Dror RO, Inoue K, Deisseroth K, Kato HE. 2022. Structural basis for
701 channel conduction in the pump-like channelrhodopsin ChRmine. Cell 185:672-
702 689.e23.

703 46. Munson-McGee JH, Lindsay MR, Sintes E, Brown JM, D'Angelo T, Brown J,
704 Lubelczyk LC, Tomko P, Emerson D, Orcutt BN, Poulton NJ, Herndl GJ,
705 Stepanauskas R. 2022. Decoupling of respiration rates and abundance in marine
706 prokaryoplankton. Nat 2022 6127941 612:764–770.

707 47. Pinhassi J, DeLong EF, Béjà O, González JM, Pedrós-Alió C. 2016. Marine
708 Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology,
709 and Ecology. Microbiol Mol Biol Rev 80:929–54.

710 48. Gushchin I, Gordeliy V. 2018. Microbial Rhodopsins. Subcell Biochem 87:19–56.

711 49. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. 1990. Genetic diversity in
712 Sargasso Sea bacterioplankton. Nature 345:60–63.

713 50. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads

714 J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur
715 EJ. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science
716 (80- ) 309:1242–5.

717 51. Giovannoni SJ. 2017. SAR11 Bacteria: The Most Abundant Plankton in the
718 Oceans. Ann Rev Mar Sci 9:231–255.

719 52. López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-
720 Valera F. 2020. The Evolutionary Success of the Marine Bacterium SAR11
721 Analyzed through a Metagenomic Perspective. mSystems 5:605–625.

722 53. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. 2013.
723 Metagenomics uncovers a new group of low GC and ultra-small marine
724 Actinobacteria. Sci Rep 3:2471.

725 54. López-Pérez M, Haro-Moreno JM, Iranzo J, Rodriguez-Valera F. 2020. Genomes
726 of the “Candidatus Actinomarinales” Order: Highly Streamlined Marine Epipelagic
727 Actinobacteria. mSystems 5:e01041-20.

728 55. Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Alexander Richter R, Valas R,
729 Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, Halpern AL, Lasken RS,
730 Nealson K, Friedman R, Craig Venter J. 2012. Genomic insights to SAR86, an
731 abundant and uncultivated marine bacterial lineage. ISME J 6:1186–1199.

732 56. Hoarfrost A, Nayfach S, Ladau J, Yooseph S, Arnosti C, Dupont CL, Pollard KS.
733 2020. Global ecotypes in the ubiquitous marine clade SAR86. ISME J 14:178–
734 188.

735 57. Giovannoni SJ, Cameron Thrash J, Temperton B. 2014. Implications of
736 streamlining theory for microbial ecology. ISME J 8:1553–1565.

737 58. Dempwolff F, Möller HM, Graumanna PL. 2012. Synthetic motility and cell shape
738 defects associated with deletions of flotillin/reggie paralogs in Bacillus subtilis and
739 interplay of these proteins with NfeD proteins. J Bacteriol 194:4652–4661.

740 59. Broser M, Spreen A, Konold PE, Peter E, Adam S, Borin V, Schapiro I, Seifert R,
741 Kennis JTM, Bernal Sierra YA, Hegemann P. 2020. NeoR, a near-infrared
742 absorbing rhodopsin. Nat Commun 2020 111 11:1–10.

743 60. Rozenberg A, Inoue K, Kandori H, Béjà O. 2021. Microbial Rhodopsins: The Last
744 Two Decades. https://doi.org/101146/annurev-micro-031721-020452 75:427–447.

745 61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM:
746 assessing the quality of microbial genomes recovered from isolates, single cells,
747 and metagenomes. Genome Res 25:1043–55.

748 62. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to
749 classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–
750 1927.

751 63. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P.
752 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat
753 Biotechnol 38:1079–1086.

754 64. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence
755 similarity searching. Nucleic Acids Res 39:W29–W37.

756 65. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and
757 high throughput. Nucleic Acids Res 32:1792–1797.

758 66. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for
759 automated alignment trimming in large-scale phylogenetic analyses.
760 Bioinformatics 25:1972–1973.

761 67. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and
762 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol
763 Biol Evol 32:268–274.

764 68. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.
765 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database
766 search programs. Nucleic Acids Res 25:3389–3402.

767 69. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010.
768 Prodigal: prokaryotic gene recognition and translation initiation site identification.
769 BMC Bioinformatics 11:119.

770 70. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using
771 DIAMOND. Nat Methods 12:59–60.

772 71. Tatusov RL, Natale DA, Garkavtsev I V, Tatusova TA, Shankavaram UT, Rao BS,
773 Kiryutin B, Galperin MY, Fedorova ND, Koonin E V. 2001. The COG database:
774 new developments in phylogenetic classification of proteins from complete
775 genomes. Nucleic Acids Res 29:22–8.

776 72. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001.

777 TIGRFAMs: a protein family resource for the functional identification of proteins.
778 Nucleic Acids Res 29:41–43.

779 73. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das
780 U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E,
781 Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C,
782 McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF,
783 Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH,
784 Yeats C. 2009. InterPro: the integrative protein signature database. Nucleic Acids
785 Res 37:D211–D215.

786 74. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P,
787 Nielsen H, Krogh A, Winther O. 2022. DeepTMHMM predicts alpha and beta
788 transmembrane proteins using deep neural networks. bioRxiv
789 2022.04.08.487609.

790 75. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos
791 KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all
792 five types of signal peptides using protein language models. Nat Biotechnol 2022
793 407 40:1023–1025.

794 76. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL,
795 Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam:
796 The protein families database in 2021. Nucleic Acids Res 49:D412–D419.

797 77. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
798 sequence data. Bioinformatics 30:2114–2120.

78. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. Nat Methods 2022 196 19:679–682.

## FIGURE LEGENDS

**Figure 1. A.** Pie charts indicating the percentage of single-amplified genomes (SAGs) in which two, one or no type-I rhodopsins were detected. Only those SAGs (from a large database collected from the tropical and subtropical euphotic ocean, Pachiadaki et al., 2019) that met the established quality criteria of ≥50% completeness and ≤5% contamination, i.e., medium to high quality draft genomes were included in the analysis. For those SAGs containing 2 rhodopsin genes, a second pie chart indicates their taxonomic affiliation **B.** Maximum likelihood phylogenetic tree of rhodopsin genes detected in the marine SAGs, together with several rhodopsin references covering all types of microbial rhodopsins. Branches have been colored for SAGs harboring two rhodopsins according to their taxonomy (blue, Pelagibacterales; purple, HIMB59 and green, *Pseudothioglobus*). Sequences belonging to the new clade of rhodopsins (Flotillin-associated rhodopsin; FArhodopsin) are highlighted in red. Branches within color background contain type-I rhodopsins from the SAG dataset. The yellow circles correspond to the SAGs containing the two type-I rhodopsin paralogs. **C)** Protein sequence alignment of Helix C and G residues for the two type I rhodopsins. Two reference proteorhodopsin proton pumps at the top. Residues involved in rhodopsin

821 function (e.g. ion translocation) and wavelength absorption are highlighted in blue and
822 orange, respectively. The lysine (K) highlighted in red (helix G) is responsible for the
823 covalent binding of retinal. Dark and light grey regions represent conserved residues
824 within the alignment.

825 **Figure 2. A)** Genomic location of the FArhodopsin and its neighboring genes of
826 Pelagibacterales SAG AG-390-L03 in relation to the reference genomes HTCC1002,
827 HTCC7211, and HIMB083. Genes colored in yellow in the upper alignment represent
828 the (flotillin-associated (FAR) gene cluster. The FArhodopsin and the flotillin are colored
829 in red and blue, respectively. **B)** Genomic alignment (amino acids) of the flotillin-
830 associated rhodopsin gene cluster found in Pelagibacterales, HIMB59 and
831 *Pseudothioglobus* genomes. The ANI value for whole genome is indicated. The region
832 containing the FArhodopsin and flotillin genes highlighted in a red box. Genes that are
833 homologous are colored with the same color. Annotation is provided for the AG-390-L03
834 genome.

835 **Figure 3. A.** Maximum likelihood phylogenetic tree of FArhodopsin sequences detected
836 from marine and freshwater datasets colored in red and blue, respectively. The full
837 description of samples can be found in Table S4 and within Methods section.
838 Sequences with 100% amino acid identity with those in the tree are not shown.
839 Sequences in the blue background branch to the right (all from freshwater
840 metagenomes) do not have lysine in the retinal binding position in helix G (Table S4).
841 Numbers within freshwater branches represents clades on which a contig was used as
842 reference for the genomic comparison shown in B. **B**. Genomic comparison of
843 freshwater FArhodopsin subclades, using the Pelagibacterales SAG AG-390-L03

844 (subclade Ia.3) as a marine reference. Numbers between contigs represents the
845 average amino acid identity (AAI, %) of shared proteins between them**.**

846 **Figure 4. A.** Relative abundance (measured in reads per kilobase of genome per
847 gigabase of metagenome -- RPKG) of the flotillin-associated rhodopsin gene clusters in
848 a depth profile metagenomic study from Western Mediterranean Sea collected in a
849 single offshore location during a period of marked thermal stratification (Haro-Moreno *et*
850 *al.* 2018). Metagenomes are grouped by depth (i) upper photic (UP, 15 and 30 m), (ii)
851 deep chlorophyll maximum (DCM, 45 and 60 m), (iii) lower photic (LP, 75 and 90 m),
852 and deep (Deep, 1,000 and 2,000 m) layers. **B.** Metagenome and Metatranscriptome
853 analysis of flotillin-associated rhodopsin gene cluster. The upper panel show read
854 coverage of the metatranscriptomic Illumina reads from the 60 m sample of the profile
855 (98% identity, 250 bp window). In the lower panel, a metagenomic recruitment plot
856 using raw reads from the samples 15m and 60m (>70% identity, >50 bp long) is
857 represented. FAR gene cluster is highlighted in red. **C.** Read coverage showing
858 individual reads in the 60 m metatranscriptome of the FAR gene cluster showing that
859 the FArhodopsin and flotillin genes are co-transcribed (there are reads covering both
860 genes simultaneously).

861 **Figure 5.** Alpha fold prediction of retinal binding pockets in marine FArhodopsins from
862 different subclades: Pelagibacterales **(A)**, HIMB59 **(B)**, and *Pseudothioglobus* **(C)**.
863 Retinal is depicted in aqua green, and only residues that are present in more than 99%
864 of the subclade family members are displayed. Residues that clash with the retinal are
865 indicated and highlighted in red.

866

## SUPPLEMENTARY MATERIAL

**Figure S1.** Complete protein sequence alignment of 6 proteorhodopsin and 6 FArhodopsin sequences. Sequences come from 6 SAGs (two of each taxa) containing both rhodopsins in their genomes. Dark and light grey regions represent conserved residues within the alignment.

**Figure S2.** Maximum-likelihood phylogenetic tree of flotillin genes retrieved from marine SAGs. Reference flotillin sequences and their accession numbers within brackets are colored in black. Sequences in red represent sequences from marine SAGs that were not found near a FArhodopsin. Sequences in blue correspond to the flotillins of the three taxa coding for the tandem flotillin-FArhodopsin. Sequences in this branch were condensed for simplicity.

**Figure S3.** Relative abundance (measured in RPKG) of the FAR gene cluster in Tara Ocean metagenomes. Each province has been divided by depth in surface (SRF), deep chlorophyll maximum (DCM) and mesopelagic (MES) regions. The bars have been colored according to the microbes (blue*, Pelagibacterales;* purple, HIMB59 and green, *Pseudothioglobus*).

**Table S1**. Summary of marine SAGs, with their taxonomic classification and the number of type-I and type-III rhodopsin genes detected.

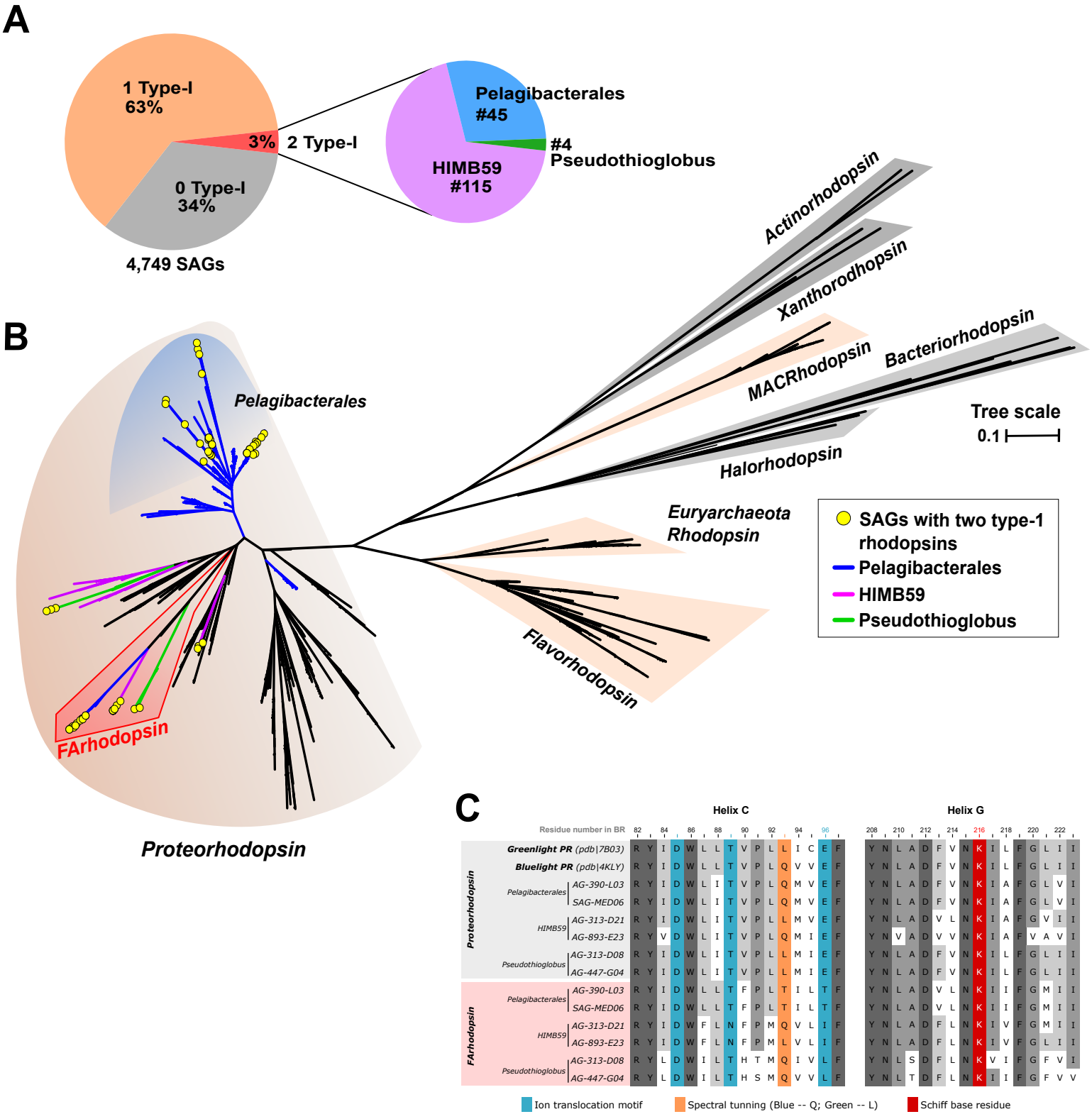
**Table S2.** Functional annotation of genes encoded within the FAR gene clusters of Pelagibacterales, HIMB59 and *Pseudothioglobus*. Only one FAR gene cluster per taxa is shown.

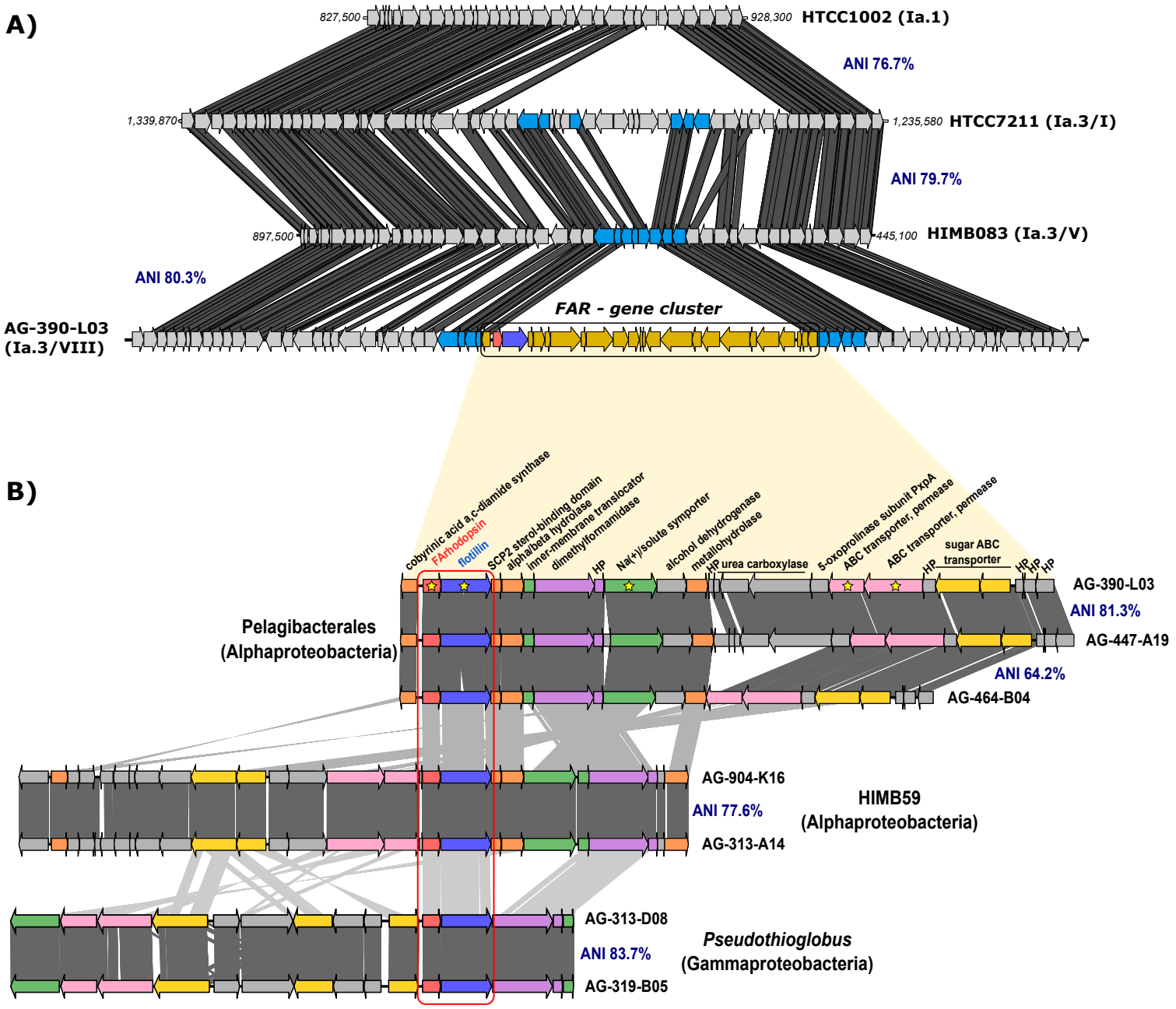
**Table S3.** Number of rhodopsin sequences per metagenomic dataset**.**

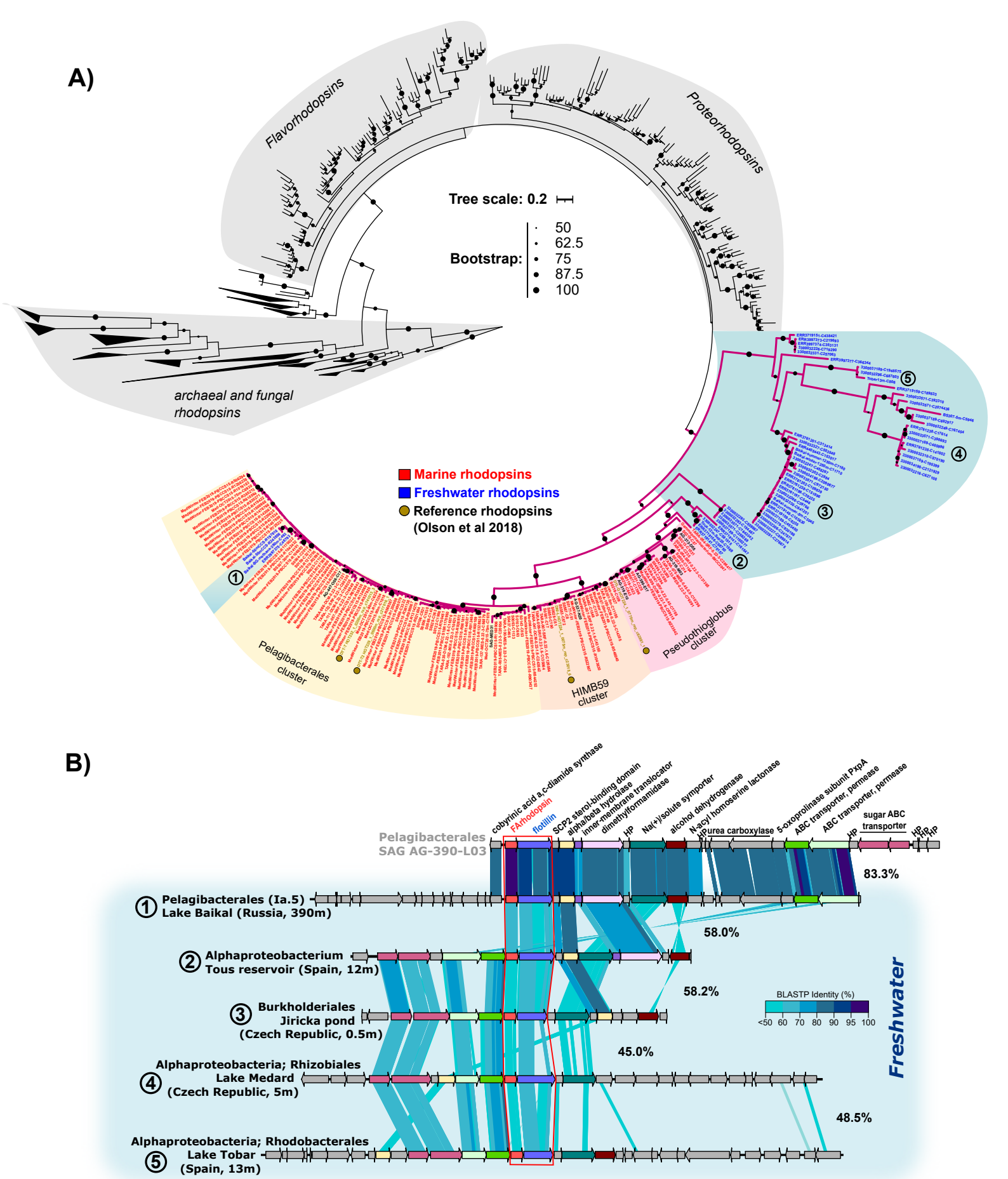
889 **Table S4.** Protein alignment of rhodopsin metagenomic sequences. Amino acids
890 involved in the ion translocation activity are colored in blue. The lysine involved in retinal
891 binding is colored in red.

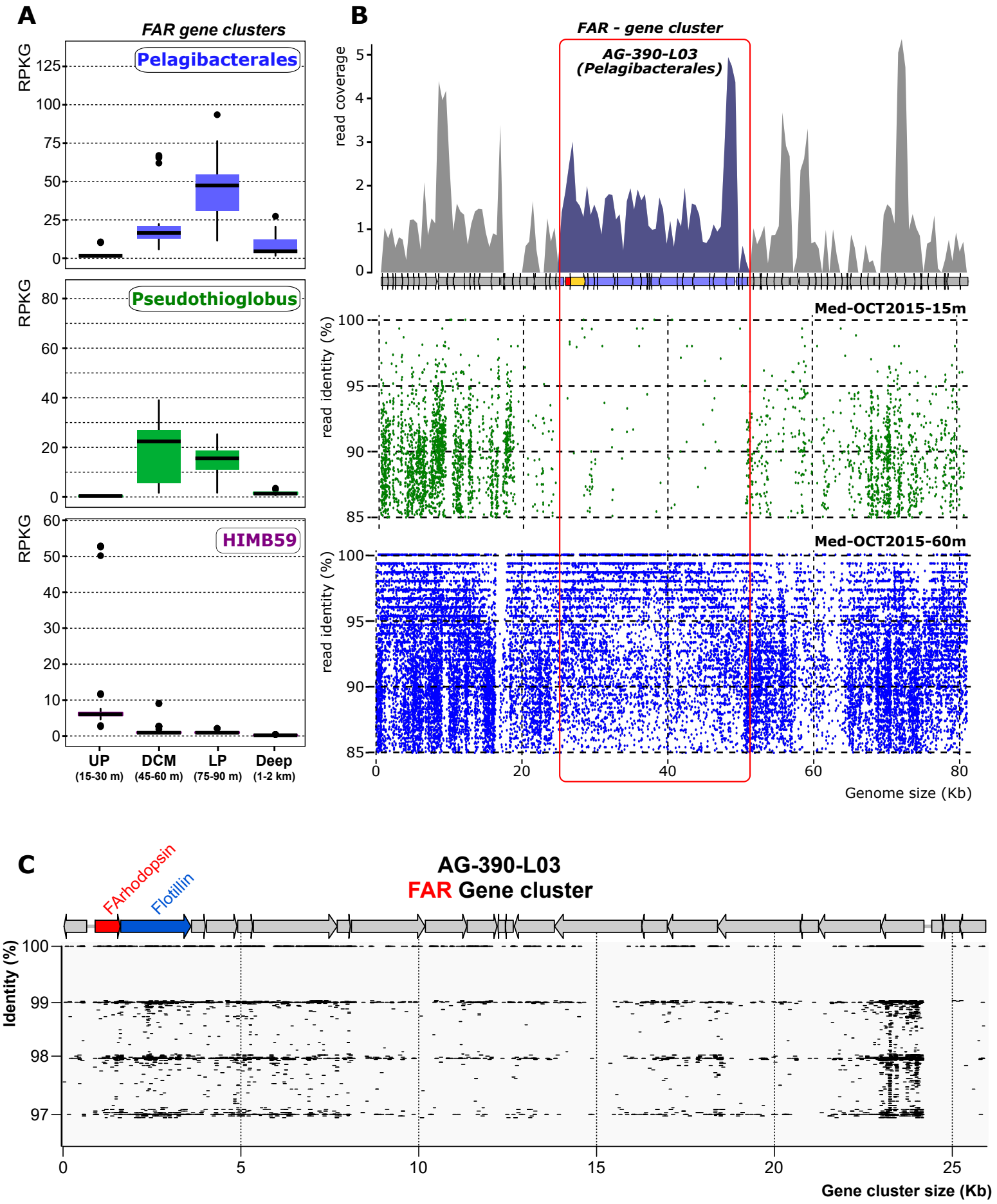
892

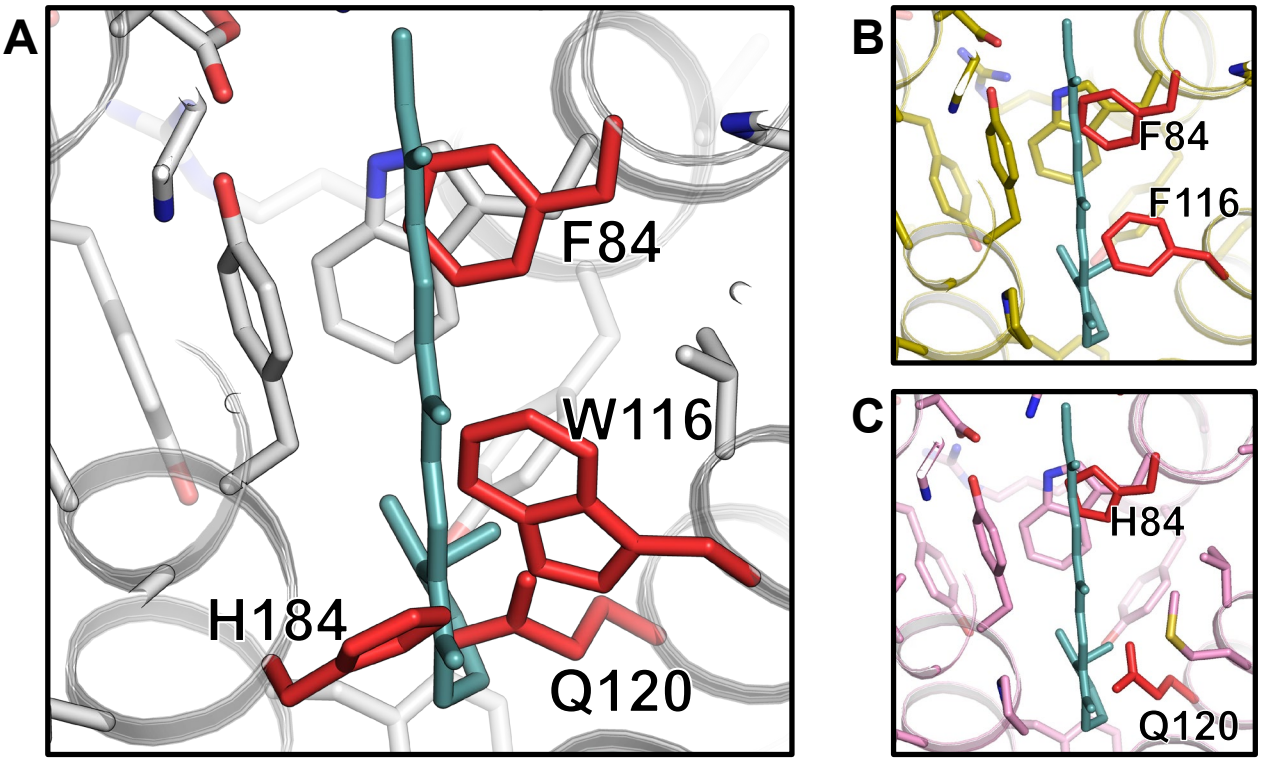
893


A
1 Type-I
63%
3%
2 Type-I
0 Type-I
34%
4,749 SAGs
Pelagibacterales
#45
#4
Pseudothioglobus
HIMB59
#115
B
Actinorhodopsin
Xanthorodhopsin
MACRhodopsin
Bacteriorhodopsin
Halorhodopsin
Tree scale
0.1
Euryarchaeota
Rhodopsin
Flavorhodopsin
Pelagibacterales
FArhodopsin
Proteorhodopsin
SAGs with two type-1 rhodopsins
Pelagibacterales
HIMB59
Pseudothioglobus
C
Helix C
Helix G
Residue number in BR
Proteorhodopsin
FArhodopsin
Greenlight PR (pdb|7B03)
Bluelight PR (pdb|4KLY)
Pelagibacterales
HIMB59
Pseudothioglobus
AG-390-L03
SAG-MED06
AG-313-D21
AG-893-E23
AG-313-D08
AG-447-G04
Ion translocation motif
Spectral tunning (Blue -- Q; Green -- L)
Schiff base residue


A)
827,500
928,300
HTCC1002 (Ia.1)
ANI 76.7%
1,339,870
1,235,580
HTCC7211 (Ia.3/I)
ANI 79.7%
897,500
445,100
HIMB083 (Ia.3/V)
ANI 80.3%
FAR - gene cluster
AG-390-L03
(Ia.3/VIII)
B)
cobyrinic acid a,c-diamide synthase
FArhodopsin
flotillin
SCP2 sterol-binding domain
alpha/beta hydrolase
inner-membrane translocator
dimethylformamidase
HP
Na(+)/solute symporter
alcohol dehydrogenase
metallohydrolase
HP
urea carboxylase
5-oxoprolinase subunit PxpA
ABC transporter, permease
ABC transporter, permease
HP
sugar ABC transporter
HP
HP
HP
AG-390-L03
ANI 81.3%
Pelagibacterales
(Alphaproteobacteria)
AG-447-A19
ANI 64.2%
AG-464-B04
AG-904-K16
ANI 77.6%
HIMB59
(Alphaproteobacteria)
AG-313-A14
AG-313-D08
ANI 83.7%
Pseudothioglobus
(Gammaproteobacteria)
AG-319-B05


A)
Flavorhodopsins
Proteorhodopsins
Tree scale: 0.2
Bootstrap:
50
62.5
75
87.5
100
archaeal and fungal rhodopsins
Marine rhodopsins
Freshwater rhodopsins
Reference rhodopsins (Olson et al 2018)
Pelagibacterales cluster
HIMB59 cluster
Pseudothioglobus cluster
B)
cobyrinic acid a,c-diamide synthase
FArhodopsin
flotillin
SCP2 sterol-binding domain
alpha/beta hydrolase
inner-membrane translocator
dimethylformamidase
HP
Na(+)/solute symporter
alcohol dehydrogenase
N-acyl homoserine lactonase
HP
urea carboxylase
5-oxoprolinase subunit PxpA
ABC transporter, permease
ABC transporter, permease
HP
sugar ABC transporter
HP
HP
HP
Pelagibacterales SAG AG-390-L03
83.3%
① Pelagibacterales (Ia.5) Lake Baikal (Russia, 390m)
58.0%
② Alphaproteobacterium Tous reservoir (Spain, 12m)
58.2%
③ Burkholderiales Jiricka pond (Czech Republic, 0.5m)
45.0%
④ Alphaproteobacteria; Rhizobiales Lake Medard (Czech Republic, 5m)
48.5%
⑤ Alphaproteobacteria; Rhodobacterales Lake Tobar (Spain, 13m)
BLASTP Identity (%)
<50 60 70 80 90 95 100
Freshwater


A
FAR gene clusters
Pelagibacterales
Pseudothioglobus
HIMB59
RPKG
UP (15-30 m)
DCM (45-60 m)
LP (75-90 m)
Deep (1-2 km)
B
FAR - gene cluster
AG-390-L03 (Pelagibacterales)
read coverage
Med-OCT2015-15m
Med-OCT2015-60m
read identity (%)
Genome size (Kb)
C
AG-390-L03
FAR Gene cluster
FArhodopsin
Flotillin
Identity (%)
Gene cluster size (Kb)


A
F84
W116
H184
Q120
B
F84
F116
C
H84
Q120