博士論文

A large-scale comparative genomic analysis to reveal adaptation strategies of marine Flavobacteriia

（大規模比較ゲノムから探る、海洋性フラボバクテリアの適応戦略）

Yohei Kumagai
熊谷洋平

2018

Contents

Chapter 1. General Introduction

Microbial ecology in the “$1000 genome” era

Due to the development of the sequencing technology in the 21st century, the sequencing cost of a whole human genome, which was about 100 million dollars in 2001, fell to be about $1,000 in 2014 (Sheridan 2014). The innovation provides two new approaches for microbial ecology: genomics and metagenomics. In 2017, the whole genome of more than 100,000 prokaryotes has been sequenced (https://www.ncbi.nlm.nih.gov/genome/) and microbial genomics provides us with great insights into ecology, physiology, and evolution of microbes (Land et al. 2015). Metagenomics enables us to understand microbial process occurring in environments and to discover novel microbial species and novel genes possessed by environmental microbes (DeLong 2005, Hiraoka et al. 2016, Tseng and Tang 2014). For example, the metagenomic analysis of groundwater that passed through a ~0.2-μm filter reveals more than 35 candidate phylum of bacteria which should comprise >15% of the bacterial domain (Brown et al. 2015, Luef et al. 2015). Metagenomics and genomics may give answers to the traditional questions in microbial ecology such as “What kind of bacteria are present in what kind of environments”, or “What kind of genes the bacterial strain in question possesses?”. Thus, in the post $1,000 genome era, we must consider how to obtain insights from large-scale sequence information (Kao et al. 2014). One of the most effective uses of genome information is the phylogenetic analysis of bacteria. Since pioneering work by Woese and Fox. (Woese and Fox 1977), bacterial phylogenetic trees were usually reconstructed using 16S rRNA sequences. However, using a combined sequence of all housekeeping genes in bacterial genomes provides a much more reliable phylogenetic tree (Segata et al. 2013). The genomic information not only improved the traditional methods but also created new methods. Comparative genomics is one of such new approaches (Galperin and Koonin 2014). The most basic concept of the comparative genome is a comparison of gene composition between different microbial groups (Lauro et al. 2009). For comparison of gene composition, clustering of genes expected to have the same function is necessary. Clustering of genes is roughly divided into a method using gene homology within a data set to be compared (O'brien et al. 2005, Uchiyama 2006) and a method using an external gene database (Huerta-Cepas et al. 2017, Kanehisa and Goto 2000). Presence/absence or quantitative difference of certain gene clusters should reflect differences in ecophysiology of bacterial strains. Moreover, comparative genomics is also useful approaches to predict functions of each gene. By comparative genomics, we can obtain information on genes in the vicinity of an unknown gene and information on genes showing a phylogenetic distribution pattern similar to that gene (Galperin and Koonin 2014, Pellegrini et al. 1999, Yamada et al. 2012). Such information can be obtained for all gene clusters, so this approach is especially useful to predict functions of genes with no homology to functionally established genes.

Discovery and distribution of bacterial rhodopsin in the environments

One of the biggest findings of metagenomics in marine environments is the discovery of proteorhodopsin (PR), one type of bacterial rhodopsins (also called as type

I rhodopsin) (Béja et al. 2000, Hugenholtz and Tyson 2008). Bacterial rhodopsin was first discovered in 1971 from halophilic archaea isolated from salt lakes and was named as bacteriorhodopsin (BR) (Oesterhelt and Stoeckenius 1971). BR plays a role as a light-driven proton pump that uses light energy to generate proton motive force (PMF). This PMF provides energy to generate ATP, thus BR-possessing bacteria can produce ATP via sunlight energy (Drachev et al. 1974). Until the discovery of PR, such light utilization mechanism had been regarded to be restricted to the hypersaline environment. However, the discovery of proteorhodopsin (PR) in marine environments changed the view on the distribution of the microbial rhodopsin (Béja et al. 2000). It is now evident that rhodopsin genes exist widely in bacterial communities of hydrosphere including both seawater (Campbell et al. 2008) and freshwater (Sharma et al. 2009). The ratio of microbes with rhodopsins is higher than that of microbes with photosynthetic systems (Finkel et al. 2013). It is estimated that up to 80% of prokaryotes may have PR in open ocean (Dubinsky et al. 2017), indicating that PR should be the most abundant rhodopsin in the aquatic environments. Rhodopsin genes not only show a wide geographical distribution but also occur in many taxonomic microbial groups. Among them, SAR11 and SAR86 groups are important because those are the most abundant in the ocean surface (Béja et al. 2000, Giovannoni et al. 2005a). Strains belonging to Flavobacteriia have been used for many culture-based experiments (Gómez-Consarnau et al. 2007, González et al. 2008, Palovaara et al. 2014, Riedel et al. 2013), and those belonging to Vibrio are used for genetic manipulation-based studies (DeLong and Beja 2010, Gómez-Consarnau et al. 2010).

Physiology of PR-possessing bacteria

As is shown above, many findings of the environmental distribution of PR have been made since 2000. However, culture-based physiological analysis of PR-possessing (PR+) bacteria is indispensable for understanding the significance of PR in the natural environment, no matter how widely distributed in the environment. Previous studies have shown that the PMF produced by PR may provide sufficient energy for ATP synthesis (Johnson et al. 2010, Yoshizawa et al. 2012) and some bacteria can utilize the PMF produced by PR for organic matter acquisition (Morris et al. 2010). Knockout-based experiment using PR+ vibrio strains showed PR promotes survival during starvation (Gómez-Consarnau et al. 2010). A particularly important finding is the ability to inorganic carbon fixation of PR+ Flavobacteriia. PR+ Flavobacteriia fix more inorganic carbon under light conditions than dark, indicating the presence of metabolic relation between energy production by PR and an anaplerotic carbon fixation reaction (González et al. 2008, Palovaara et al. 2014). Under the light, anaplerotic CO2 fixation contributing up to 30% of the total cellular carbon of Flavobacteriia (Palovaara et al. 2014). These findings suggest that PR may have a contribution to the global carbon cycle, if such a metabolic relation is widespread among PR possessing microorganisms.

PR-lacking and PR- possessing strains of marine Flavobacteriia

Both environmental distributions and physiological studies of PR suggest its important role for adaptation to sunlit environments. However, the growing understanding of PR function provokes another fundamental question—if the possession of PR is so advantageous as bonus “solar panels” for microbes, why are there so many PR-lacking (PR–) prokaryotes in the marine photic zone (Yoshizawa et al. 2012)? This question is important to reveal the advantage and disadvantage of light utilization. Comparative genomics is a potentially useful approach to answering such questions because genomes fundamentally reflect microbial ecophysiology (Cordero et al. 2012, Fernández-Gomez et al. 2013, Lauro et al. 2009, Thrash et al. 2014). That is, systematic differences between PR－ and PR+ prokaryote genomes might provide clues for understanding differences in the lifestyles of these microbes. Genomic differences revealed in a previous study showed that PR－ Flavobacteriia have significantly larger genomes than PR+ Flavobacteriia, although the ecophysiological reasons for this phenomenon remain enigmatic (Fernández-Gomez et al. 2013).

Purpose of this thesis

The purpose of this thesis is to clarify the advantage and disadvantage to possess PR by applying comparative genome analysis. I focus on marine Flavobacteriia because of the following reasons. First, there are a considerable number of isolates and genome data are available. Second, it is already evident that PR possessing species are spreading in the phylogenetic tree of this group. Third, some physiological and genetic data are accumulating for some species.

In this thesis, I sequenced 21 marine Flavobacteriia genomes and constructed genomic dataset including 41 PR− and 35 PR+ marine Flavobacteriia (Chapter 2) and using this genomic dataset, I did phylogenetic profiling analysis and statistically detected the genes with biased distribution in PR– strains or PR+ strains (Chapter 3). To validate my result of comparative genomics, I conducted several experimental assays (Chapter 4). With all the results in this thesis, I will discuss on the ecology, physiology, and evolution of ocean surface bacteria (Chapter 5).

Chapter 2. Construction of genome dataset of marine Flavobacteriia

Introduction

After the genome sequence of *Polaribacter* sp. MED152 in 2008, which is the first report of the whole genome of PR+ bacteria (González et al. 2008), many genomes of the PR+ bacteria have been sequenced. The complete genome of SAR11 was also sequenced in 2007 after the success of their cultivation (Giovannoni et al. 2005b). On the other hand, a complete genome of SAR86, another widespread bacterium possessing PR, has not yet been sequenced because it is uncultured yet and there are only multiple low-quality draft genomes have been obtained by single-cell genomics (Dupont et al. 2012). These facts indicate that the culturability is still the bottleneck for bacterial genome sequencing. Marine Flavobacteriia are relatively easy to be cultivated among PR+ bacteria, thus among PR+ microbes, the genomes of marine Flavobacteriia are so far the most sequenced to date (Fernández-Gomez et al. 2013, González et al. 2011, Kumagai et al. 2014, Kwon et al. 2013, Riedel et al. 2012, Riedel et al. 2013, Yoshizawa et al. 2014).

With the growing numbers of the genome sequences of marine Flavobacteriia, comparative genomics is a reasonable approach to understand the ecophysiological difference between PR– and PR+ strains. The previous study comparing 4 strains (2 PR– and 2 PR+ strains) of marine Flavobacteriia reported several characteristics of PR+ strains, such as smaller genome sizes and higher numbers of genes involved in anaplerotic CO_2 fixation (Fernández-Gomez et al. 2013). However, their analysis using small number of genomes makes it difficult to obtain statistically reliable correlation. Preparing enough number of genomes for both PR– and PR+ strains solve this problem and enables us to conduct more detailed comparative works. Based on these backgrounds, I sequenced 21 genomes of marine Flavobacteriia and constructed a genomic dataset of marine Flavobacteriia with available genomic data.

Material and Methods

Sample preparation and genome sequencing

Table 2-1 shows the summary of 21 marine Flavobacteriia strains whose genomes were sequenced in this study. Seven *Polaribacter* (*P. butkevichii* KCTC 12100^T, *P. gangjinensis* KCTC 22729^T, *P. glomeratus* ATCC 43844^T, *P. sejongensis* KCTC 23670^T, *P. reichenbachii* KCTC 23969^T, *P. porphyrae* NBRC 108759^T, and *P. filamentus* ATCC 700397^T) and six *Nonlabens* (*N. agnitus* JCM 17109^T, *N. arenilitoris* KCTC 32109^T, *N. sediminis* NBRC 100970^T, *N. spongiae* JCM 13191^T, *N. tegetincola* JCM 12886^T, and *N. xylanidelens* DSM 16809^T) type strains were provided by the Korean Collection for Type Cultures (KCTC), NITE Biological Resource Center (NBRC), Japan Collection of Microorganisms (JCM), and Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ). The other eight strains were isolated from environmental samples in 2009 (Yoshizawa et al. 2012): Four strains from surface seawater at Western North Pacific Station S (30°40'N, 138°00'E) during KT-09-11 cruise of R/V 'Tansei Maru' (Atmosphere and Ocean Research Institute, The University of Tokyo and Japan Agency for Marine-Earth Science and Technology (JAMSTEC)) (*Aureicoccus marinus* SG-18^T, *Tenacibaculum* sp. SG-28, *Tenacibaculum* sp. SZ-18, and *Gilvibacter* sp. SZ-19), two strain from surface seawater at Western North Pacific Station S1 (30°11'N, 145°05'E) during MR10-01 cruise of R/V 'Mirai' (JAMSTEC) (*Nonlabens marinus* S1-08 ^{T}, *Aureitalea marina* NBRC 107741^{T}), two strains from sea ice in Saroma-ko Lagoon (44°07'N, 143°58'E) (*Polaribacter* spp. SA4-10 and SA4-12), and one strain from surface seawater at Sagami Bay Station P (35°00'N, 139°20'E) during KT-09-11 cruise (*Winogradskyella* sp. PC-19). All strains were cultivated using half strength ZoBell's 2216E medium.

Genomic DNA samples were extracted by the standard phenol-chloroform method (Neumann et al. 1992). Genomes of two strains were sequenced using a 454 FLX+ System (Roche) and an Ion PGM System (Thermo Fisher Scientific) and assembled using Newbler assembler v2.7 software (Roche). Genomes of eleven strains were sequenced using a 454 FLX+ System and a MiSeq (Illumina) platform and assembled using Newbler assembler v2.7 software. Genomes of the other eight strains were sequenced using a PacBio RS II (Pacific Biosciences) instrument and assembled using Sprai v0.9.5.1.3 (http://zombie.cb.k.u-tokyo.ac.jp/sprai/) and subsequent manual curation. All sequencings were performed by following manufacturers' protocols, and all assembling steps were performed using default parameters. The isolation of Flavobacteriia was done by Dr. Susumu Yoshizawa and the cultivation and DNA extraction was done by Mr. Yu Nakajima. Sequencing and assembling genomes using a 454 FLX+ System and a MiSeq (Illumina) platform were done by Prof. Hayashi Tetsuya and Dr. Yoshitoshi Ogura.

Dataset preparation and assessment of genome completeness

I downloaded 55 genomes of marine Flavobacteriia from the NCBI RefSeq database (Pruitt et al. 2005) (Supplementary Table 2-S1). During the quality check of the sequenced genomes, I found that several scaffolds of *P. sejongensis* KCTC 23670^T and *P. reichenbachii* KCTC 23969^T genomes were likely to be contaminated. I randomly selected ten CDSs from the six scaffolds that coded CDSs and identified their origins by sequence similarity searches against the UniProt database (The UniProt Consortium 2014) (downloaded in April 2017, results in Table 2-2). The origin of each scaffold was consistently estimated at the genus level, and only the largest scaffold from each genome was concluded to be from the *Polaribacter* strains. I assessed the completeness of these two scaffolds after removing other scaffolds using Benchmarking Universal Single-Copy Orthologs (BUSCO, version 3.0.0) (Simão et al. 2015) and a 443 orthologue dataset that is conserved in the class Bacteroidetes, and obtained high scores (97.7% for *P. sejongensis* KCTC 23670^T and 98.4% for *P. reichenbachii* KCTC 23969^T). The genome sizes of these two genomes were also reasonable. The completeness of all 76 genomes was also assessed using BUSCO on the Bacteroidetes orthologue dataset.

Results

Marine Flavobacteriia genome sequencing and dataset preparation

To obtain a large, unbiased, and polyphyletic (phylogenetically dispersed) genomic dataset, genomes of 21 marine Flavobacteriia strains were sequenced. These strains contained eight *Polaribacter* type strains, seven *Nonlabens* type strains, and eight strains that were isolated from Saroma-ko Lagoon (Hokkaido, Japan), Sagami Bay (Kanagawa, Japan), and the western North Pacific Ocean (Supplementary Table S2-1). I subsequently downloaded 55 genomes of marine Flavobacteriia from the NCBI RefSeq database (Pruitt et al. 2005) and constructed a genomic dataset of 76 marine Flavobacteriia strains, 41 and 35 of which were PR－ and PR+ strains, respectively (Supplementary Table S2-1; their sampling sites are visualized in Fig. 2-1). All genomes were subjected to in-house annotation of their ribosomal RNAs, transfer RNAs, and protein-coding sequences (CDSs). To evaluate the quality of the 76 genomes, their completeness was estimated using BUSCO software (Simão et al. 2015). The scores averaged 98.0%, and the completeness of five genomes was less than 95.0%. The lowest BUSCO score was that of *Salinibacter ruber* DSM 13855 ^{T}, which had acquired many genes from hyperhalophilic archaea (Mongodin et al. 2005). Excluding the five genomes with less than 95.0% completion did not affect the conclusions of this study.

Discussion

In this chapter, I sequenced 21 genomes of marine Flavobacteriia and acquired complete genomes of 4 strains (*Polaribacter* sp. SA4-10, *Polaribacter* sp. SA4-12, *Gilvibacter* sp. SZ-19 and *Winogradskyella* sp. PC–19) (Table 2-1). Regarding the 17 draft genomes, I acquired 14 high-quality draft genomes with 5 or fewer contigs and 3 draft genomes with 30 or fewer contigs (Table 2-1). Average BUSCO score of newly sequenced genomes is 96.82% and high BUSCO scores ensure the high quality of obtained genomes. Two strains contained contaminated contigs, however, I successfully removed those contigs because contaminated contigs belong to phylogenetically distant species from Flavobacteriia (Table 2-2). With available genome data from NCBI database and newly sequenced genome, I acquired 76 marine Flavobacteriia genome including 41 PR– and 35 PR+ strains. The sampling points of strains included in the genome dataset cover a wide range of environments such as Antarctica, Mediterranean Sea, Pacific Ocean, Japan Sea, but there was no difference in the geographical distribution of PR– bacteria and PR+ strains (Fig. 2-1). Therefore, it is expected that biases due to differences in the isolation source of each strain will not affect downstream analysis. It should also be noted that many of the strains contained in this genomic dataset are isolated from coastal areas and only 6 strains were from open ocean (Fig. 2-1 and Supplementary Table S2-1). In addition, this data set includes a more fundamental bias of using only cultivable strains. To avoid these sampling biases, long-read sequencing technology will be one solution because the advancing of such technology allows us to acquire complete genomes of uncultured bacteria from environmental samples.

Table 2-1. List of 21 marine Flavobacteriia genomes that were sequenced in this study.

Species	Strain	Type strain	Total scaffold size	Number of contigs	Total read number	Coverage	Sequence method	Assemble method	BioSample ID	PR
Tenacibaculum sp.	SG-28 (NBRC 107667)	-	2801347	17	648278	55.5	454FLX+ and Ion PGM	Newbler v2.7	SAMN06075357	+
Aureicoccus marinus	SG-18 (NBRC 108814)	T	3052917	2	664251	49.6	454FLX+ and Ion PGM	Newbler v2.7	SAMN06075358	+
Winogradskyella sp.	PC-19 (NBRC 107664)	-	2957311	2	1280430	115	454FLX+ and MiSeq	Newbler v2.7	SAMN06133349	+
Tenacibaculum sp.	SZ-18 (NBRC 107760)	-	4023590	9	972366	68	454FLX+ and MiSeq	Newbler v2.7	SAMN06133350	+
Aureitalea marina	NBRC 107741	T	3074655	4	18215	39.92	PacBio RS II	Sprai v0.9.5.1.3	SAMN06074325	+
Gilvibacter sp.	SZ-19 (NBRC 107666)	-	3097621	2	1028954	86	454FLX+ and MiSeq	Newbler v2.7	SAMN06133352	+
Polaribacter sp.	SA4-10 (NBRC 107119)	-	3433642	3	776100	65	454FLX+ and MiSeq	Newbler v2.7	SAMN06133347	+
Polaribacter sp.	SA4-12 (NBRC 108842)	-	3970876	5	1264970	82	454FLX+ and MiSeq	Newbler v2.7	SAMN06133351	-
Polaribacter butkevichii	KCTC 12100	T	4085573	4	1062036	72	454FLX+ and MiSeq	Newbler v2.7	SAMN06133457	-
Polaribacter gangjinensis	KCTC 22729	T	2943061	2	1391804	130	454FLX+ and MiSeq	Newbler v2.7	SAMN06133456	-
Polaribacter glomeratus	ATCC 43844	T	4064562	4	716134	55	454FLX+ and MiSeq	Newbler v2.7	SAMN06133459	+
Polaribacter sejongensis	KCTC 23670	T	4413491	1 (7)	1284906	42	454FLX+ and MiSeq	Newbler v2.7	SAMN06133458	-
Polaribacter reichenbachii	KCTC 23969	T	4081294	1 (7)	1000286	35	454FLX+ and MiSeq	Newbler v2.7	SAMN06133461	-
Polaribacter porphyrae	NBRC 108759	T	3904103	2	1192068	87	454FLX+ and MiSeq	Newbler v2.7	SAMN06133460	-
Polaribacter filamentus	ATCC 700397	T	4173678	3	15320	24.85	PacBio RS II	Sprai v0.9.5.1.3	SAMN06074323	+
Nonlabens agnitus	JCM 17109	T	3215882	4	18283	37.05	PacBio RS II	Sprai v0.9.5.1.3	SAMN06074326	+
Nonlabens arenilitoris	KCTC 32109	T	3323003	3	15506	29.8	PacBio RS II	Sprai v0.9.5.1.3	SAMN06075340	-
Nonlabens sediminis	NBRC 100970	T	2883442	5	13430	37.72	PacBio RS II	Sprai v0.9.5.1.3	SAMN06075339	+
Nonlabens spongiae	JCM 13191	T	3393235	2	16831	34.24	PacBio RS II	Sprai v0.9.5.1.3	SAMN06075354	-
Nonlabens tegetincola	JCM 12886	T	3028293	21	9489	26.17	PacBio RS II	Sprai v0.9.5.1.3	SAMN06075350	+
Nonlabens xylanidelens	DSM 16809	T	3552991	28	11834	22.64	PacBio RS II	Sprai v0.9.5.1.3	SAMN06075351	-

Table 2-2. Estimated origins of scaffolds of *Polaribacter sejongensis* KCTC 23670^{T} and *Polaribacter reichenbachii* KCTC 23969^{T} genomes.

Strain name	Scaffold ID	Scaffold size	Number of CDSs	Estimated genus to which each scaffold belongs	Contamination
Polaribacter sejongensis KCTC 23670	1	4526271	3664	*Polaribacter*	No
	2	3734458	3763	*Bacillus*	Yes
	3	7391	0	—	—
	4	4279	0	—	—
	5	3297	0	—	—
	6	3039	0	—	—
	7	2044	0	—	—
Polaribacter reichenbachii KCTC 23969	1	4122594	3491	*Polaribacter*	No
	2	2400331	2144	*Lacinutrix*	Yes
	3	2005328	1690	*Lacinutrix*	Yes
	4	43261	27	*Lacinutrix*	Yes
	5	5595	0	—	—
	6	5398	0	—	—
	7	2781	0	—	—

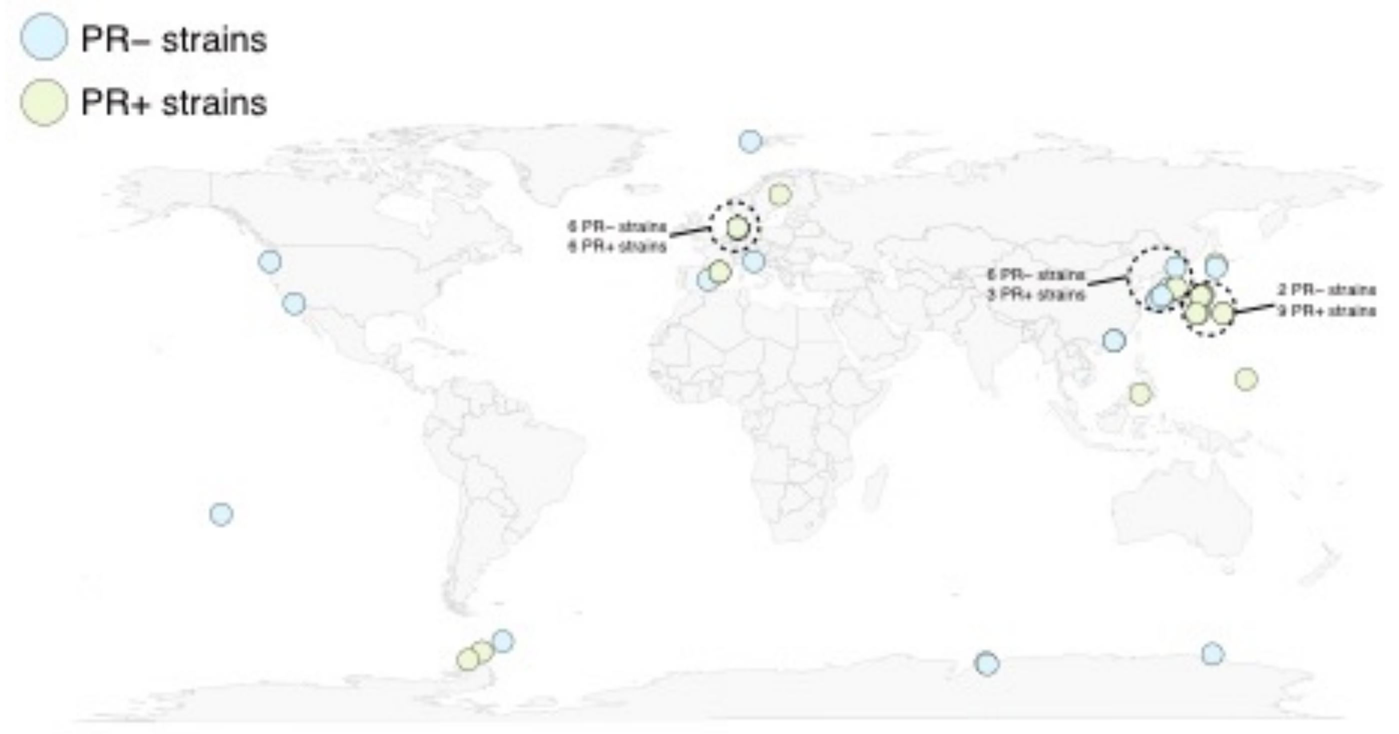


Figure 2-1. Sampling sites of 54 flavobacterial strains in genomic dataset.

Circles indicate the sampling sites of flavobacterial strains in my genomic dataset (Yellow: PR–, Purple: PR+). The information of latitude and longitude were obtained from IMG database (Markowitz et al. 2013) except for newly sequenced strains in this study. Among 76 Flavobacteriia, no information of latitude and longitude were available for 22 strains. This figure was visualized using “maps” package of R software (https://cran.r-project.org/web/packages/maps/maps.pdf).

Chapter 3. Phylogenetic profiling analysis of marine Flavobacteriia

Introduction

In the previous chapter, I constructed the genome dataset of marine Flavobacteriia. In this chapter, I perform comparative genomic analysis by phylogenetic profiling analysis based on the genomic dataset. Phylogenetic profiling is a method of comparative genome analysis to detect the relationship between genes by the similarity to a phylogenetic distribution pattern of each gene (Pellegrini et al. 1999, Sun et al. 2005). In many cases, this method is used to detect associated gene pairs such as genes coding protein complex (Pagel et al. 2004), genes working on the same pathway (Snitkin et al. 2006), and genes with regulatory relationships (Rodionov and Gelfand 2005). To apply the phylogenetic profiling method, polyphyletic distribution of genes is required because if you analyze a gene that exists only in a specific phylogenetic group, it simply detects a group of genes specifically distributed in a certain phylogenetic group (Fig. 3-1 (a, b)). It has been reported that the horizontal gene transfer (HGT) of PR gene occurred multiple times in the evolutionary history of marine Flavobacteriia, and as a result, PR is predicted to show polyphyletic distribution (Pinhassi et al. 2016, Yoshizawa et al. 2012). Therefore, PR should be a good target of phylogenetic profiling analysis (Sharma et al. 2006). In this analysis, I also analyze the inverse correlation of each gene (Fig. 3-1 (c)) to know the adaptive strategies of PR– strains.

Methods

Functional annotation of genes

All 76 genomes were annotated by the following procedure. Ribosomal and transfer RNA genes were annotated using RNAmmer v1.2 (Lagesen et al. 2007) and tRNAscan-SE v1.3.1 (Lowe and Eddy 1997), respectively, with their default settings. Subsequently, I masked the ribosomal and transfer RNA gene sequences with "N" and predicted CDSs with Prodigal v2.50 (Hyatt et al. 2010) with default settings.

The functional annotation of the 258,135 CDSs was performed by eggNOG-mapper (Huerta-Cepas et al. 2017) and the bactNOG dataset in the eggNOG database version 4.5 (Huerta-Cepas et al. 2015), using the DIAMOND algorism for mapping and setting the taxonomic scope to Bacteroidetes. This approach resulted in functional annotation of 184,623 (71.5%) of the CDSs to 14,361 eggNOG orthologue groups, excluding function-unknown orthologue groups (i.e., groups whose annotations contained any of the terms "NA", "unknown", or "DUF").

Amino acid sequences of CDSs that were annotated to the rhodopsin orthologue group (ENOG05CSB) were aligned using MAFFT version 7.212 (Katoh et al. 2002) with the linsi algorithm and default parameters. The alignments were curated using trimAl version 1.2 (Capella-Gutiérrez et al. 2009) with the option "–gt 1". The best substitution model of each alignment was selected by using prottest3 (Darriba et al. 2011). The maximum-likelihood method was performed using RAxML version 7.2.8 (Stamatakis 2006) and 1,000 bootstrap replicates. The other settings were set at their default values.

Phylogenetic classification of rhodopsins as PR, NaR, and ClR genes was conducted as described in previous study (Yoshizawa et al. 2014).

Reconstruction of the genomic phylogenetic tree

As outgroups, genomes of two strains of the class Bacteroidetes (*Cytophaga hutchinsonii* ATCC 33406^T and *Salinibacter ruber* DSM 13855^T) were additionally downloaded from the NCBI RefSeq database. The prediction and annotation of their CDSs were conducted in the same manner as the other genomes. I selected 155 ENOG orthologue groups so that each genome contained exactly one CDS that was annotated to each of those orthologue groups. Their amino acid sequences were aligned using MAFFT and curated using trimAl as described above. The best substitution model of each alignment was selected by using prottest3. The alignments of 155 eggNOG orthologue groups were concatenated and subjected to phylogenetic tree reconstruction using RAxML with the best substitution model for each protein column and 1,000 bootstrap replicates. The other settings were set to their default values.

Genome size and gene content analysis

The difference in the genome sizes of the PR－ and PR+ genomes was statistically evaluated by applying Student's *t*-test to the total scaffold sizes of the two groups using R software. I used all complete and draft genome data to compare their genome sizes. Orthologue group distributions that were biased to PR－ or PR+ genomes were identified by applying the Brunner-Munzel test (Brunner and Munzel

2000) to the numbers of CDSs in each of the 14,361 eggNOG orthologue groups. To correct for multiple testing, Storey's approach (Storey and Tibshirani 2003) was used with a cut-off false discovery rate of 0.05.

Gene proximity analysis

A gene proximity network analysis was conducted to group the detected genes into functionally related clusters. The gene proximity network was constructed by connecting any orthologue group pair that are located within 20 kb of each other in at least ten genomes in my dataset using in-house python script. Network was visualized using Cytoscape software(Shannon et al. 2003).

Analysis of RNR gene classes

To identify the classes of RNR genes, all CDSs that were annotated with the ENOG05BZH (ribonucleotide reductase) were fed into domain-level annotation using the NCBI conserved domain search (Marchler-Bauer et al. 2014). For phylogenetic analysis of RNR genes, RNR genes of *Lactobacillus leichmannii* (GenBank: AAA03078) and *Escherichia coli* H736 (GenBank: EGI11882) were downloaded from GenBank to serve as representatives of class II and class I genes, respectively. CDSs were aligned, and the alignments were curated by the same methods described above. The best-fit substitution model was selected by using prottest3 at its default settings. The maximum-likelihood method was performed using RAxML and 1,000 bootstrap replicates. The other settings were set to their default values.

Analysis of Tara Oceans dataset

The Tara Oceans dataset, containing gene abundance FPKM value, oxygen concentration, and sampling depth data, was downloaded from http://ocean-microbiome.embl.de/companion.html (Sunagawa et al. 2015). Correlation analysis was conducted using the "psych" package in R (https://pbil.univ-lyon1.fr/CRAN/web/packages/psych/). Curve fitting was also done by using locally weighted scatterplot smoothing with its default options.

Results

Functional annotation and confirmation of polyphyletic PR distribution

The CDSs were functionally annotated using eggNOG-mapper(Huerta-Cepas et al. 2017) and the bactNOG dataset in the eggNOG database (Huerta-Cepas et al. 2015). Among the 258,135 CDSs in total, 71.5% were assigned to any eggNOG orthologue group by ignoring function-unknown groups. I further classified the CDSs that were assigned to the rhodopsin orthologue group (ENOG05CSB) as PR, Na^+-pumping rhodopsin (NaR), and Cl^--pumping rhodopsin (ClR) genes by phylogenetic analysis (Fig. 3-2). All NaR-possessing strains had additional PR genes, whereas two ClR-possessing strains (*Nonlabens spongiae* JCM 13191^T and *Psychroserpens* sp. Hel_I_66) were revealed to lack PR genes. I treated these two ClR-possessing strains as PR+ strains in the following analyses because the inward Cl^--pumping activity of ClR also generates membrane potential; however, the conclusions were not affected even if they were treated as PR－ strains.

I then reconstructed a genomic phylogenetic tree of the 76 marine Flavobacteriia strains by applying the maximum-likelihood method to the concatenated protein sequence dataset of 155 conserved CDSs that were present in each strain in exactly one copy. To root the tree, genomes of two strains of the phylum Bacteroidetes were added to the dataset as outgroups. I confirmed that the PR－ and PR+ strains had polyphyletic distributions on the reconstructed phylogenetic tree, fulfilling the second condition for a comparative genomic study (Fig. 3-3).

Detection of genes significantly biased to either PR－ or PR+ genomes

I first compared the genome sizes of PR－ and PR+ marine Flavobacteriia strains. As consistent with previous findings (Fernández-Gomez et al. 2013), the PR－ genomes were significantly larger than the PR+ genomes (p-value = 4.7E-3, Fig. 3-3 and Fig. 3-4 (a)). To further investigate the ecophysiological background that might cause this difference in genome size, I compared their CDS numbers in each eggNOG functional category (Huerta-Cepas et al. 2015) (Fig. 3-4 (b)). I found that except for several categories that are generally rare in bacteria, the numbers of CDSs were consistently larger in the PR－ genomes than in the PR+ genomes, regardless of their functional categories. This result suggests that the observed genome-size difference is not due to acquisitions (in the PR－ strains) or losses (in the PR+ strains) of gene sets involved in specific metabolic and/or cellular systems but rather due to net acceleration of genome size expansion (in the PR－ strains) or reduction (in the PR+ strains).

Next, I investigated if there are specific eggNOG orthologue groups that had particularly biased distributions. A statistical test detected 86 and 43 (129 in total) orthologue groups whose distributions were significantly biased to the PR－ and PR+ genomes, respectively (q-value < 0.05, Supplementary Tables S3-1 and S3-2). Except for the trivial case of the PR gene itself, the most significant case was the enrichment of the beta-carotene dioxygenase (*blh*) gene (ENOG05FTR) in the PR+ genomes. This result is quite reasonable because the *blh* gene is involved in the synthesis of retinal, the chromophore of PR.

One unexpected finding was that most of the genes involved in anaplerotic inorganic carbon fixation were not included in the orthologue groups that showed biased distributions to the PR+ genomes (except for the *sbtA* gene (ENOG05EGC), Fig. 3-5). In a previous study, PR+ Flavobacteriia were argued to have significantly more genes involved in anaplerotic inorganic carbon fixation (Fernández-Gomez et al. 2013) for PR-coupled carbon fixation and light-promoted growth (González et al. 2008, Palovaara et al. 2014). I assume that the previously observed larger proportion of those genes in PR+ genomes might be due to the smaller genome size of PR+ strains (i.e., denominators in the calculation). Instead, based on the universal occurrence pattern of those genes, I assume that the fixation of inorganic carbonic acid by anaplerotic carbon fixation would be a common feature among marine Flavobacteriia.

Proximity analysis of genes biased to PR－ or PR+ genomes

I conducted gene proximity analysis of the 129 orthologue groups that showed biased distributions in PR– or PR+ strains because genes that are near each other in genomes likely have related functions (e. g. genes involved in same biological pathways) (Huynen et al. 2003). A gene proximity network was constructed by connecting any orthologue group pair that are located within 20 kb of each other in at least ten genomes in my dataset (Fig. 3-6). A typical example of such proximal relation was seen between the rhodopsin and *blh* genes, which are often adjacently coded for concerted expression (Pinhassi et al. 2016). I note that the two ClR-possessing PR– strains code the *blh* genes next to their ClR genes.

Three large clusters were formed in the gene proximity network. Among them, two clusters were composed of genes that were enriched in the PR－ genomes: The first was composed of genes for anaerobic nitrous-oxide metabolism, and the second was composed of genes for synthesis and transport of aryl polyenes (APEs) (Fig. 3-6). The third cluster was composed of photolyase and photolyase-related genes, which were enriched in the PR+ genomes. These three large clusters were assumed to especially reflect lifestyles to which PR－ and PR+ Flavobacteriia species have adapted.

Signs of adaptation of PR－ Flavobacteriia to anaerobic conditions

Despite a predominance of function-unknown genes in the 129 orthologue groups that showed biased distributions, I discovered one interesting trend therein: The genes enriched in the PR－ genomes showed several signs of adaptation to microaerobic or anaerobic conditions although Flavobacteriia species are usually considered to be strictly aerobic (Kirchman 2002).

The PR－ genomes coded significantly more nitrous oxide reductase (*nosZ,* ENOG05EQJ) and nitrous oxide metabolism (*nosY,* ENOG05J39) genes than those of the PR+ genomes (Fig. 3-3, q-value = 3.7E-2 and 3.7E-2, respectively). These genes, which were members of the first cluster that was formed in the gene-proximity network (Fig. 3-6), function in bacterial anaerobic N_2O respiration (Chan et al. 1997, Coyle et al. 1985), which uses nitrous oxide as a terminal electron acceptor at reduced oxygen concentrations (Bauer et al. 2006). Second, the PR－ genomes had more class II ribonucleotide reductase (RNR) genes (ENOG05BZH) (Fig. 3-3, q-value = 9.9E-5). RNR proteins catalyse the synthesis of deoxyribonucleotides from ribonucleotides and are grouped into three classes according to their subunit types (Nordlund and Reichard 2006). NCBI conserved domain searches (Marchler-Bauer et al. 2014) and a phylogenetic analysis (Fig. 3-7) showed significant enrichment of the class II RNR genes in the PR－ genomes (PR－: 23/41, PR+: 2/35) occurred. Class II RNRs do not depend on oxygen for their catalytic function, whereas class I RNRs function under aerobic conditions (Nordlund and Reichard 2006). A catalase gene, *katE* (ENOG05CH6), was also enriched in the PR－ genomes (Fig. 3-3, q-value = 4.9E-3). This gene was reported to modulate reactive oxygen stress when cells that usually live in anaerobic conditions are exposed to oxygen. Expression of *katE* increases under anaerobic conditions in *E. coli* (Schellhorn and Hassan 1988), and the *katE* protein is the only H_2O_2-removing enzyme that is present in an obligate anaerobic Bacteroidetes, *Bacteroides thetaiotaomicron* (Imlay 2013, Mishra and Imlay 2013). In addition, the PR－ genomes almost always had *cbb*$_3$-type cytochrome oxidase genes (ENOG05EUH), whereas the PR+ genomes did not (Supplementary Fig. S3-1, q-value = 1.6E-2). The *cbb*$_3$-type cytochrome oxidases have a very high affinity for O_2 so that their organisms can respire under microaerobic conditions (Preisig et al. 1993), and they should enable Flavobacteriia to survive in transiently low-O_2 microniches (González et al. 2011).

Enrichment of UV-screening pigment synthesis genes in PR－ genomes

The second cluster in the gene-proximity network contained 16 genes for the synthesis and transport of APEs and was enriched in the PR－ genomes (Fig. 3-6). Most notably, almost all genes in this cluster were not only significantly but also exclusively found in the PR－ genomes (Fig. 3-8 (a)). The genes in this cluster corresponded well to those previously reported in an APE-producing gene cluster in the *Flavobacterium johnsoniae* ATCC 17061^T genome (Cimermancic et al. 2014) (Fig. 3-8 (b)). These data strongly suggested that production of APEs is a unique feature of PR－ marine Flavobacteriia.

APEs (Fig. 3-8 (c)) protect bacterial cells from UV and visible light by localizing to outer membranes (Goel et al. 2002, Wang et al. 2013). This localization to outer membranes contrasts with the localization of carotenoids to inner membranes (Irschik and Reichenbach 1978) but resembles that of scytonemin, a cyanobacterial UV-screening extracellular phenolic pigment (Gao and Garcia-Pichel 2011). When proteins that synthesize the dialkylresorcinol (DAR) moiety are present (e.g., in *F. johnsoniae* cells), APEs are esterified with the DAR moiety and converted to flexirubin-type pigments (FTPs) (Fig. 3-8 (d)). FTPs are well-studied yellow-to-orange pigments specific to Bacteroidetes and have been used as a chemosystematic marker for taxonomic studies because of its polyphyletic distribution (Fautz and Reichenbach 1980, Reichenbach et al. 1980). FTPs also absorb UV and visible light(Reichenbach et al. 1980, Venil et al. 2014) and localize to outer membranes (Irschik and Reichenbach 1978), and can be detected by a flexirubin test (Fautz and Reichenbach 1980). A strain that has APE synthesis genes and the *darA* (ENOG08K4P) and *darB* (ENOG05CXX) genes, *Aquimarina muelleri* DSM 19832^T, was reported to respond positively to the flexirubin test (Nedashkovskaya et al. 2005).

Finally, the third cluster formed in the gene-proximity network contained photolyase and photolyase-related genes and was enriched in the PR+ genomes (Fig. 3-6). Photolyase is an enzyme that uses visible light energy to repair DNA damage caused by UV light (Sancar 1994). Specifically, the PR+ genomes coded significantly more genes of a photolyase paralogue (ENOG05CVP) than the PR－ genomes did (Fig. 3-3, PR－: 1.9, PR+: 2.9 copies per genome on average).

Discussion

In this chapter, I conducted a comparative genomic analysis of PR－ and PR+ marine Flavobacteriia. The large and unbiased genomic dataset enabled me to clarify their differences, which appeared to be related to fundamentally different life styles and ecophysiological strategies. In addition, the polyphyletic distribution of PR genes (Fig.

3-3) and genomic traces indicated that PR genes have not only been gained but also lost during evolution (Fig. 3-9), suggesting that the conditions that have made each of the PR– and PR+ lifestyles advantageous have not been stable during the course of evolution. The approach adopted in this study can be further applied to provide broader insights into microbial ecology in the future—the more genomes I have, the more powerful comparative genomic approaches become possible. Moreover, if enough number of genomes will be sequenced, this approach should be useful not only prokaryotes but also eukaryotes.

My results suggest that PR－ and PR+ marine Flavobacteriia adopt contrasting strategies to address UV damage: The former produces APEs or FTPs to avoid UV damage, whereas the latter produces photolyase to efficiently repair themselves after UV damage (Fig. 3-10). I propose that PR+ Flavobacteriia accept both UV damage and cost of repairing UV-damaged DNA so that they can take advantage of light energy by using PR in their inner membranes. On the other hand, PR－ Flavobacteriia avoid the UV damage by blocking the UV light and thus must abandon utilization of light energy. To confirm the generality of my finding across different taxonomic groups, I analysed the distribution patterns of rhodopsin and APE synthesis genes in all prokaryotes. While both rhodopsin and APE synthesis genes are distributed across diverse phyla, I observed their completely exclusive distribution patterns, i.e., no strain possesses both rhodopsin and APE genes (Table 3-1). In accord with the analogy in which PR functions as microbial "solar panels", I propose a "solar-panel or parasol" hypothesis, in which APEs and FTPs are regarded as cellular "parasols," although it should be noted that correlation does not necessarily mean causality and experimental verification would be required. In this framework, I can choose to either charge solar-powered devices or use parasols to avoid tanning but cannot do both simultaneously.

Notably, these two different strategies for the handling of UV damage may also explain the smaller genome size of PR+ Flavobacteriia. First, UV damage itself would accelerate the net rate of genome size reduction in the PR+ strains via induced double strand breaks and nonsense mutations (Brash and Haseltine 1982). Second, stronger selection pressure to minimize the DNA repair cost would also lead to the smaller genome size in PR+ Flavobacteriia. In contrast, PR－ Flavobacteriia would receive less DNA damage and bear less cost for maintaining DNA; thus, they may be able to maintain a larger genome.

The evidence for the adaptation of PR－ Flavobacteriia to conditions that are characterized by genes associated with anaerobic lifestyles provides another perspective on their ecophysiological adaptation (Fig. 3-10). Because molecular oxygen is required to synthesize retinal (Kim et al. 2009), PR+ bacteria are expected to prefer aerobic environments. To directly confirm this relationship between rhodopsins and oxygen, we re-analysed the shotgun metagenomic data of Tara Oceans samples (Sunagawa et al. 2015) and observed a positive correlation between rhodopsin gene abundance and oxygen concentration, even after normalizing for the effects of sampling depths (Pearson's partial correlation = 0.61, *n* = 133) (Fig. 3-11). Although Flavobacteriia are generally thought to be aerobic, it may be noted that a species in the family Flavobacteriaceae (*Muricauda ruestringensis* DSM 13258^{T}) has nitrous oxide reductase genes (*nosZ* and *nosY*) and was reported to be facultative anaerobic (Bruns et al. 2001). I also note that the presence of Flavobacteriia is significant in environments with nanomolar oxygen concentrations and that nitrous oxide reductase genes are more abundant in particle-associated microbial communities than in free-living communities (Ganesh et al. 2014). Thus, the interiors of macroscopic organic aggregates (also known as marine snows) in the upper ocean, which is known as a typical niche of Flavobacteriia (Buchan et al. 2014, Kirchman 2002), are an environment where facultative anaerobic PR－ microbes may predominate because their microaerobic (and nutrient-rich) conditions likely decrease the advantage of possessing PR (Ploug et al. 1997). A possible niche of the facultative anaerobic PR– Flavobacteriia with UV protective pigments might be the surface layer of the eastern tropical north Pacific ocean, whose oxygen concentration is <10 μM even in the near-surface layer (Paulmier and Ruiz-Pino 2009). We should keep in mind that the distribution of PR+ and PR– Flavobacteriia cannot simply be extended to bacteria in the ocean surface layer in general. For example, it is known that flavobacterial abundance is inversely correlated with the bacteria of the SAR11 group, which is most dominant PR+ bacteria in the ocean surface layer, and the potential niche of PR+ Flavobacteriia and SAR11 group bacteria is expected to be different (Williams et al. 2013).

Table 3-1. Numbers of genomes that code rhodopsin and APE synthesis genes. Genomes that code more than 50 % of APE synthesis genes (eggNOG ID: 05C84, 05EWP, 05F6T, 05IAS, 05VME, 0636E, 07T6G, 08M6I, 05DZS, 05EYS, 05H6H, 05M49, 05YNY, 07T3I, 08H1G, and 08XF1) were regarded as APE-synthesizing (APE+) strains. The data were obtained from the eggNOG database version 4.5 (Huerta-Cepas et al. 2015).

Taxonomic group	Rhodopsin+		Rhodopsin−	
	APE+	APE−	APE+	APE−
Bacteroidetes	0	15	33	125
Chloroflexi	0	1	0	10
Cyanobacteria	0	1	0	60
Deinococcusthermus	0	3	0	13
Firmicutes	0	2	0	690
Planctomycetes	0	1	0	6
Alphaproteobacteria	0	12	0	278
Betaproteobacteria	0	2	0	197
Gammaproteobacteria	0	10	0	730
Deltaproteobacteria	0	0	4	51
Sphingobacteriia	0	1	2	7

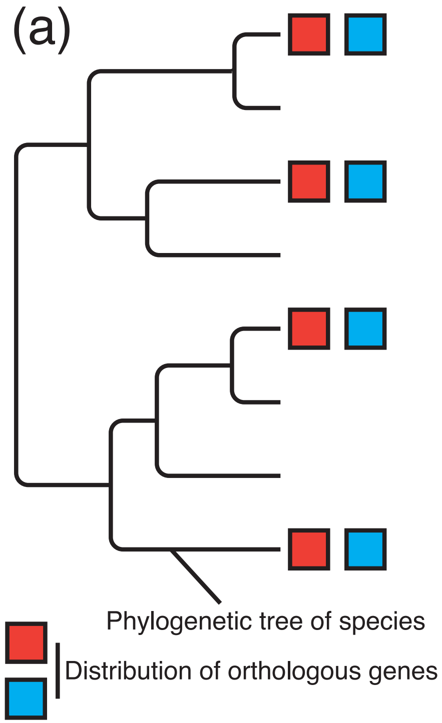

(a)
(b)
(c)
Phylogenetic tree of species
Distribution of orthologous genes

Figure 3-1. Phylogenetic distribution pattern of two genes. (a) Two genes show similar distribution pattern and polyphyletic distribution. (b) Two genes show similar distribution pattern and monophyletic distribution. (c) Two genes show opposite distribution pattern and polyphyletic distribution.

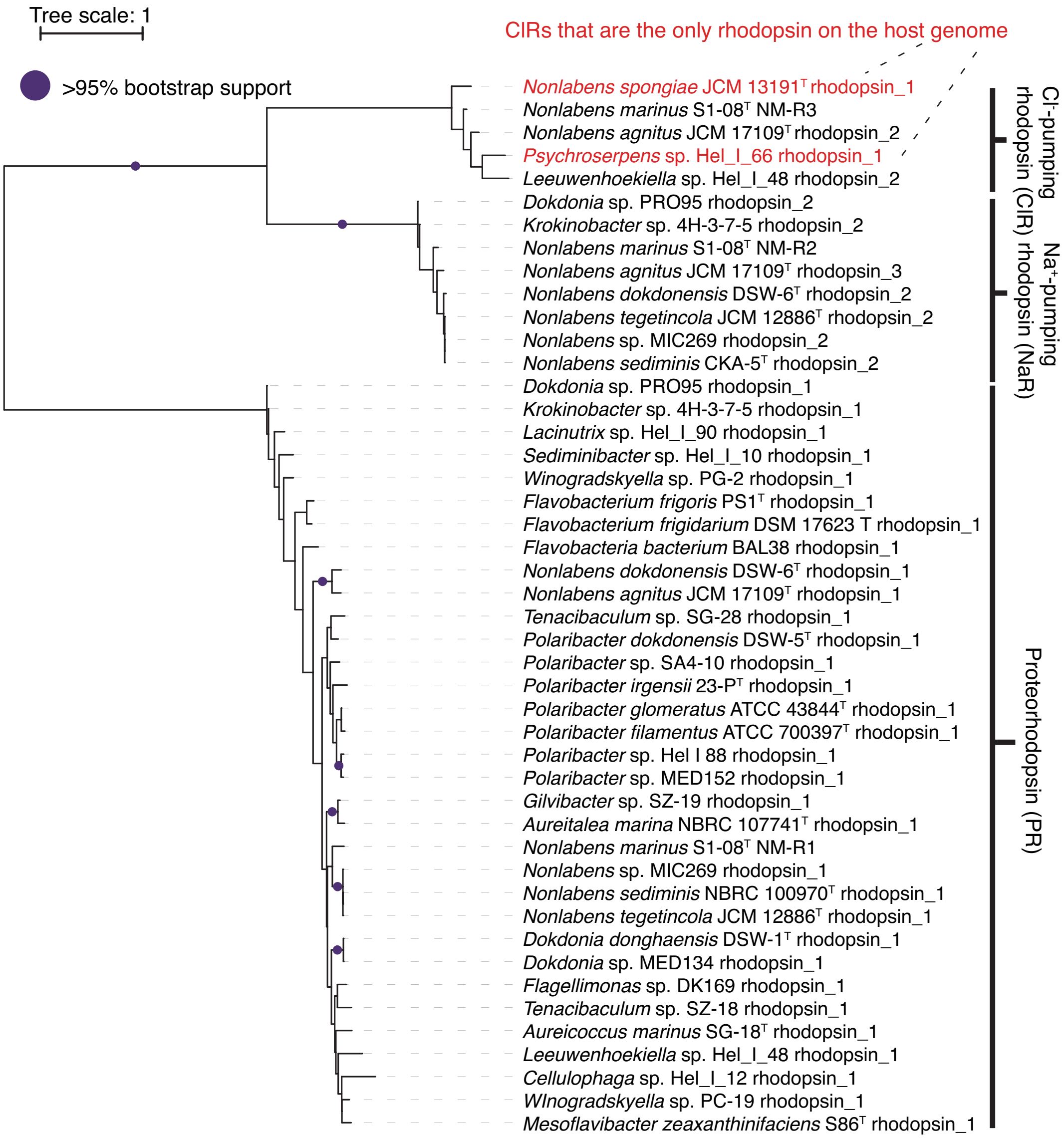

Tree scale: 1
>95% bootstrap support
ClRs that are the only rhodopsin on the host genome
Nonlabens spongiae JCM 13191T rhodopsin_1
Nonlabens marinus S1-08T NM-R3
Nonlabens agnitus JCM 17109T rhodopsin_2
Psychroserpens sp. Hel_I_66 rhodopsin_1
Leeuwenhoekiella sp. Hel_I_48 rhodopsin_2
Dokdonia sp. PRO95 rhodopsin_2
Krokinobacter sp. 4H-3-7-5 rhodopsin_2
Nonlabens marinus S1-08T NM-R2
Nonlabens agnitus JCM 17109T rhodopsin_3
Nonlabens dokdonensis DSW-6T rhodopsin_2
Nonlabens tegetincola JCM 12886T rhodopsin_2
Nonlabens sp. MIC269 rhodopsin_2
Nonlabens sediminis CKA-5T rhodopsin_2
Dokdonia sp. PRO95 rhodopsin_1
Krokinobacter sp. 4H-3-7-5 rhodopsin_1
Lacinutrix sp. Hel_I_90 rhodopsin_1
Sediminibacter sp. Hel_I_10 rhodopsin_1
Winogradskyella sp. PG-2 rhodopsin_1
Flavobacterium frigoris PS1T rhodopsin_1
Flavobacterium frigidarium DSM 17623 T rhodopsin_1
Flavobacteria bacterium BAL38 rhodopsin_1
Nonlabens dokdonensis DSW-6T rhodopsin_1
Nonlabens agnitus JCM 17109T rhodopsin_1
Tenacibaculum sp. SG-28 rhodopsin_1
Polaribacter dokdonensis DSW-5T rhodopsin_1
Polaribacter sp. SA4-10 rhodopsin_1
Polaribacter irgensii 23-PT rhodopsin_1
Polaribacter glomeratus ATCC 43844T rhodopsin_1
Polaribacter filamentus ATCC 700397T rhodopsin_1
Polaribacter sp. Hel I 88 rhodopsin_1
Polaribacter sp. MED152 rhodopsin_1
Gilvibacter sp. SZ-19 rhodopsin_1
Aureitalea marina NBRC 107741T rhodopsin_1
Nonlabens marinus S1-08T NM-R1
Nonlabens sp. MIC269 rhodopsin_1
Nonlabens sediminis NBRC 100970T rhodopsin_1
Nonlabens tegetincola JCM 12886T rhodopsin_1
Dokdonia donghaensis DSW-1T rhodopsin_1
Dokdonia sp. MED134 rhodopsin_1
Flagellimonas sp. DK169 rhodopsin_1
Tenacibaculum sp. SZ-18 rhodopsin_1
Aureicoccus marinus SG-18T rhodopsin_1
Leeuwenhoekiella sp. Hel_I_48 rhodopsin_1
Cellulophaga sp. Hel_I_12 rhodopsin_1
WInogradskyella sp. PC-19 rhodopsin_1
Mesoflavibacter zeaxanthinifaciens S86T rhodopsin_1
Cl--pumping rhodopsin (ClR)
Na+-pumping rhodopsin (NaR)
Proteorhodopsin (PR)

Figure 3-2. Phylogenetic tree of rhodopsin genes.

A maximum-likelihood tree of rhodopsin genes (CDSs annotated to bactNOG05CSB). The closed circles indicate branches with 95% bootstrapping support. Gene names in red indicate genes within genomes which contain Cl^{-}-pumping rhodopsin (ClR) genes as their only rhodopsin gene. The tree was visualized using iTol v 3.3.2.

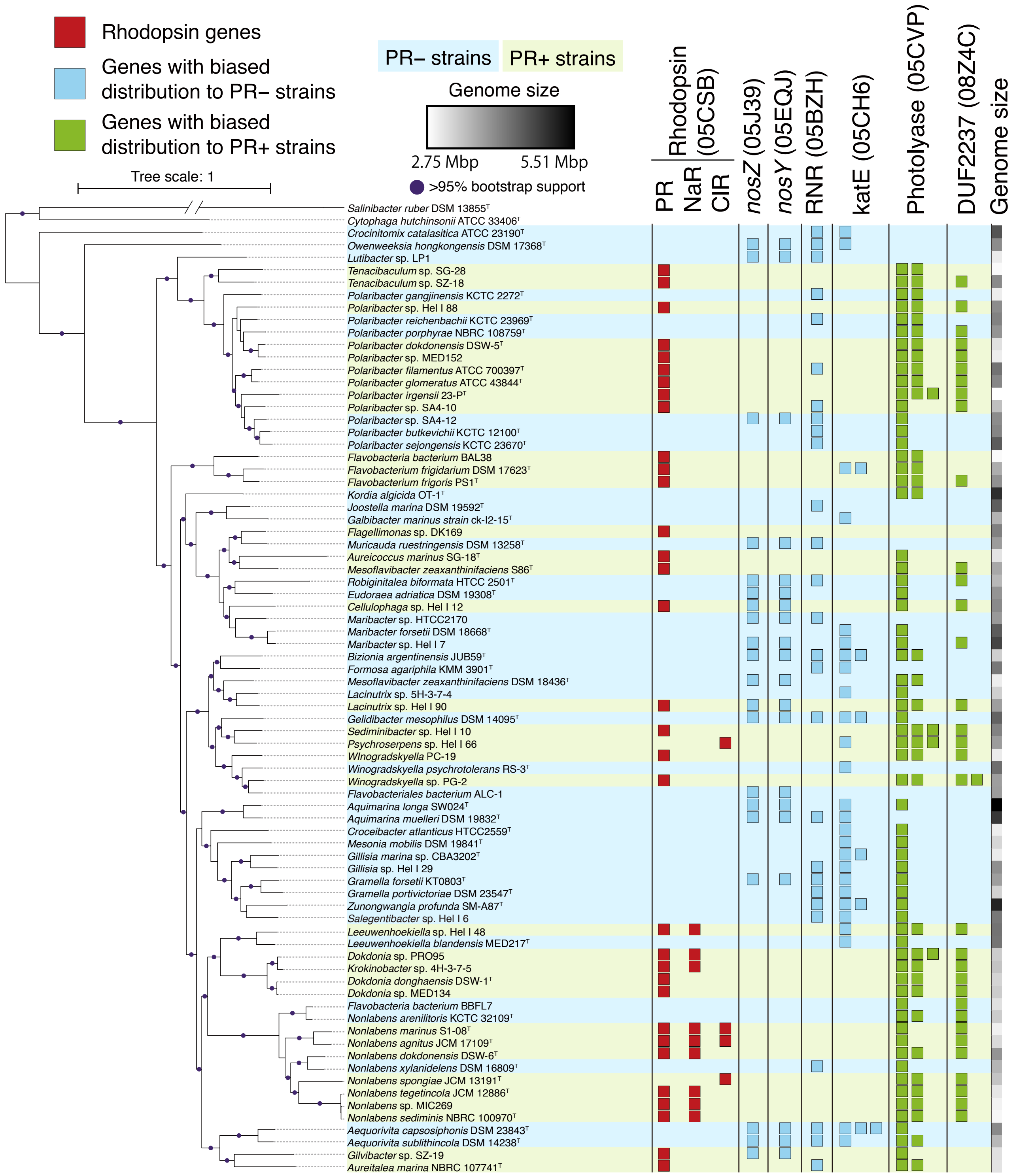


Figure 3-3. Phylogenetic tree and distributions of genes biased in PR－ or PR+ genomes.

(Left) A maximum-likelihood genomic tree based on 155 CDSs that were conserved across the 76 marine Flavobacteriia genomes. The closed circles indicate branches with 95% and more bootstrapping support. Blue and green background colours indicate PR－ and PR+ strains, respectively. The tree was visualized using iTol v 3.3.2 (Letunic and Bork 2016).

(Right) Number of genes coded by each genome is represented by the numbers of closed squares. Red: Rhodopsin genes. Light blue: eggNOG orthologue groups that showed distributions that were particularly biased to PR－ genomes. Light green: Those particularly biased to PR+ genomes. For RNR (05BZH) and photolyase (05CVP) orthologs, one gene in each strain is not shown because all strains except for *Lutibacter* sp. LP1 possess at least one copy of those orthologs. Note that distributions of genes for the synthesis and transport of APEs are shown in Fig. 3-8. Genome sizes of each strains were visualized as a grey scale heatmap.

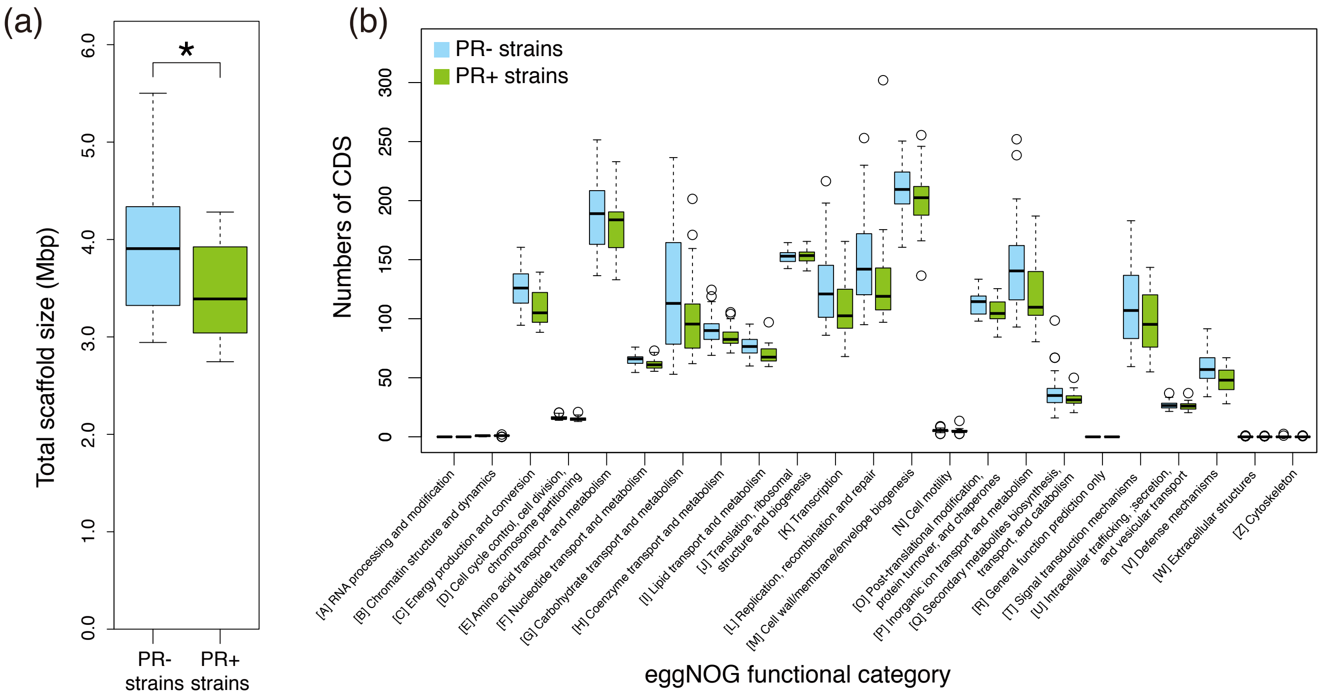


Figure 3-4. Genome sizes and quantities of CDSs in PR－ and PR+ marine Flavobacteriia.

(a) Total scaffold sizes of PR－ (blue) and PR+ (green) genomes. The bottom, central line, and top of the box plots represent the first, second, and third interquartile ranges (IQR), respectively. Whiskers represent the lowest and highest values within 1.5 × IQR from the first and third quartiles, respectively. (b) Quantities of CDSs in each eggNOG functional category in the box plot drawn in the same manner. Circles represent outliers beyond the whiskers.

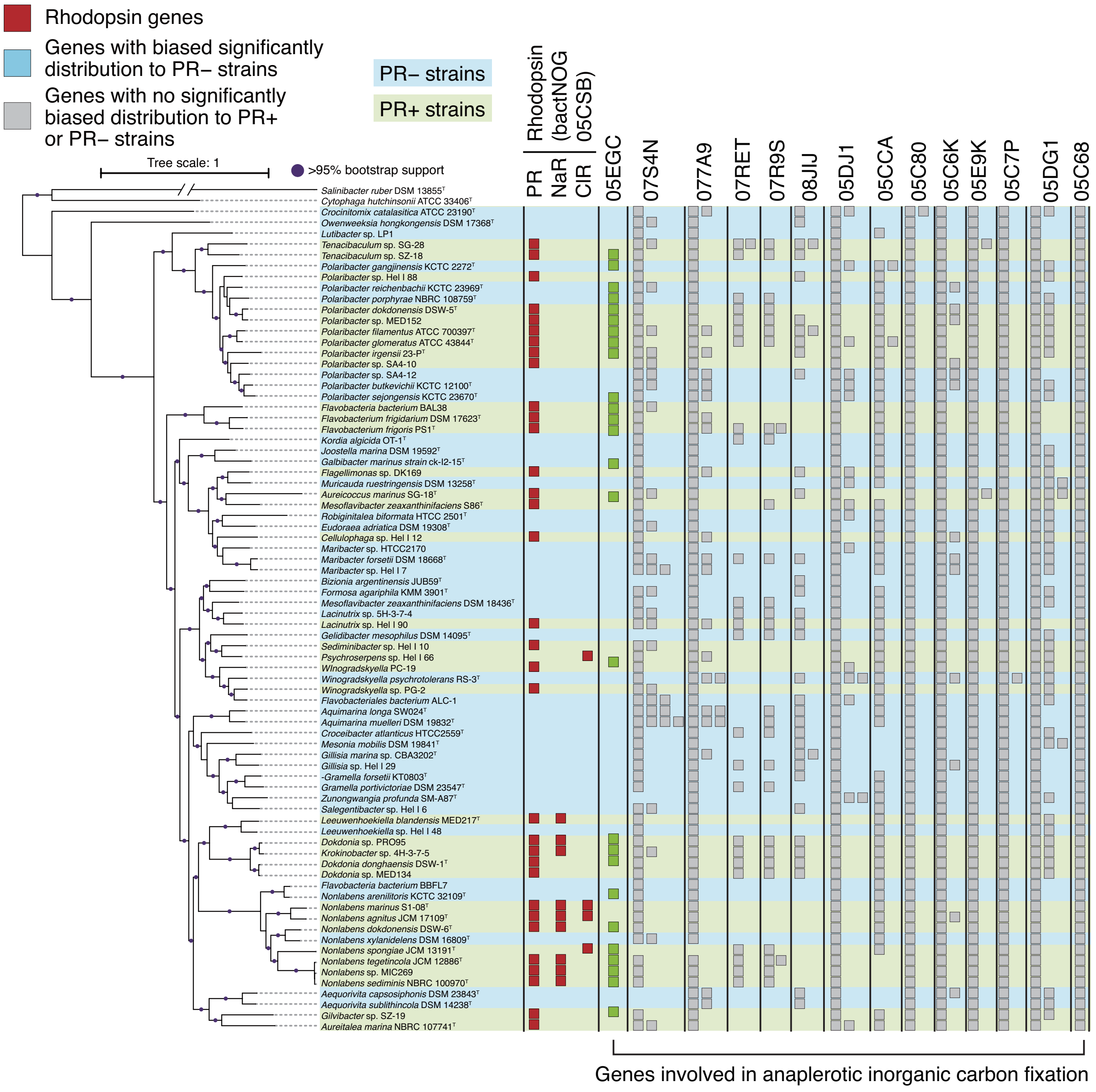


Figure 3-5. Distributions of genes involved in anaplerotic inorganic carbon fixation.

The genomic phylogenetic tree is from Fig. 3-3. The closed circles indicate branches with 95% bootstrapping support. Blue and green background colors indicate the PR− and PR+ strains, respectively. The number of genes encoded by each genome is illustrated by the number of closed squares. Red: rhodopsin genes. Light green and grey: bactNOG orthologue groups that are involved in anaplerotic inorganic carbon fixation (05EGC: *sbtA*, 07S4N: *bicA*, 077A9: carbonic anhydrase, 07RET: isocitrate lyase, 07R9S: malate synthase, 08JIJ: pyruvate carboxylase, 05DJ1: phosphoenolpyruvate carboxykinase, 05CCA: phosphoenolpyruvate carboxylase, 05C80: malate dehydrogenase, 05C6K: malate dehydrogenase, 05E9K: isocitrate dehydrogenase, and 05C7P: 2-oxoglutarate dehydrogenase subunit E1). The 05EGC group is coloured in light green because it showed distributions that were significantly biased to the PR+ genomes.

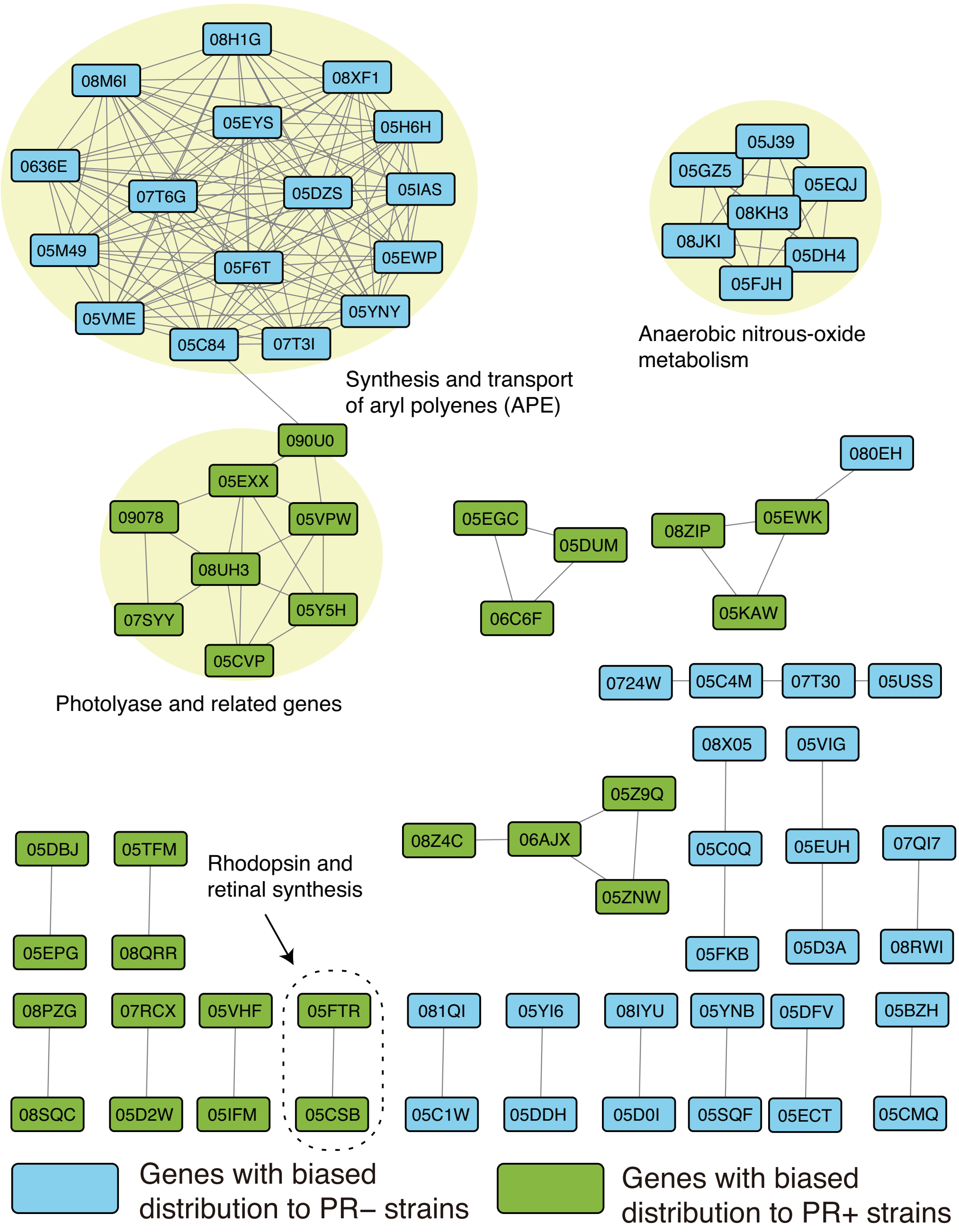


Figure 3-6. Gene proximity network of orthologue groups that showed biased distributions. Boxes in light blue and light green represent eggNOG orthologue groups that are biased to the PR－ and PR+ genomes, respectively. Note that 49 of the 129 orthologue groups showed no proximal relations and are absent from this figure. Three large clusters are indicated by the light-yellow ellipses. The PR and *blh* genes are indicated by the ellipse with a dashed border. Gene annotations are available in Supplementary Tables S4 and S5.

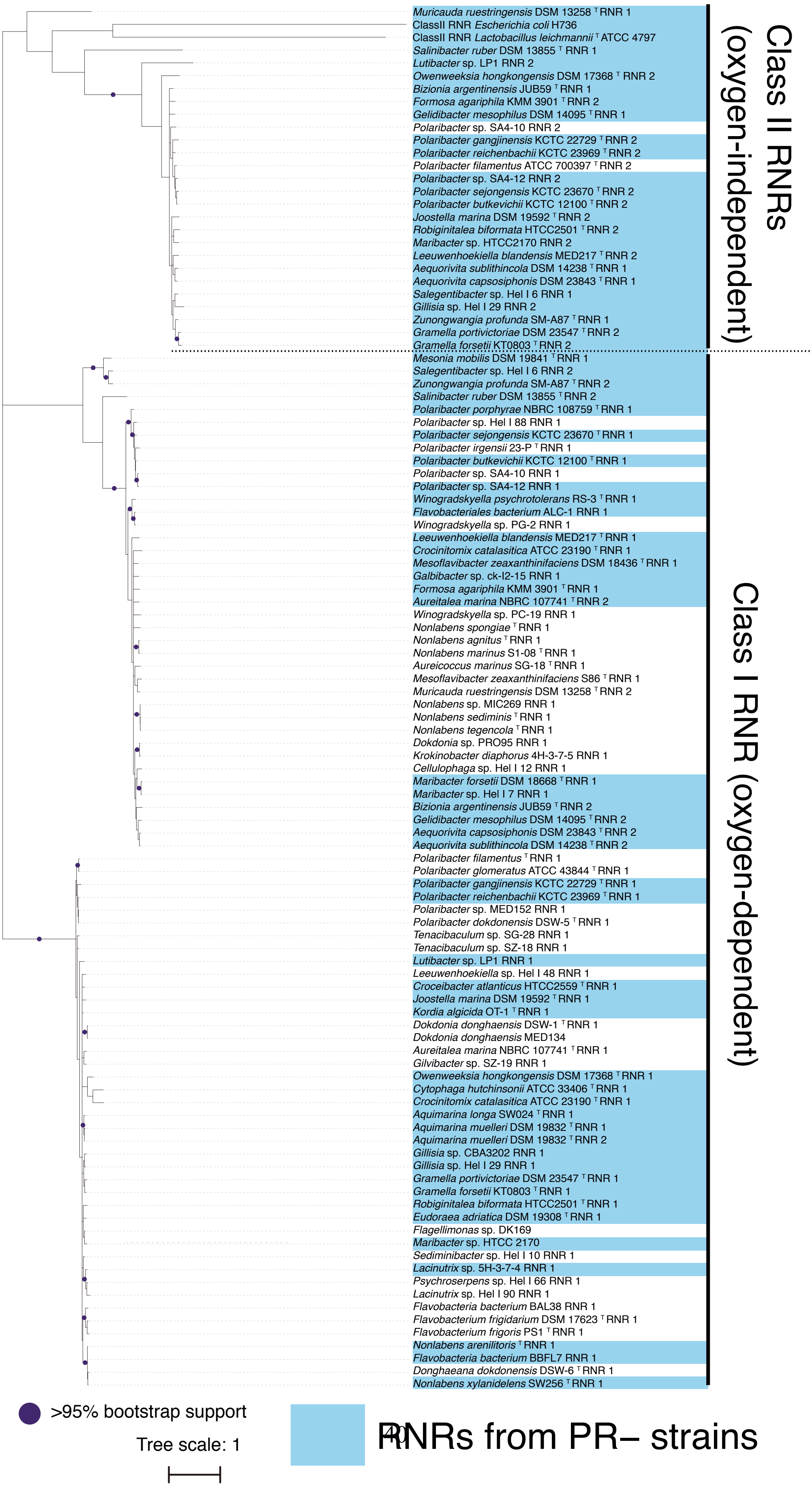

Muricauda ruestringensis DSM 13258 T RNR 1
ClassII RNR Escherichia coli H736
ClassII RNR Lactobacillus leichmannii T ATCC 4797
Salinibacter ruber DSM 13855 T RNR 1
Lutibacter sp. LP1 RNR 2
Owenweeksia hongkongensis DSM 17368 T RNR 2
Bizionia argentinensis JUB59 T RNR 1
Formosa agariphila KMM 3901 T RNR 2
Gelidibacter mesophilus DSM 14095 T RNR 1
Polaribacter sp. SA4-10 RNR 2
Polaribacter gangjinensis KCTC 22729 T RNR 2
Polaribacter reichenbachii KCTC 23969 T RNR 2
Polaribacter filamentus ATCC 700397 T RNR 2
Polaribacter sp. SA4-12 RNR 2
Polaribacter sejongensis KCTC 23670 T RNR 2
Polaribacter butkevichii KCTC 12100 T RNR 2
Joostella marina DSM 19592 T RNR 2
Robiginitalea biformata HTCC2501 T RNR 2
Maribacter sp. HTCC2170 RNR 2
Leeuwenhoekiella blandensis MED217 T RNR 2
Aequorivita sublithincola DSM 14238 T RNR 1
Aequorivita capsosiphonis DSM 23843 T RNR 1
Salegentibacter sp. Hel I 6 RNR 1
Gillisia sp. Hel I 29 RNR 2
Zunongwangia profunda SM-A87 T RNR 1
Gramella portivictoriae DSM 23547 T RNR 2
Gramella forsetii KT0803 T RNR 2
Mesonia mobilis DSM 19841 T RNR 1
Salegentibacter sp. Hel I 6 RNR 2
Zunongwangia profunda SM-A87 T RNR 2
Salinibacter ruber DSM 13855 T RNR 2
Polaribacter porphyrae NBRC 108759 T RNR 1
Polaribacter sp. Hel I 88 RNR 1
Polaribacter sejongensis KCTC 23670 T RNR 1
Polaribacter irgensii 23-P T RNR 1
Polaribacter butkevichii KCTC 12100 T RNR 1
Polaribacter sp. SA4-10 RNR 1
Polaribacter sp. SA4-12 RNR 1
Winogradskyella psychrotolerans RS-3 T RNR 1
Flavobacteriales bacterium ALC-1 RNR 1
Winogradskyella sp. PG-2 RNR 1
Leeuwenhoekiella blandensis MED217 T RNR 1
Crocinitomix catalasitica ATCC 23190 T RNR 1
Mesoflavibacter zeaxanthinifaciens DSM 18436 T RNR 1
Galbibacter sp. ck-I2-15 RNR 1
Formosa agariphila KMM 3901 T RNR 1
Aureitalea marina NBRC 107741 T RNR 2
Winogradskyella sp. PC-19 RNR 1
Nonlabens spongiae T RNR 1
Nonlabens agnitus T RNR 1
Nonlabens marinus S1-08 T RNR 1
Aureicoccus marinus SG-18 T RNR 1
Mesoflavibacter zeaxanthinifaciens S86 T RNR 1
Muricauda ruestringensis DSM 13258 T RNR 2
Nonlabens sp. MIC269 RNR 1
Nonlabens sediminis T RNR 1
Nonlabens tegencola T RNR 1
Dokdonia sp. PRO95 RNR 1
Krokinobacter diaphorus 4H-3-7-5 RNR 1
Cellulophaga sp. Hel I 12 RNR 1
Maribacter forsetii DSM 18668 T RNR 1
Maribacter sp. Hel I 7 RNR 1
Bizionia argentinensis JUB59 T RNR 2
Gelidibacter mesophilus DSM 14095 T RNR 2
Aequorivita capsosiphonis DSM 23843 T RNR 2
Aequorivita sublithincola DSM 14238 T RNR 2
Polaribacter filamentus T RNR 1
Polaribacter glomeratus ATCC 43844 T RNR 1
Polaribacter gangjinensis KCTC 22729 T RNR 1
Polaribacter reichenbachii KCTC 23969 T RNR 1
Polaribacter sp. MED152 RNR 1
Polaribacter dokdonensis DSW-5 T RNR 1
Tenacibaculum sp. SG-28 RNR 1
Tenacibaculum sp. SZ-18 RNR 1
Lutibacter sp. LP1 RNR 1
Leeuwenhoekiella sp. Hel I 48 RNR 1
Croceibacter atlanticus HTCC2559 T RNR 1
Joostella marina DSM 19592 T RNR 1
Kordia algicida OT-1 T RNR 1
Dokdonia donghaensis DSW-1 T RNR 1
Dokdonia donghaensis MED134
Aureitalea marina NBRC 107741 T RNR 1
Gilvibacter sp. SZ-19 RNR 1
Owenweeksia hongkongensis DSM 17368 T RNR 1
Cytophaga hutchinsonii ATCC 33406 T RNR 1
Crocinitomix catalasitica ATCC 23190 T RNR 1
Aquimarina longa SW024 T RNR 1
Aquimarina muelleri DSM 19832 T RNR 1
Aquimarina muelleri DSM 19832 T RNR 2
Gillisia sp. CBA3202 RNR 1
Gillisia sp. Hel I 29 RNR 1
Gramella portivictoriae DSM 23547 T RNR 1
Gramella forsetii KT0803 T RNR 1
Robiginitalea biformata HTCC2501 T RNR 1
Eudoraea adriatica DSM 19308 T RNR 1
Flagellimonas sp. DK169
Maribacter sp. HTCC 2170
Sediminibacter sp. Hel I 10 RNR 1
Lacinutrix sp. 5H-3-7-4 RNR 1
Psychroserpens sp. Hel I 66 RNR 1
Lacinutrix sp. Hel I 90 RNR 1
Flavobacteria bacterium BAL38 RNR 1
Flavobacterium frigidarium DSM 17623 T RNR 1
Flavobacterium frigoris PS1 T RNR 1
Nonlabens arenilitoris T RNR 1
Flavobacteria bacterium BBFL7 RNR 1
Donghaeana dokdonensis DSW-6 T RNR 1
Nonlabens xylanidelens SW256 T RNR 1
Class II RNRs
(oxygen-independent)
Class I RNR (oxygen-dependent)
● >95% bootstrap support
Tree scale: 1
RNRs from PR− strains

Figure 3-7. Phylogenetic tree of ribonucleotide reductase (RNR) genes.

The RNR classes, which are consistent with the phylogenetic tree organization, were identified by using an NCBI conserved domain search. The light blue background colour indicates those genes in PR－ genomes (otherwise, PR+ genomes). Any branch with less than 30% bootstrap support was removed, and those with more than 95% bootstrap support are represented by purple circles.

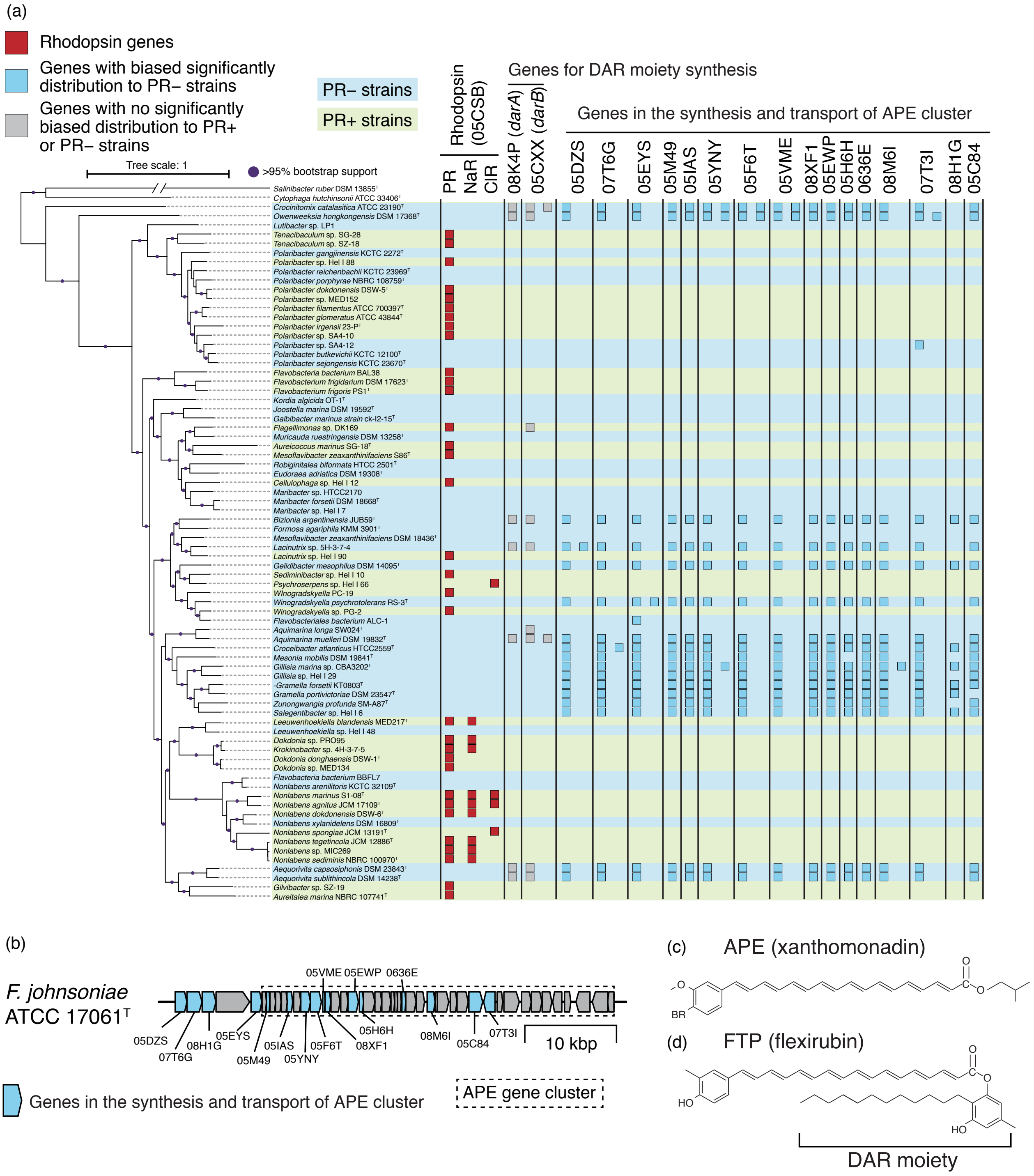

(a)
Rhodopsin genes
Genes with biased significantly distribution to PR– strains
Genes with no significantly biased distribution to PR+ or PR– strains
PR– strains
PR+ strains
Tree scale: 1
>95% bootstrap support
Rhodopsin
PR
NaR (05CSB)
ClR
Genes for DAR moiety synthesis
08K4P (darA)
05CXX (darB)
Genes in the synthesis and transport of APE cluster
05DZS
07T6G
05EYS
05M49
05IAS
05YNY
05F6T
05VME
08XF1
05EWP
05H6H
0636E
08M6I
07T3I
08H1G
05C84
Salinibacter ruber DSM 13855T
Cytophaga hutchinsonii ATCC 33406T
Crocinitomix catalasitica ATCC 23190T
Owenweeksia hongkongensis DSM 17368T
Lutibacter sp. LP1
Tenacibaculum sp. SG-28
Tenacibaculum sp. SZ-18
Polaribacter gangjinensis KCTC 2272T
Polaribacter sp. Hel I 88
Polaribacter reichenbachii KCTC 23969T
Polaribacter porphyrae NBRC 108759T
Polaribacter dokdonensis DSW-5T
Polaribacter sp. MED152
Polaribacter filamentus ATCC 700397T
Polaribacter glomeratus ATCC 43844T
Polaribacter irgensii 23-PT
Polaribacter sp. SA4-10
Polaribacter sp. SA4-12
Polaribacter butkevichii KCTC 12100T
Polaribacter sejongensis KCTC 23670T
Flavobacteria bacterium BAL38
Flavobacterium frigidarium DSM 17623T
Flavobacterium frigoris PS1T
Kordia algicida OT-1T
Joostella marina DSM 19592T
Galbibacter marinus strain ck-I2-15T
Flagellimonas sp. DK169
Muricauda ruestringensis DSM 13258T
Aureicoccus marinus SG-18T
Mesoflavibacter zeaxanthinifaciens S86T
Robiginitalea biformata HTCC 2501T
Eudoraea adriatica DSM 19308T
Cellulophaga sp. Hel I 12
Maribacter sp. HTCC2170
Maribacter forsetii DSM 18668T
Maribacter sp. Hel I 7
Bizionia argentinensis JUB59T
Formosa agariphila KMM 3901T
Mesoflavibacter zeaxanthinifaciens DSM 18436T
Lacinutrix sp. 5H-3-7-4
Lacinutrix sp. Hel I 90
Gelidibacter mesophilus DSM 14095T
Sediminibacter sp. Hel I 10
Psychroserpens sp. Hel I 66
WInogradskyella PC-19
Winogradskyella psychrotolerans RS-3T
Winogradskyella sp. PG-2
Flavobacteriales bacterium ALC-1
Aquimarina longa SW024T
Aquimarina muelleri DSM 19832T
Croceibacter atlanticus HTCC2559T
Mesonia mobilis DSM 19841T
Gillisia marina sp. CBA3202T
Gillisia sp. Hel I 29
-Gramella forsetii KT0803T
Gramella portivictoriae DSM 23547T
Zunongwangia profunda SM-A87T
Salegentibacter sp. Hel I 6
Leeuwenhoekiella blandensis MED217T
Leeuwenhoekiella sp. Hel I 48
Dokdonia sp. PRO95
Krokinobacter sp. 4H-3-7-5
Dokdonia donghaensis DSW-1T
Dokdonia sp. MED134
Flavobacteria bacterium BBFL7
Nonlabens arenilitoris KCTC 32109T
Nonlabens marinus S1-08T
Nonlabens agnitus JCM 17109T
Nonlabens dokdonensis DSW-6T
Nonlabens xylanidelens DSM 16809T
Nonlabens spongiae JCM 13191T
Nonlabens tegetincola JCM 12886T
Nonlabens sp. MIC269
Nonlabens sediminis NBRC 100970T
Aequorivita capsosiphonis DSM 23843T
Aequorivita sublithincola DSM 14238T
Gilvibacter sp. SZ-19
Aureitalea marina NBRC 107741T
(b)
F. johnsoniae
ATCC 17061T
05VME 05EWP 0636E
05DZS
08H1G
07T6G
05EYS
05M49
05IAS
05YNY
05F6T
08XF1
05H6H
08M6I
05C84
07T3I
10 kbp
Genes in the synthesis and transport of APE cluster
APE gene cluster
(c) APE (xanthomonadin)
BR
O
(d) FTP (flexirubin)
HO
HO
DAR moiety

Figure 3-8. Analysis of genes involved in the synthesis and transport of APEs. (a) Distributions of genes involved in synthesis and transport of APEs and DAR. The genomic phylogenetic tree is from Fig. 3-3. The closed circles indicate branches with 95% bootstrapping support. Blue and green background colours indicate PR－ and PR+ strains, respectively. Number of genes coded by each

genome is illustrated by the number of closed squares. Red: rhodopsin genes. Grey: eggNOG orthologue groups involved in the synthesis of the DAR moiety (05CXX: *darB* and 08K4P: *darA*) and did not show biased distributions. Light blue: eggNOG orthologue groups in the cluster related to the synthesis and transport of APEs (05DZS: phenylacetate-CoA ligase, 07T6G: acyl-coenzyme A 6-aminopenicillanic acid acyl-transferase, 05EYS: glycosyl transferase, family 2, 05M49: dehydratase, 05IAS: outer membrane lipoprotein carrier protein LolA, 05YNY: synthase, 05F6T: synthase, 05VME, acyl carrier protein, 08XF1: NA, 05EWP: synthase, 05H6H: thioesterase, 0636E: flexirubin-type pigment biosynthesis acyl carrier protein, 08M6I: lipid A biosynthesis acyltransferase, 07T31: 5'-nucleotidase, 08H1G: phospholipid glycerol acyltransferase, and 05C84: histidine ammonia-lyase) that showed distributions that were significantly biased to the PR– genomes. (b) Syntenic map of the APE gene cluster of *F. johnsoniae* ATCC 17061^{T}. Pentagons represent genes, and pentagon lengths are proportional to the gene lengths. Genes in the cluster related to the synthesis and transport of APEs in Fig. 3-6 are shown in light blue. The previously reported APE gene cluster of *F. johnsoniae* ATCC 17061^{T} (between *Fjoh_1080* and *Fjoh_1115* genes) is represented by a rectangle with a dashed border. (c) Structure of a representative APE molecule (xanthomonadin). (d) Structure of a representative FTP molecule (flexirubin).

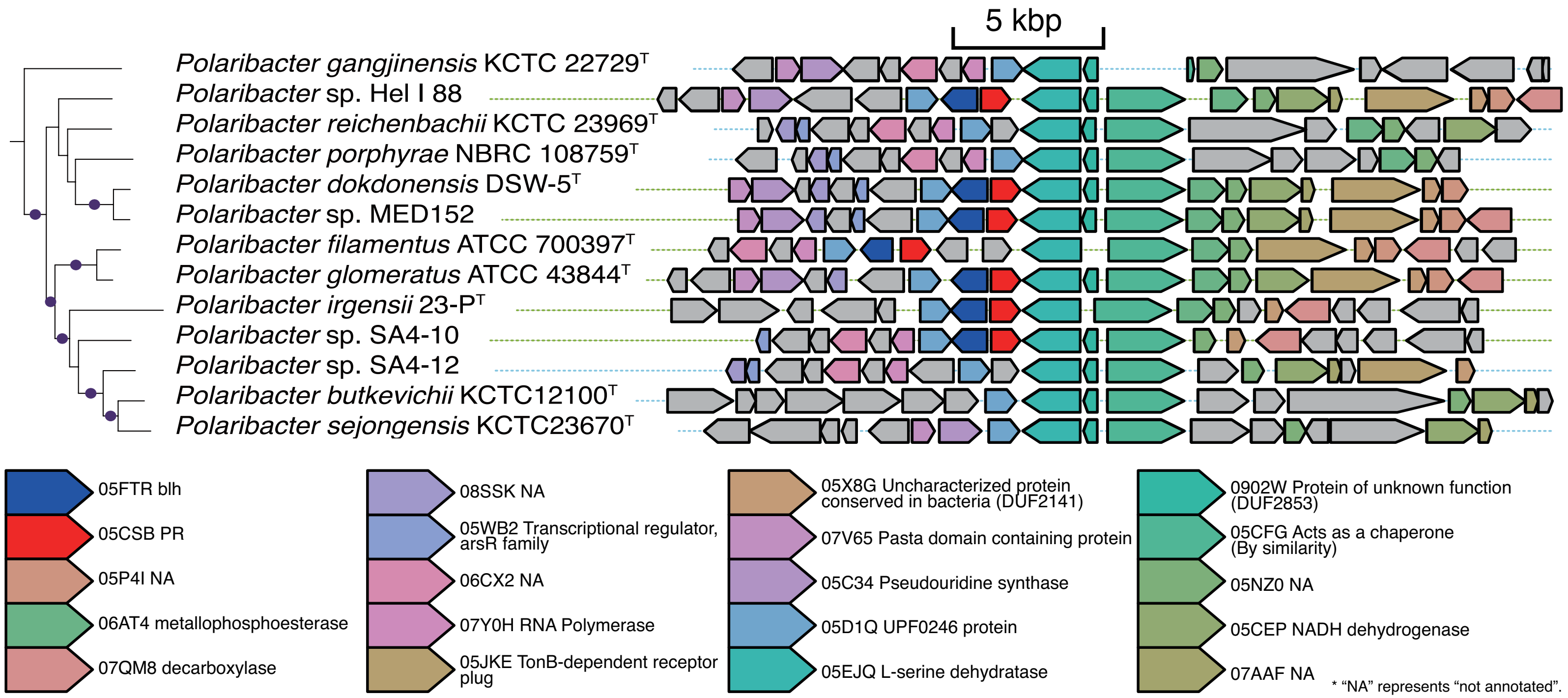

5 kbp
Polaribacter gangjinensis KCTC 22729^T
Polaribacter sp. Hel I 88
Polaribacter reichenbachii KCTC 23969^T
Polaribacter porphyrae NBRC 108759^T
Polaribacter dokdonensis DSW-5^T
Polaribacter sp. MED152
Polaribacter filamentus ATCC 700397^T
Polaribacter glomeratus ATCC 43844^T
Polaribacter irgensii 23-P^T
Polaribacter sp. SA4-10
Polaribacter sp. SA4-12
Polaribacter butkevichii KCTC12100^T
Polaribacter sejongensis KCTC23670^T
05FTR blh
05CSB PR
05P4I NA
06AT4 metallophosphoesterase
07QM8 decarboxylase
08SSK NA
05WB2 Transcriptional regulator, arsR family
06CX2 NA
07Y0H RNA Polymerase
05JKE TonB-dependent receptor plug
05X8G Uncharacterized protein conserved in bacteria (DUF2141)
07V65 Pasta domain containing protein
05C34 Pseudouridine synthase
05D1Q UPF0246 protein
05EJQ L-serine dehydratase
0902W Protein of unknown function (DUF2853)
05CFG Acts as a chaperone (By similarity)
05NZ0 NA
05CEP NADH dehydrogenase
07AAF NA
* “NA” represents “not annotated”.

Figure 3-9. Syntenic map of region around PR genes in the *Polaribacter* genomes.

A part of the genomic phylogenetic tree is from Fig. 3-3. The closed circles indicate branches with 95% bootstrapping support. Blue and green horizontal dotted lines indicate PR－ and PR+ strains, respectively. Pentagons represent genes, and pentagon lengths are proportional to the gene lengths. The conserved genes are shown in colours based on their annotation. Other genes are shown in grey. Because it is highly unlikely that the PR and *blh* genes were independently and repeatedly acquired next to the L-serine dehydratase gene by chance, the PR and *blh* genes are assumed to have been lost during evolution.

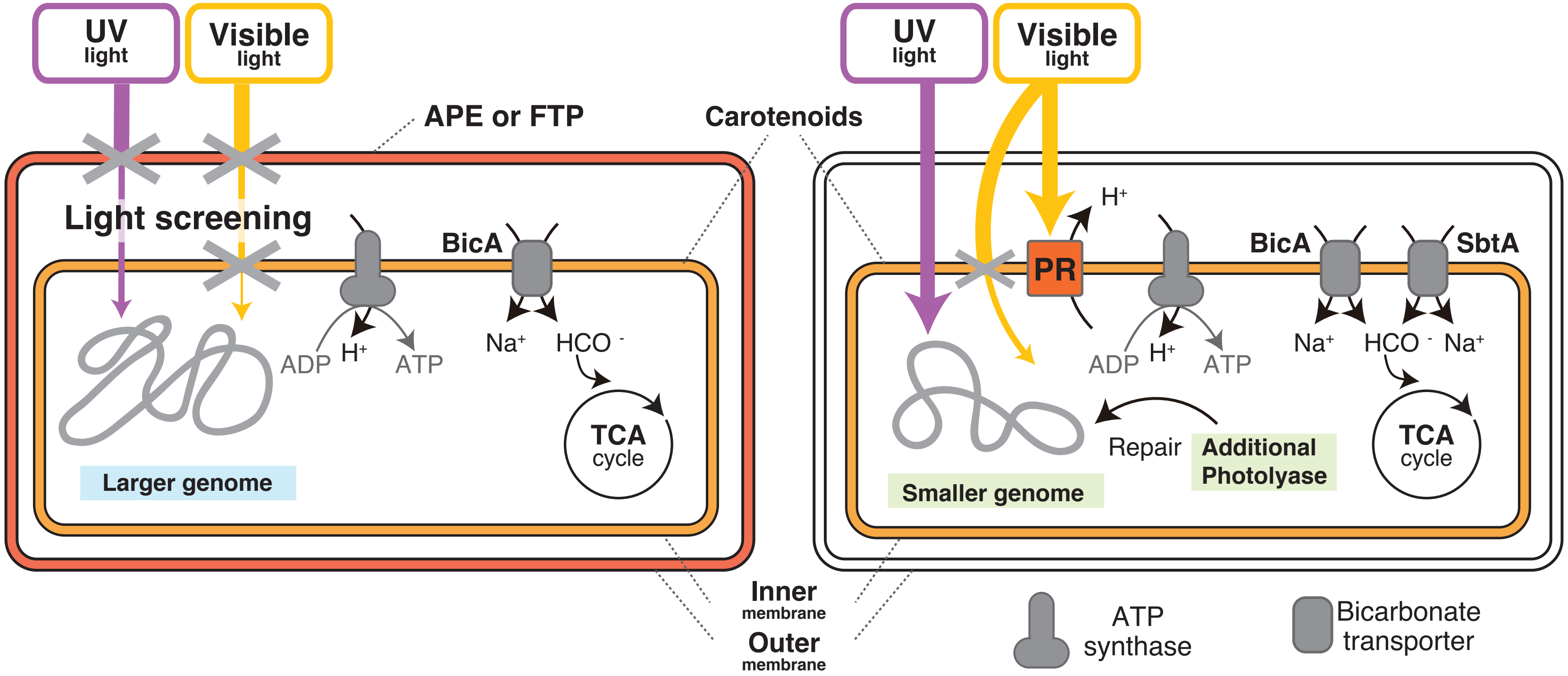


Figure 3-10. Schematic figure of the adaptive strategies of PR－ and PR+ Flavobacteriia.

The background colours in light blue and green represent characteristics of the PR－ and PR+ marine Flavobacteriia, respectively. PR－ strains have APEs or FTPs in the outer membrane to block UV and visible light. On the other hand, PR+ strains have neither APEs nor FTPs, but their PR can utilize visible light in the inner membrane. UV light that reaches the DNA in PR+ strains causes DNA damage, which is repaired by photolyases but leads to the smaller genome size of the PR+ strains. One gene that is involved in anaplerotic inorganic carbon fixation (*sbtA*) is biased to PR+ genomes. The PR－ strains show signs of adaptation to anaerobic conditions

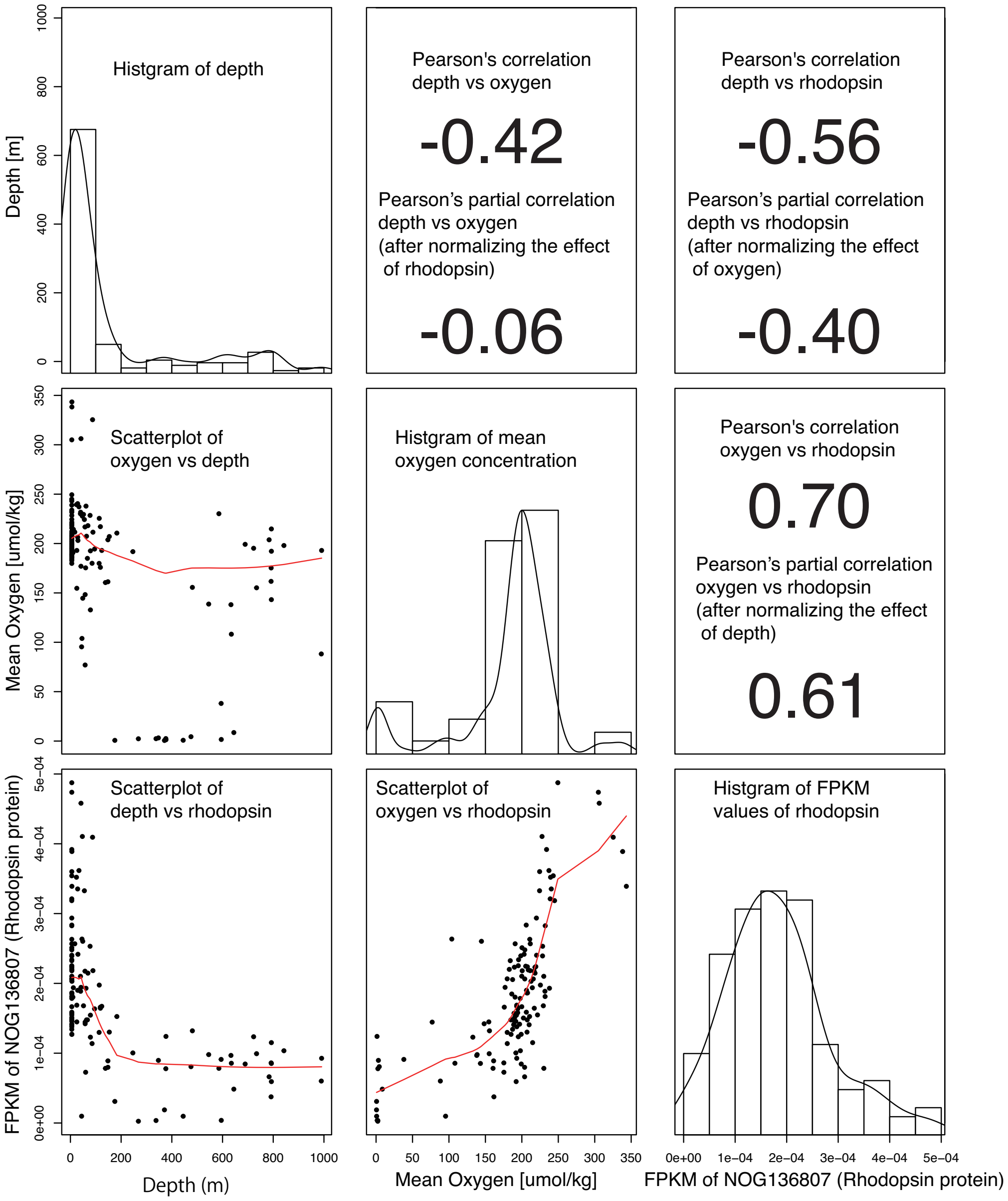


Figure 3-11. Relationships between depth, oxygen concentration, and abundance of rhodopsin genes in the metagenomic dataset of Tara Oceans samples.

The rhodopsin gene abundances were quantified by their Fragments Per Kilobase of exon per Million fragments (FPKM) values. Histograms of the three values, scatterplots, Pearson's correlation values, and Pearson's partial correlation values are shown.

Chapter 4. Experimental validation of phylogenetic profiling results

Introduction

In chapter 3, I have discussed based on the individual annotation of detected genes. However, comparative genomics is also a powerful technique for predicting the function of genes with no similarity to any gene yet studied (Haft 2015). Especially, in the field of environmental microbiology, there are many unannotated hypothetical genes and such post-homology methods are required (DeLong 2009). My comparative genomic analysis reveals the different adaptation strategies to light between PR– and PR+ Flavobacteriia, however, it should be considered that statistical correlation and inverse correlation are not equal to causation. Experimental verification is necessary to know whether the obtained findings truly reflect the ecophysiology of bacteria. In order to find out whether the detected gene is really related to light utilization, I did a genetic manipulation-based experiment of DUF2237, a functionally unknown gene biased in PR+ bacteria in this chapter. I selected this gene because it showed the highest correlation to PR gene and this gene is distributed among other prokaryotes with different light utilizing system, diverse rhodopsins and oxygenic and anoxygenic photosynthesis. Because of the difficulty of genetic manipulation of Flavobacteriia, I used model cyanobacteria *Synechocystis* sp. PCC 6803-P for this experiment (Yoshihara and Ikeuchi 2004).

Material and Methods

Construction of DUF2237 deletion mutant

The DUF2237 gene of *Synechocystis* sp. PCC 6803-P (i.e., *slr*1628) (Nakamura et al. 1998) was inactivated by replacing it with a chloramphenicol resistance cassette. A DNA sequence that contained the region that is 500 bp upstream of the DUF2237 gene, a chloramphenicol resistance cassette, and 500 bp downstream of DUF2237 was artificially synthesized and inserted into a pEX-A2 vector (Eurofins Genomics). Knockout strains were generated by transforming this plasmid into PCC 6803-P cells, grown at 30°C under continuous white light with an intensity of 50 μmol m^{-2} s^{-1}, and selected on plates with BG-11 medium (Stanier et al. 1971) that contains 20 μg $m1^{-1}$ chloramphenicol. Because PCC 6803-P cells contain multiple genomes in each cell, the segregation between the wild-type and DUF2237-knockout genomes was examined by PCR with DUF2237-upstream (5'-AAT CTC TGC TAG GTT TGG -3') and DUF2237-downstream (5'-AAC TCT GGT AGC TGT TCC-3') primers after 3 days of growth on the BG-11 plates.

Phototaxis assay using DUF2237 deletion mutant

For the phototaxis assay, wild-type and DUF2237-knockout cells were collected in the exponential phase, suspended in BG-11 liquid medium at an optical density of 0.1, and spotted onto 1.5% agarose BG-11 plates four times per strain. The spotted plates were incubated under unidirectional white light with an intensity of 22 μmol m^{-2} s^{-1} at 30°C for seven days, and the distances of colony movements were measured.

Results

DUF2237 had biased distribution in prokaryotes possessing rhodopsins or photosystem II

The orthologue group that showed the second most biased distribution contained the DUF2237 genes (q-value = 3.9E-10), which were functionally unknown and enriched in the PR+ genomes (Fig. 3-3 and Supplementary Table S3-1). Using the Microbial Genome Database for Comparative Analysis (MBGD) (Uchiyama 2003), I found that DUF2237 genes (MBGD ID 4444) are broadly distributed across 11 phyla, and many Cyanobacteria, phototrophic bacteria, and rhodopsin-containing Euryarchaeota have this gene. The sequence of the DUF2237 gene is highly conserved across different phyla (Fig. 4-1 (a)). MBGD analysis showed that DUF2237 is possessed by 72 and 66% of prokaryotes that have photosystem II (*pufM*/*psbA*/*pufL*, MBGD ID 2841) and rhodopsin genes (MBGD ID 22185 and 4672), respectively, whereas only 17% of all prokaryotes have DUF2237 (Fig. 4-1 (b)). This bias was not just because Cyanobacteria tend to have DUF2237 (i.e., phylogenetic constraint); I confirmed that excluding Cyanobacteria did not diminish the observed bias (Fig. 4-1 (b)). These observations strongly suggested that DUF2237 has a conserved function that is related to phototrophy.

DUF2237 is involved in cyanobacterial phototaxis

To experimentally confirm the functional importance of DUF2237, I knocked its gene out of *Synechocystis* sp. PCC6803-P (Yoshihara et al. 2000). The DUF2237-knockout strain did not show any apparent difference in proliferation speed and other phenotypes under standard laboratory culture conditions; however, in phototaxis assays, the DUF2237-knockout strain showed significantly less movement than the wild-type strain, which exhibits positive phototaxis under unidirectional white light (p-value = 2.9E-04, Fig. 4-1 (c, d)).

Discussion

The phototaxis-related function of DUF2237 is consistent with the strong correlation between the presence of DUF2237 and phototrophy because phototaxis should be beneficial to organisms that utilize light. While cyanobacterial phototaxis is a phenotype in which many proteins are involved (e.g., light sensing, signal transduction, transcriptional regulation, and pilus formation proteins)

(Bhaya 2004, Yoshihara et al. 2000, Yoshihara and Ikeuchi 2004) and further analyses are required to clarify the molecular basis of the DUF2237 function, this result proves that my comparative genomics approach is powerful enough to find genes that reflect microbial ecophysiology. The biased distribution of DUF2237 in the prokaryotes with different light utilization mechanisms (photosynthesis and rhodopsin) suggests that they have common mechanisms for light adaptation. Further investigation of the molecular function of DUF2237 should provide a clue to understanding the general characteristics of light utilizing prokaryotes.

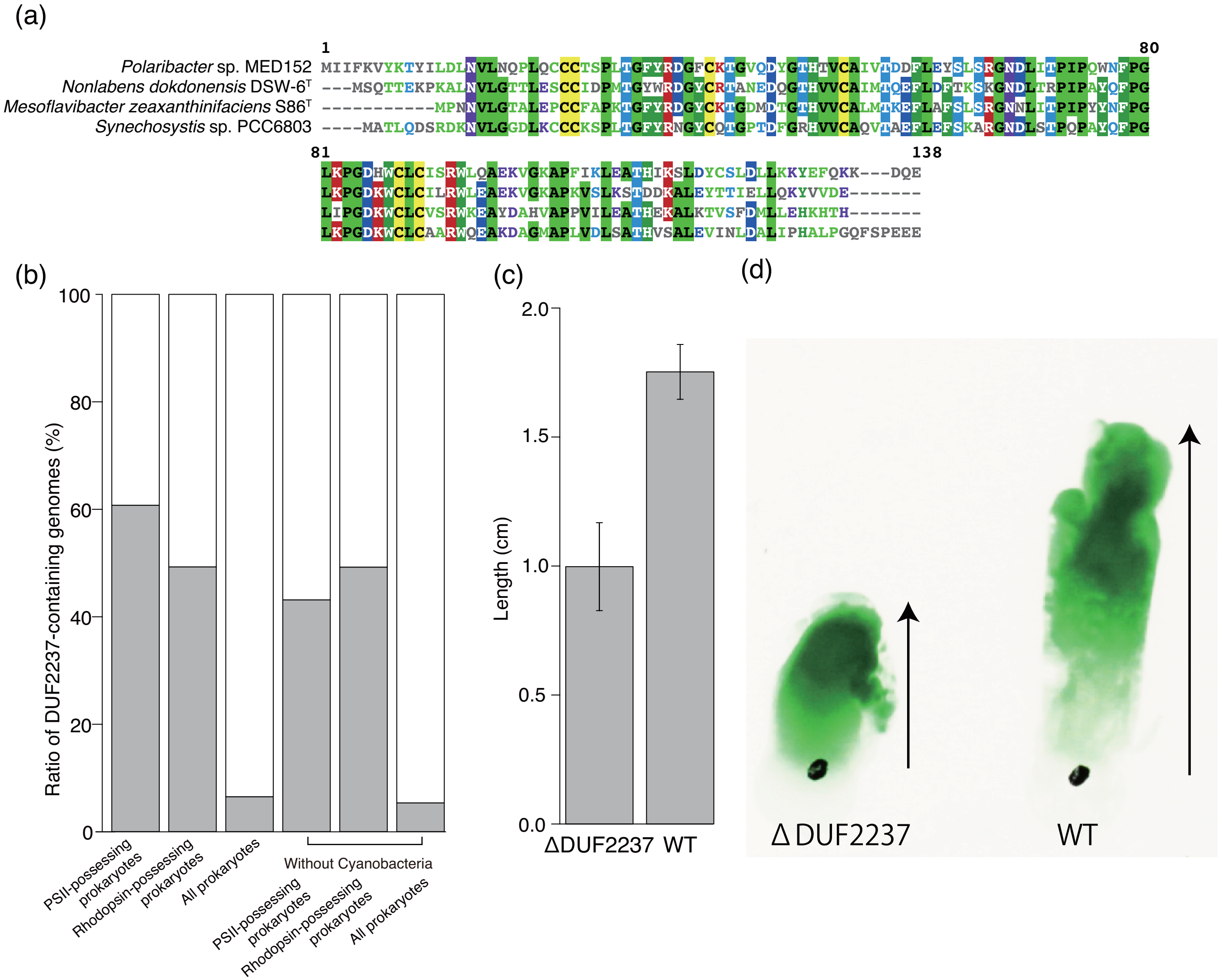

(a)
1
80
Polaribacter sp. MED152 MIIFKVYKTYILDLNVLNQPLQCCCTSPLTGFYRDGFCKTGVQDYGTHTVCAIVTDDFLEYSLSRGNDLITPIPQWNFPG
Nonlabens dokdonensis DSW-6T ---MSQTTEKPKALNVLGTTLESCCIDPMTGYWRDGYCRTANEDQGTHVVCAIMTQEFLDFTKSKGNDLTRPIPAYQFPG
Mesoflavibacter zeaxanthinifaciens S86T -----------MPNNVLGTALEPCCFAPKTGFYRDGYCKTGDMDTGTHVVCALMTKEFLAFSLSRGNNLITPIPYYNFPG
Synechosystis sp. PCC6803 ----MATLQDSRDKNVLGGDLKCCCKSPLTGFYRNGYCQTGPTDFGRHVVCAQVTAEFLEFSKARGNDLSTPQPAYQFPG
81
138
LKPGDHWCLCISRWLQAEKVGKAPFIKLEATHIKSLDYCSLDLLKKYEFQKK---DQE
LKPGDKWCLCILRWLEAEKVGKAPKVSLKSTDDKALEYTTIELLQKYVVDE-------
LIPGDKWCLCVSRWKEAYDAHVAPPVILEATHEKALKTVSFDMLLEHKHTH-------
LKPGDKWCLCAARWQEAKDAGMAPLVDLSATHVSALEVINLDALIPHALPGQFSPEEE
(b)
100
80
60
40
20
0
Ratio of DUF2237-containing genomes (%)
PSII-possessing prokaryotes
Rhodopsin-possessing prokaryotes
All prokaryotes
Without Cyanobacteria
PSII-possessing prokaryotes
Rhodopsin-possessing prokaryotes
All prokaryotes
(c)
2.0
1.5
1.0
0.5
0.0
Length (cm)
ΔDUF2237
WT
(d)
Δ DUF2237
WT

Figure 4-1**.** Analysis of the DUF2237 gene. (a) Multiple alignment of DUF2237 amino-acid sequences of Flavobacteriia strains and *Synechocystis* sp. PCC6803. Amino acids with background colours indicate residues with >50% consensus. The multiple alignment was conducted by using MAFFT with the linsi algorithm and its default options and was visualized by using MView. (b) Biased distribution of DUF2237 genes to phototrophs. The bar chart represents the ratios of DUF2237-possessing strains in the PSII-possessing prokaryotes (n = 51), rhodopsin-possessing prokaryotes (n = 52), all prokaryotes (n = 547), PSII-possessing non-cyanobacterial prokaryotes (n = 20), rhodopsin-possessing non-cyanobacterial prokaryotes (n = 44), and all non-cyanobacterial prokaryotes (n = 515). The data were obtained from the MBGD database13. (c) Phototaxis assay of the DUF2237 gene knockout strain of *Synechocystis* sp. PCC6803-P. The bar graph shows the distances of colony movements from the spotted points under unidirectional light. Four replicated experiments were conducted for each strain. Statistical significance was examined by Student's *t*-test (p-value = 2.9E-0) (d) A photo of plate culture during phototaxis assays.

Chapter 5 General discussion

Importance of "Solar-panel and parasol" strategies to understand the ecology of marine Flavobacteriia

In this thesis, I sequenced 21 genomes of marine Flavobacteriia and conducted comparative genomics. A particularly important finding is a trade-off relationship between light utilization and light protection, indicating that there are two different approaches for adaptation to sunlit environments in marine Flavobacteria. Traditionally, the research of oceanic Flavobacteriia mainly focused on their high abundance in macroscopic organic aggregates (also known as marine snows) (Buchan et al. 2014, DeLong et al. 1993, Kirchman 2002). On the other hand, since the promotion of light-dependent growth as a bacterium with PR has been confirmed for the first time using Flavobacteriia (Gómez-Consarnau et al. 2007), many studies began to focus on aspects of Flavobacteriia as a model of light-utilizing bacteria. However, the interiors of macroscopic organic aggregates in the upper ocean is nutrient-rich conditions and likely decrease the advantage of possessing PR, thus these two characteristics of Flavobacteriia seems incompatible. Previous metagenomics study of free-living and particle-attached microbial assemblages within a coastal ecosystem reported that there are more PR+ Flavobacteriia in free-living communities than particle-attached communities (Smith et al. 2013). My results provide the reason of such different niches of Flavobacteriia from ecophysiological perspective. Both light conditions and oxygenic conditions should contribute the formation of flavobacterial communities.

The usefulness of phylogenetic profiling analysis on the field of microbial ecology

This research should be a good example of phylogenetic profiling analysis in the field of microbial ecology. For bacteria with β-carotene, only the retinal synthesis gene *blh* and the rhodopsin gene are necessary to obtain functional bacterial rhodopsin (Pinhassi et al. 2016, Sharma et al. 2006). Because of this structural simplicity, it is known that many HGT events of bacterial rhodopsin have occurred in the evolutional history of prokaryotes. Considering genetic mobility and function of PR, Sharma et al. claimed bacterial rhodopsin should be a good model to study HGT-driven bacterial evolution (Sharma et al. 2006). The polyphyletic distribution of PR genes caused by HGT events is necessary to conduct phylogenetic profiling analysis. However, there are many other genes showing polyphyletic distributions on the prokaryotic phylogeny (Martiny et al. 2013) and phylogenetic profiling analysis should be a powerful tool to study the bacterial evolution and adaptation.

From a methodological perspective, it is also important to conduct annotation-independent method and did a statistical evaluation on all genes including functional unknown genes. How to estimate the function of genes with no homology with functionary established genes is one of the largest problems in the $1000 genome era (DeLong 2009, Galperin and Koonin 2014). The example of experimental validation of light-related function of DUF2237 shows that functional prediction of unknown genes by phylogenetic profiling analysis is useful in handling an enormous number of functional unknown genes possessed by environmental prokaryotes.

Bacterial genome reduction and "Solar-panel and parasol" strategies

The relationship between light utilization and bacterial genome reduction clarified in this study provides insights in the field of bacterial evolution. SAR11 and SAR86 group bacteria, which are the most dominant bacterial groups in the ocean surface, have extremely small genome sizes and the reason of it being a mystery in the study of bacterial evolution (Dupont et al. 2012, Giovannoni et al. 2005b). One reasonable explanation is a hypothesis called "the black queen hypothesis" that some of the ocean surface bacteria are dependent on metabolites of other community members to reduce the number of necessary genes (Morris et al. 2012, Morris 2015). However, the black queen hypothesis is a theory to explain "how" bacteria shrink the genome size, and there was no argument as to "why" genome reduction is necessary. Both SAR11 and SAR86 group bacteria have similar characteristics with PR+ Flavobacteriia in that they both have PR and many photolyase genes (Dupont et al. 2012, Giovannoni et al. 2005b), and they should also have small genomes as compensation for light utilization. Unlike the case of Flavobacteriia, PR– strains of SAR11 and SAR86 have not found yet and their high dependency on light utilization possibly resulted in their extreme small genome size. To obtain further insight, the influence of existence of rhodopsin gene on genome size evolution should be evaluated by more mathematically accurate way. The probabilistic evaluation of the acquisition/deletion event of rhodopsin gene and the change in genome size enables quantitative assessment of the extent to which rhodopsin is involved in the change in genome size.

Concluding remarks and future perspectives

The "solar panel and parasol hypothesis" shows that the adaptation to light is important for understanding the ecology and evolution of marine bacteria. Although this scheme should be able to expand other prokaryotic group, marine Flavobacteriia is the best model group to understand my hypothesis because of their ecological abundance and culturability. To get further insights of my hypothesis, the establishment of genetic manipulation methods of marine Flavobacteriia and knockout-based analysis of PR and APE/FTP are required. Another important question provoked by my hypothesis is which adaptation strategies to light is more adaptive in what kind of environment. To answer this question, further investigation is required that the distribution of rhodopsin and outer membrane pigments synthesis gene in the particle-attaching/free-living fraction in the ocean surface layer.

Acknowledgements

Firstly, I would like to express my gratitude and deep appreciation to my supervisor Prof. Kazuhiro Kogure, for his giving me fruitful suggestions, material support and continuous guidance during my study. My sincere thanks also go to Prof. Wataru Iwasaki and Prof. Susumu Yoshizawa for their great support for entire my research. I also appreciate Prof. Chuya Shinzato and Prof. Shigeaki Kojima for their work as a sub-chief examiner of my doctoral thesis and their helpful advice for this thesis. I thank Dr. Mai Watanabe, Mr. Yu Nakajima, Dr. Shu-Kuan Wong, Dr. Minoru Ijichi, and Ms. Masumi Hasegawa for assisting in experiments. I thank Prof. Masahira Ikeuchi, Dr. Tsukasa Fukunaga, Dr. Haruka Ozaki, Prof. Koji Hamasaki, Prof. Rei Narikawa, Dr. Motomu Matsui, Mr. Satoshi Hiraoka, Dr. Hiroshi Kiyota, Dr. Daisuke Nakane, and Prof. Edward F. DeLong for providing helpful suggestions. I wish to express my gratitude to Prof. Yoshitoshi Ogura, Prof. Tetsuya Hayashi, Prof. Kenshiro Oshima, and Prof. Masahira Hattori for their help to sequencing bacterial genome. The members and staff of #bioinfowakate provides me good opportunity to discuss about my research with researchers in different fields of biology. Some computations were performed on the NIG supercomputer at ROIS National Institute of Genetics, so I'm grateful to the staff of the NIG supercomputer.

References

Bauer M, Kube M, Teeling H, Richter M, Lombardot T, Allers E *et al* (2006). Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii'reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol* **8:** 2201-2213.

Béja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902-1906.

Bhaya D (2004). Light matters: phototaxis and signal transduction in unicellular cyanobacteria. *Mol Microbiol* **53:** 745-754.

Brash DE, Haseltine WA (1982). UV-induced mutation hotspots occur at DNA damage hotspots. *Nature* **298:** 189-192.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A *et al* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523:** 208-211.

Brunner E, Munzel U (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical J* **42:** 17-25.

Bruns A, Rohde M, Berthe-Corti L (2001). Muricauda ruestringensis gen. nov., sp. nov., a facultatively anaerobic, appendaged bacterium from German North Sea intertidal sediment. *Int J Syst Evol Microbiol* **51:** 1997-2006.

Buchan A, LeCleir GR, Gulvik CA, González JM (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* **12:** 686-698.

Campbell BJ, Waidner LA, Cottrell MT, Kirchman DL (2008). Abundant proteorhodopsin genes in the North Atlantic Ocean. *Environ Microbiol* **10:** 99-109.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972-1973.

Chan Y-K, McCormic WA, Watson RJ (1997). A new nos gene downstream from nosDFY is essential for dissimilatory reduction of nitrous oxide by Rhizobium (Sinorhizobium) meliloti. *Microbiology* **143:** 2817-2824.

Cimermancic P, Medema MH, Claesen J, Kurita K, Brown LCW, Mavrommatis K *et al* (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158:** 412-421.

The Uniprot Consortium (2014). UniProt: a hub for protein information. *Nucleic Acids Res* **43:** D204-D212.

Cordero OX, Ventouras L-A, DeLong EF, Polz MF (2012). Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc Natl Acad Sci USA* **109:** 20059-20064.

Coyle CL, Zumft WG, Kroneck PM, Korner H, Jakob W (1985). Nitrous oxide reductase from denitrifying Pseudomonas perfectomarina. Purification and properties of a novel multicopper enzyme. . *FEBS J* **153:** 459-467.

Darriba D, Taboada GL, Doallo R, Posada D (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27:** 1164-1165.

DeLong EF, Franks DG, Alldredge AL (1993). Phylogenetic diversity of aggregate ‐ attached vs. free ‐ living marine bacterial assemblages. *Limnol Oceanogr* **38:** 924-934.

DeLong EF (2005). Microbial community genomics in the ocean. *Nat Rev Microbiol* **3:** 459-469.

DeLong EF (2009). The microbial ocean from genomes to biomes. *Nature* **459:** 200-206.

DeLong EF, Beja O (2010). The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. *PLoS Biol* **8:** e1000359.

Drachev L, Jasaitis A, Kaulen A, Kondrashin A, Liberman E, Nemecek I *et al* (1974). Direct measurement of electric current generation by cytochrome oxidase, H+-ATPase and bacteriorhodopsin. *Nature* **249:** 321-324.

Dubinsky V, Haber M, Burgsdorf I, Saurav K, Lehahn Y, Malik A *et al* (2017). Metagenomic analysis reveals unusually high incidence of proteorhodopsin genes in the ultraoligotrophic Eastern Mediterranean Sea. *Environ Microbiol* **19:** 1077-1090.

Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Richter RA, Valas R *et al* (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* **6:** 1186-1199.

Fautz E, Reichenbach H (1980). A simple test for flexirubin-type pigments. *FEMS Microbiol Lett* **8:** 87-91.

Fernández-Gomez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM *et al* (2013). Ecology of marine Bacteroidetes: a comparative genomics approach. *ISME J* **7:** 1026.

Finkel OM, Béjà O, Belkin S (2013). Global abundance of microbial rhodopsins. *ISME J* **7:** 448-451.

Galperin MY, & Koonin EV (2014). Comparative genomics approaches to identifying functionally related genes. *International Conference on Algorithms for Computational Biology*. 1-24

Ganesh S, Parris DJ, DeLong EF, Stewart FJ (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J* **8:** 187.

Gao Q, Garcia-Pichel F (2011). Microbial ultraviolet sunscreens. *Nature reviews Microbiology* **9:** 791.

Giovannoni SJ, Bibbs L, Cho J-C, Stapels MD, Desiderio R, Vergin KL *et al* (2005a). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438:** 82-85.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al* (2005b). Genome streamlining in a cosmopolitan oceanic bacterium. *science* **309:** 1242-1245.

Goel AK, Rajagopal L, Nagesh N, Sonti RV (2002). Genetic locus encoding functions involved in biosynthesis and outer membrane localization of xanthomonadin in Xanthomonas oryzae pv. oryzae. *J Bacteriol* **184:** 3539-3548.

Gómez-Consarnau L, González JM, Coll-Lladó M, Gourdon P, Pascher T, Neutze R *et al* (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445:** 210.

Gómez-Consarnau L, Akram N, Lindell K, Pedersen A, Neutze R, Milton DL *et al* (2010). Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biol* **8:** e1000358.

González JM, Fernández-Gómez B, Fernàndez-Guerra A, Gómez-Consarnau L, Sánchez O, Coll-Lladó M *et al* (2008). Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *Proc Natl Acad Sci USA* **105:** 8724-8729.

González JM, Pinhassi J, Fernández-Gómez B, Coll-Lladó M, González-Velázquez M, Puigbò P *et al* (2011). Genomics of the proteorhodopsin-containing marine flavobacterium Dokdonia sp. strain MED134. *Appl Environ Microbiol* **77:** 8676-8686.

Haft DH (2015). Using comparative genomics to drive new discoveries in microbiology. *Curr Opin Microbiol* **23:** 189-196.

Hiraoka S, Yang C-c, Iwasaki W (2016). Metagenomics and bioinformatics in microbial ecology: current status and beyond. *Microbes Environ* **31:** 204-212.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC *et al* (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44:** D286-D293.

Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C *et al* (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol***:** msx148.

Hugenholtz P, Tyson GW (2008). Microbiology: metagenomics. *Nature* **455:** 481-483.

Huynen MA, Snel B, von Mering C, Bork P (2003). Function prediction and protein networks. *Curr Opin Cell Biol* **15:** 191-198.

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.

Imlay JA (2013). The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. *Nature reviews Microbiology* **11:** 443.

Irschik H, Reichenbach H (1978). Intracellular location of flexirubins in Flexibacter elegans (Cytophagales). *Biochimica et Biophysica Acta (BBA)-Biomembranes* **510:** 1-10.

Johnson ET, Baron DB, Naranjo B, Bond DR, Schmidt-Dannert C, Gralnick JA (2010). Enhancement of survival and electricity production in an engineered bacterium by light-driven proton pumping. *Appl Environ Microbiol* **76:** 4123-4129.

Kanehisa M, Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Kao RR, Haydon DT, Lycett SJ, Murcia PR (2014). Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol* **22:** 282-291.

Katoh K, Misawa K, Kuma Ki, Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059-3066.

Kim Y-S, Kim N-H, Yeom S-J, Kim S-W, Oh D-K (2009). In vitro characterization of a recombinant Blh protein from an uncultured marine bacterium as a β-carotene 15, 15′-dioxygenase. *J Biol Chem* **284:** 15781-15793.

Kirchman DL (2002). The ecology of Cytophaga–Flavobacteria in aquatic environments. *FEMS Microbiol Ecol* **39:** 91-100.

Kumagai Y, Yoshizawa S, Oshima K, Hattori M, Iwasaki W, Kogure K (2014). Complete genome sequence of Winogradskyella sp. strain PG-2, a proteorhodopsin-containing marine flavobacterium. *Genome announcements* **2:** e00490-00414.

Kwon S-K, Kim BK, Song JY, Kwak M-J, Lee CH, Yoon J-H *et al* (2013). Genomic makeup of the marine flavobacterium Nonlabens (Donghaeana) dokdonensis and identification of a novel class of rhodopsins. *Genome biology and evolution* **5:** 187-199.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35:** 3100-3108.

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H *et al* (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics* **15:** 141-161.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S *et al* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106:** 15527-15533.

Letunic I, Bork P (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44:** W242-W245.

Lowe TM, Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955-964.

Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC *et al* (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature communications* **6**.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY *et al* (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43:** D222-D226.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M *et al* (2013). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* **42:** D560-D567.

Martiny AC, Treseder K, Pusch G (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J* **7:** 830-838.

Mishra S, Imlay JA (2013). An anaerobic bacterium, Bacteroides thetaiotaomicron, uses a consortium of enzymes to scavenge hydrogen peroxide. *Mol Microbiol* **90:** 1356-1371.

Mongodin EF, Nelson K, Daugherty S, Deboy R, Wister J, Khouri H *et al* (2005). The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 18147-18152.

Morris JJ, Lenski RE, Zinser ER (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3:** e00036-00012.

Morris JJ (2015). Black Queen evolution: the role of leakiness in structuring microbial communities. *Trends Genet* **31:** 475-482.

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4:** 673.

Nakamura Y, Kaneko T, Hirosawa M, Miyajima N, Tabata S (1998). CyanoBase, a www database containing the complete nucleotide sequence of the genome of Synechocystis sp. strain PCC6803. *Nucleic Acids Res* **26:** 63-67.

Nedashkovskaya OI, Kim SB, Lysenko AM, Frolova GM, Mikhailov VV, Lee KH *et al* (2005). Description of Aquimarina muelleri gen. nov., sp. nov., and proposal of the reclassification of [Cytophaga] latercula Lewin 1969 as Stanierella latercula gen. nov., comb. nov. *Int J Syst Evol Microbiol* **55:** 225-229.

Neumann B, Pospiech A, Schairer HU (1992). Rapid isolation of genomic DNA from gram-negative bacteria. *Trends Genet* **8:** 332-333.

Nordlund P, Reichard P (2006). Ribonucleotide reductases. *Annu Rev Biochem* **75:** 681-706.

O'brien KP, Remm M, Sonnhammer EL (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33:** D476-D480.

Oesterhelt D, Stoeckenius W (1971). Rhodopsin-like protein from the purple membrane of Halobacterium halobium. *Nature* **233:** 149-152.

Pagel P, Wong P, Frishman D (2004). A domain interaction map based on phylogenetic profiling. *J Mol Biol* **344:** 1331-1346.

Palovaara J, Akram N, Baltar F, Bunse C, Forsberg J, Pedrós-Alió C *et al* (2014). Stimulation of growth by proteorhodopsin phototrophy involves regulation of central metabolic pathways in marine planktonic bacteria. *Proc Natl Acad Sci USA* **111:** E3650-E3658.

Paulmier A, Ruiz-Pino D (2009). Oxygen minimum zones (OMZs) in the modern ocean. *Prog Oceanogr* **80:** 113-128.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96:** 4285-4288.

Pinhassi J, DeLong EF, Béjà O, González JM, Pedrós-Alió C (2016). Marine bacterial and archaeal ion-pumping rhodopsins: genetic diversity, physiology, and ecology. *Microbiol Mol Biol Rev* **80:** 929-954.

Ploug H, Kühl M, Buchholz-Cleven B, Jørgensen BB (1997). Anoxic aggregates-an ephemeral phenomenon in the pelagic environment? *Aquat Microb Ecol* **13:** 285-294.

Preisig O, Anthamatten D, Hennecke H (1993). Genes for a microaerobically induced oxidase complex in Bradyrhizobium japonicum are essential for a nitrogen-fixing endosymbiosis. *Proc Natl Acad Sci USA* **90:** 3309-3313.

Pruitt KD, Tatusova T, Maglott DR (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501-D504.

Reichenbach H, Kohl W, Böttger-Vetter A, Achenbach H (1980). Flexirubin-type pigments in Flavobacterium. *Arch Microbiol* **126:** 291-293.

Riedel T, Held B, Nolan M, Lucas S, Lapidus A, Tice H *et al* (2012). Genome sequence of the Antarctic rhodopsins-containing flavobacterium Gillisia limnaea type strain (R-8282 T). *Standards in Genomic Sciences* **7:** 107.

Riedel T, Gómez-Consarnau L, Tomasch J, Martin M, Jarek M, González JM *et al* (2013). Genomics and physiology of a marine flavobacterium encoding a proteorhodopsin and a xanthorhodopsin-like protein. *PLoS One* **8:** e57487.

Rodionov DA, Gelfand MS (2005). Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. *Trends Genet* **21:** 385-389.

Sancar A (1994). Structure and function of DNA photolyase. *Biochemistry* **33:** 2-9.

Schellhorn HE, Hassan HM (1988). Transcriptional regulation of katE in Escherichia coli K-12. *J Bacteriol* **170:** 4286-4292.

Segata N, Börnigen D, Morgan XC, Huttenhower C (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications* **4:** 2304.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-2504.

Sharma AK, Spudich JL, Doolittle WF (2006). Microbial rhodopsins: functional versatility and genetic mobility. *Trends Microbiol* **14:** 463-469.

Sharma AK, Sommerfeld K, Bullerjahn GS, Matteson AR, Wilhelm SW, Jezbera J *et al* (2009). Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria. *ISME J* **3:** 726-737.

Sheridan C (2014). Illumina claims [dollar] 1,000 genome win. Nature Research.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31:** 3210-3212.

Smith MW, Allen LZ, Allen AE, Herfort L, Simon HM (2013). Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Frontiers in microbiology* **4**.

Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C (2006). Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics* **7:** 420.

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

Stanier R, Kunisawa R, Mandel M, Cohen-Bazire G (1971). Purification and properties of unicellular blue-green algae (order Chroococcales). *Bacteriological reviews* **35:** 171.

Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100:** 9440-9445.

Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T *et al* (2005). Refined phylogenetic profiles method for predicting protein–protein interactions. *Bioinformatics* **21:** 3409-3415.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al* (2015). Structure and function of the global ocean microbiome. *Science* **348:** 1261359.

Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF *et al* (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J* **8:** 1440-1451.

Tseng C-H, Tang S-L (2014). Marine microbial metagenomics: from individual to the environment. *International journal of molecular sciences* **15:** 8878-8892.

Uchiyama I (2003). MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res* **31:** 58-62.

Uchiyama I (2006). Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic Acids Res* **34:** 647-658.

Venil CK, Zakaria ZA, Usha R, Ahmad WA (2014). Isolation and characterization of flexirubin type pigment from Chryseobacterium sp. UTM-3 T. *Biocatal Agric Biotechnol* **3:** 103-107.

Wang Y, Qian G, Li Y, Wang Y, Wang Y, Wright S *et al* (2013). Biosynthetic mechanism for sunscreens of the biocontrol agent Lysobacter enzymogenes. *PLoS One* **8:** e66633.

Williams TJ, Wilkins D, Long E, Evans F, DeMaere MZ, Raftery MJ *et al* (2013). The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environ Microbiol* **15:** 1302-1317.

Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* **74:** 5088-5090.

Yamada T, Waller AS, Raes J, Zelezniak A, Perchat N, Perret A *et al* (2012). Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Mol Syst Biol* **8:** 581.

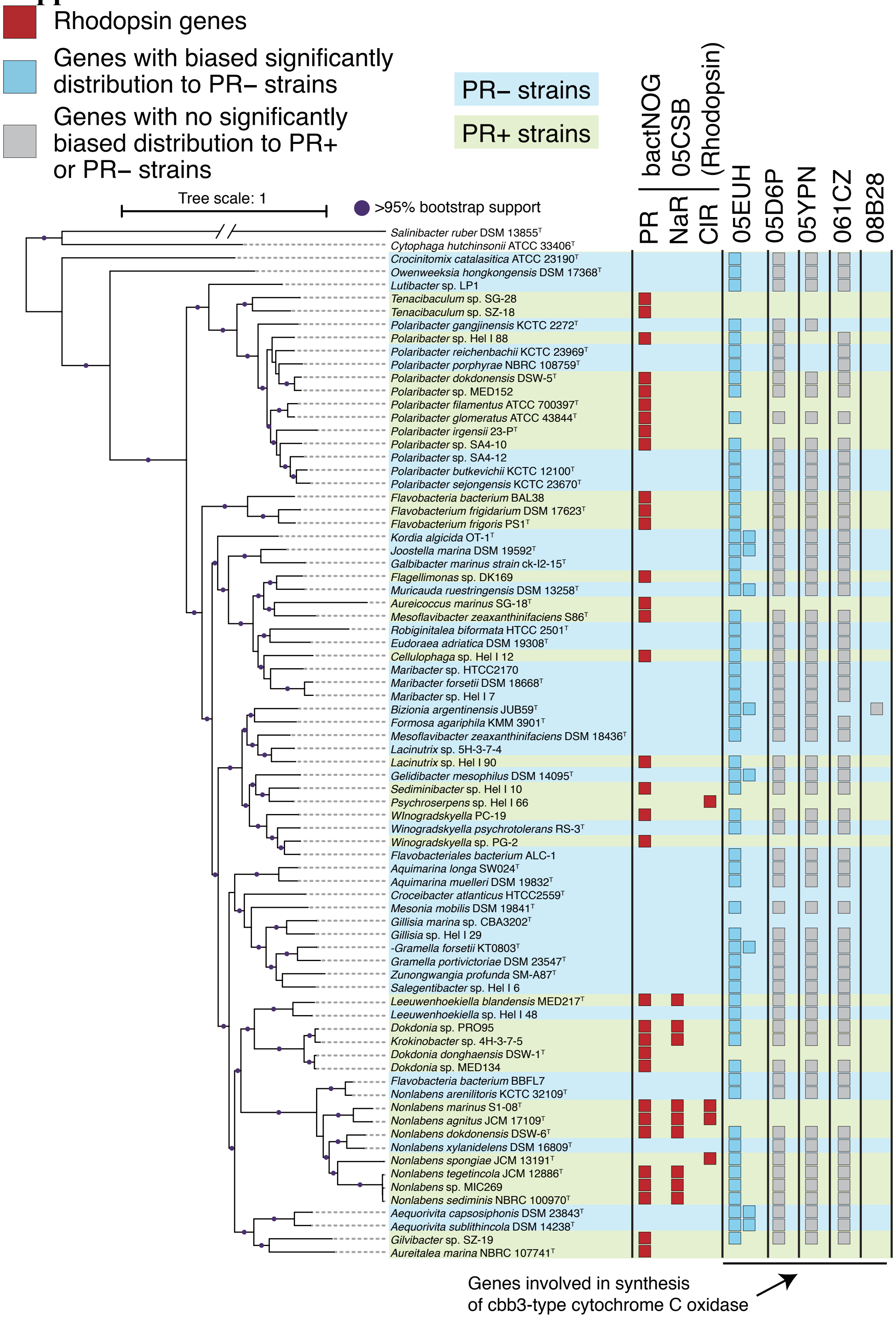
Yoshihara S, Suzuki F, Fujita H, Geng XX, Ikeuchi M (2000). Novel putative photoreceptor and regulatory genes required for the positive phototactic movement of the unicellular motile cyanobacterium Synechocystis sp. PCC 6803. *Plant Cell Physiol* **41:** 1299-1304.

Yoshihara S, Ikeuchi M (2004). Phototactic motility in the unicellular cyanobacterium Synechocystis sp. PCC 6803. *Photoch Photobio Sci* **3:** 512-518.

Yoshizawa S, Kawanabe A, Ito H, Kandori H, Kogure K (2012). Diversity and functional analysis of proteorhodopsin in marine Flavobacteria. *Environ Microbiol* **14:** 1240-1248.

Yoshizawa S, Kumagai Y, Kim H, Ogura Y, Hayashi T, Iwasaki W *et al* (2014). Functional characterization of flavobacteria rhodopsins reveals a unique class of light-driven chloride pump in bacteria. *Proc Natl Acad Sci USA* **111:** 6732-6737.

Supplemental Information



Rhodopsin genes
Genes with biased significantly distribution to PR– strains
Genes with no significantly biased distribution to PR+ or PR– strains
PR– strains
PR+ strains
Tree scale: 1
>95% bootstrap support
PR | bactNOG
NaR | 05CSB
ClR | (Rhodopsin)
05EUH
05D6P
05YPN
061CZ
08B28
Salinibacter ruber DSM 13855T
Cytophaga hutchinsonii ATCC 33406T
Crocinitomix catalasitica ATCC 23190T
Owenweeksia hongkongensis DSM 17368T
Lutibacter sp. LP1
Tenacibaculum sp. SG-28
Tenacibaculum sp. SZ-18
Polaribacter gangjinensis KCTC 2272T
Polaribacter sp. Hel I 88
Polaribacter reichenbachii KCTC 23969T
Polaribacter porphyrae NBRC 108759T
Polaribacter dokdonensis DSW-5T
Polaribacter sp. MED152
Polaribacter filamentus ATCC 700397T
Polaribacter glomeratus ATCC 43844T
Polaribacter irgensii 23-PT
Polaribacter sp. SA4-10
Polaribacter sp. SA4-12
Polaribacter butkevichii KCTC 12100T
Polaribacter sejongensis KCTC 23670T
Flavobacteria bacterium BAL38
Flavobacterium frigidarium DSM 17623T
Flavobacterium frigoris PS1T
Kordia algicida OT-1T
Joostella marina DSM 19592T
Galbibacter marinus strain ck-I2-15T
Flagellimonas sp. DK169
Muricauda ruestringensis DSM 13258T
Aureicoccus marinus SG-18T
Mesoflavibacter zeaxanthinifaciens S86T
Robiginitalea biformata HTCC 2501T
Eudoraea adriatica DSM 19308T
Cellulophaga sp. Hel I 12
Maribacter sp. HTCC2170
Maribacter forsetii DSM 18668T
Maribacter sp. Hel I 7
Bizionia argentinensis JUB59T
Formosa agariphila KMM 3901T
Mesoflavibacter zeaxanthinifaciens DSM 18436T
Lacinutrix sp. 5H-3-7-4
Lacinutrix sp. Hel I 90
Gelidibacter mesophilus DSM 14095T
Sediminibacter sp. Hel I 10
Psychroserpens sp. Hel I 66
WInogradskyella PC-19
Winogradskyella psychrotolerans RS-3T
Winogradskyella sp. PG-2
Flavobacteriales bacterium ALC-1
Aquimarina longa SW024T
Aquimarina muelleri DSM 19832T
Croceibacter atlanticus HTCC2559T
Mesonia mobilis DSM 19841T
Gillisia marina sp. CBA3202T
Gillisia sp. Hel I 29
-Gramella forsetii KT0803T
Gramella portivictoriae DSM 23547T
Zunongwangia profunda SM-A87T
Salegentibacter sp. Hel I 6
Leeuwenhoekiella blandensis MED217T
Leeuwenhoekiella sp. Hel I 48
Dokdonia sp. PRO95
Krokinobacter sp. 4H-3-7-5
Dokdonia donghaensis DSW-1T
Dokdonia sp. MED134
Flavobacteria bacterium BBFL7
Nonlabens arenilitoris KCTC 32109T
Nonlabens marinus S1-08T
Nonlabens agnitus JCM 17109T
Nonlabens dokdonensis DSW-6T
Nonlabens xylanidelens DSM 16809T
Nonlabens spongiae JCM 13191T
Nonlabens tegetincola JCM 12886T
Nonlabens sp. MIC269
Nonlabens sediminis NBRC 100970T
Aequorivita capsosiphonis DSM 23843T
Aequorivita sublithincola DSM 14238T
Gilvibacter sp. SZ-19
Aureitalea marina NBRC 107741T
Genes involved in synthesis of cbb3-type cytochrome C oxidase

Figure 3-S1. Distributions of *cbb*$_3$-type cytochrome oxidase genes.

The genomic phylogenetic tree is from Fig. 1. The closed circles indicate branches with 95% bootstrapping support. Blue and green horizontal dotted lines indicate PR－ and PR+ strains, respectively. The number of genes encoded by each genome is illustrated by the number of closed squares. Red: rhodopsin genes. Light blue and grey: bactNOG orthologue groups of *cbb*$_3$-type cytochrome oxidase genes (05EUH: *cbb*$_3$-type cytochrome *c* oxidase subunit I, 05D6P: a *cbb*$_3$-type cytochrome *c* oxidase complex protein, 05YPN: *cbb*$_3$-type cytochrome oxidase maturation protein, 061CZ: *cbb*$_3$-type cytochrome oxidase component FixQ, and 08B28: *cbb*$_3$-type cytochrome oxidase component FixQ). The 05EUH group is coloured in light blue because it showed distributions that were significantly biased to the PR– genomes.

Table 2-S1. List of 76 marine Flavobacteriia genomes. The isolation sites of the publicly available genome sequences were acquired from the IMG database[1].

NCBI Accession No.	Species	Strain	Type strain	PR	BUSCO score	Genome size	Isolation site
NZ_AJUG00000000	*Joostella marina*	DSM 19592	T	-	99.10%	4508243	Coastal seawater in the East Sea of Korea, at a depth of 100 m
NZ_JHZZ00000000.1	*Polaribacter sp.*	Hel_I_88	-	+	98.20%	3996527	Seawater
NC_015638.1	*Lacinutrix sp.*	5H-3-7-4	-	-	98.90%	3296168	Subseafloor sediments at Suruga Bay (Japan)from a depth of 41 m
NZ_ABHI00000000.1	*Flavobacteriales bacterium*	ALC-1	-	-	98.50%	3825707	Scripps Pier La Jolla CA
NC_018013.1	*Aequorivita sublithincola*	DSM 14238	T	-	99.10%	3520671	Sea water from Vestfold Hills Antarctica
NZ_AULQ00000000.1	*Mesoflavibacter zeaxanthinifaciens*	DSM 18436	T	-	98.60%	2965434	Shallow seawater sample
NZ_AMSG00000000.1	*Galbibacter sp.*	ck-I2-15	-	-	98.40%	3572447	Deep sea sediment
NZ_JHXV00000000.1	*Crocinitomix catalasitica*	ATCC 23190	T	-	96.80%	4619089	Under frozen sand, Auke Bay, AK
NZ_CP011373.1	*Nonlabens sp.*	MIC269	-	+	98.70%	2884293	Koror Island, Palau
NZ_CP009301.1	*Dokdonia donghaensis*	MED134	-	+	98.60%	3302548	Northwestern Mediterranean Sea surface water (0.5 m depth), collected one km off the coast of Catalonia at the Blanes Bay Microbial Observatory
NC_015945.1	*Muricauda ruestringensis*	DSM 13258	T	-	99.80%	3842422	Germany, North Sea coast
NZ_AANC00000000.1	*Leeuwenhoekiella blandensis*	MED217	T	-	97.50%	4238065	Spain,Bay of Blanes, NW Mediterranean Sea at a depth of 1m
NC_020156.1	*Donghaeana dokdonensis*	DSW-6	T	+	99.10%	3914632	Sea water sampled in Takeshima Island

NZ_ABIB00000000.1	*Kordia algicida*	OT-1	T	-	98.60%	5019836	1m depth at Masan Bay in South Sea Korea
NZ_ARNE00000000.1	*Eudoraea adriatica*	DSM 19308	T	-	99.70%	3906474	Coastal waters of the Adriatic Sea
CP013355.1	*Lutibacter* sp.	LP1	-	-	98.40%	2966978	Microbial mat situated on a chimney wall at the Lokis Castle deep-sea hydrothermal vent site
NZ_JHZY00000000.1	*Leeuwenhoekiella* sp.	Hel_I_48	-	+	97.90%	4281274	Seawater
NZ_ATMR00000000.1	*Winogradskyella psychrotolerans*	RS-3	T	-	90.70%	4337031	Marine
NZ_JHZW00000000.1	*Maribacter* sp.	Hel_I_7	-	-	99.50%	4775040	Seawater
NZ_LCTZ00000000.1	*Flagellimonas* sp.	DK169	-	+	98.60%	4132279	Takeshima Island
NC_014472.1	*Maribacter* sp.	HTCC2170	-	-	99.30%	3868304	Coastal area of Newport Oregon at a depth of 10m
NZ_AUHD01000000	*Gelidibacter mesophilus*	DSM 14095	T	-	99.10%	4430503	Mediterranean sea water
NZ_AAOG00000000.1	*Polaribacter irgensii*	23-P	T	+	97.30%	2745458	Surface water from the Penola Strait Antarctica
NC_013222.1	*Robiginitalea biformata*	HTCC2501	T	-	98.60%	3530383	Seawater taken at a depth of 10m from the Sargasso Sea
NZ_AUML00000000.1	*Aquimarina muelleri*	DSM 19832	T	-	99.40%	4900431	Seawater sample
NZ_AFOE00000000.1	*Mesoflavibacter zeaxanthinifaciens*	S86	-	+	98.40%	3704661	Seawater of Chuuk State in Micronesia
NC_015496.1	*Krokinobacter diaphorus*	4H-3-7-5	-	+	99.10%	3389993	Subseafloor sediments at Suruga Bay (Japan) from a depth of 31.4 meters
NC_014041.1	*Zunongwangia profunda*	SM-A87	T	-	98.60%	5128187	Deep-sea sediment; China, East China Sea, southern Okinawa Trough area
NZ_JQLH00000000.1	*Maribacter forsetii*	DSM 18668	T	-	98.90%	4514366	Seawater

NC_014230.1	*Croceibacter atlanticus*	HTCC2559	T	-	99.10%	2952962	Sargasso Sea at a depth of 250 meters
NC_016599.1	*Owenweeksia hongkongensis*	DSM 17368	T	-	97.30%	4000057	Sea water (sand filtered); China, Hong Kong
NZ_AVQK00000000.1	*Aquimarina longa*	SW024	T	-	99.80%	5501201	Surface seawater
NZ_AFXZ00000000.1	*Bizionia argentinensis*	JUB59	T	-	98.80%	3279329	Surface seawater in Antarctica.
NZ_AUBG00000000.1	*Aequorivita capsosiphonis*	DSM 23843	T	-	98.60%	4039217	Seawater
NZ_AUDO00000000.1	*Flavobacterium frigidarium*	DSM 17623	T	+	98.60%	3627910	Marine sediment
NZ_JQLP00000000.1	*Gillisia* sp.	Hel_I_29	-	-	99.70%	3959304	Seawater
NZ_JHZX01000001.1	*Sediminibacter* sp.	Hel_I_10	-	+	99.10%	4106053	Seawater
NZ_AHKF00000000.1	*Flavobacterium frigoris*	PS1	-	+	99.10%	3934101	Marine
NZ_AAPD00000000.1	*Flavobacteria bacterium*	BBFL7	-	-	98.00%	3083153	SIO pier water
NC_008571.1	*Gramella forsetii*	KT0803	T	-	99.30%	3798465	Concentrated seawater collected from the German Bight in the North Sea
NZ_AJLT00000000.1	*Gillisia* sp.	CBA3202	-	-	86.70%	2981404	Marine
NZ_AAXX00000000.1	*Flavobacteria bacterium*	BAL38	-	+	97.10%	2806989	4m depth of Baltic proper
NZ_AUHF00000000.1	*Gramella portivictoriae*	DSM 23547	T	-	99.40%	3264369	Marine sediment
NC_020830.1	*Polaribacter* sp.	MED152	-	+	98.60%	2961474	NW Mediterranean Sea from a sample taken at a depth of 1m
NZ_AUHX00000000.1	*Mesonia mobilis*	DSM 19841	T	-	99.30%	3200157	Seawater
NZ_HG315671.1	*Formosa agariphila*	KMM 3901	T	-	99.60%	4228350	Marine sample
NZ_LGBR00000000.1	*Polaribacter dokdonensis*	DSW-5	T	+	99.10%	3087076	Seawater
NZ_JSAQ00000000.1	*Dokdonia donghaensis*	DSW-1	T	+	98.60%	3219590	Seawater

NZ_JUHB00000000.1	*Cellulophaga* sp.	Hel_I_12	-	+	99.50%	4031126	Seawater
NZ_JUGU01000001.1	*Psychroserpens* sp.	Hel_I_66	-	only ClR	97.70%	3842990	Seawater
NZ_JQNQ00000000.1	*Salegentibacter* sp.	Hel_I_6	-	-	99.60%	4212161	Seawater
NZ_JYNQ00000000.1	*Lacinutrix* sp.	Hel_I _90	-	+	98.90%	3819763	Seawater
NZ_ANPJ00000000.1	*Dokdonia* sp.	PRO95	-	+	98.20%	3303993	Seawater
NZ_AP014548.1	*Nonlabens marinus*	S1-08	T	+	98.40%	2915920	Seawater
NZ_AP014583.1	*Winogradskyella* sp.	PG-2	-	+	99.10%	3811479	Seawater
Newly sequenced	*Tenacibaculum* sp.	SG-28	-	+	85.60%	2801347	Surface seawater at Western North Pacific Station S (30°40'N, 138°00'E)
Newly sequenced	*Aureicoccus marinus*	SG-18	T	+	88.90%	3052917	Surface seawater at Western North Pacific Station S (30°40'N, 138°00'E)
Newly sequenced	*Polaribacter* sp.	SA4-10	-	+	97.70%	3435762	Sea ice in Saroma-ko Lagoon (44°07'N, 143°58'E)
Newly sequenced	*Winogradskyella* sp.	PC-19	-	+	99.10%	2977423	Surface seawater at Sagami Bay Station P (35°00'N, 139°20'E)
Newly sequenced	*Tenacibaculum* sp.	SZ-18	-	+	97.70%	4024179	Surface seawater at Western North Pacific Station S (30°40'N, 138°00'E)
Newly sequenced	*Gilvibacter* sp.	SZ-19	-	+	97.10%	3100111	Surface seawater at Western North Pacific Station S (30°40'N, 138°00'E)
Newly sequenced	*Polaribacter* sp.	SA4-12	-	-	98.20%	3990161	Sea ice in Saroma-ko Lagoon (44°07'N, 143°58'E)

Newly sequenced	*Polaribacter butkevichii*	KCTC 12100	T	-	98.60%	4085573	Seawater
Newly sequenced	*Polaribacter gangjinensis*	KCTC 22729	T	-	98.40%	2943061	Seawater
Newly sequenced	*Polaribacter glomeratus*	ATCC 43844	T	+	98.90%	4064562	Seawater
Newly sequenced	*Polaribacter sejongensis*	KCTC 23670	T	-	97.70%	4526271	Seawater
Newly sequenced	*Polaribacter reichenbachii*	KCTC 23969	T	-	98.40%	4122594	Seawater
Newly sequenced	*Polaribacter porphyrae*	NBRC 108759	T	-	98.20%	3904103	Seawater
Newly sequenced	*Nonlabens agnitus*	JCM 17109	T	+	99.10%	3178896	Seawater
Newly sequenced	*Nonlabens arenilitoris*	KCTC 32109	T	-	97.70%	3323003	Seawater
Newly sequenced	*Nonlabens sediminis*	NBRC 100970	T	+	97.70%	2835711	Seawater
Newly sequenced	*Nonlabens spongiae*	JCM 13191	T	only ClR	98.20%	3393235	Seawater
Newly sequenced	*Nonlabens tegetincola*	JCM 12886	T	+	98.20%	3028293	Seawater
Newly sequenced	*Nonlabens xylanidelens*	DSM 16809	T	-	98.70%	3552991	Seawater
Newly sequenced	*Aureitalea marina*	NBRC 107741	T	+	94.00%	3074655	Surface seawater at Western North Pacific Station S1 (30°11'N, 145°05'E
Newly sequenced	*Polaribacter filamentus*	ATCC 700397	T	+	95.20%	4281931	Seawater
NC_008255.1	*Cytophaga hutchinsonii*	ATCC 33406	T	-	96.60%	4433218	Seawater
NC_007677.1	*Salinibacter ruber*	DSM 13855	T	XR, SR1, SR2	76.10%	3551823	Saltern crystallizer ponds in Spain

Table 3-S1. List of eggNOG orthologue groups with distributions biased to PR− Flavobacteriia.

Annotation	bactNOG ID	q-value
Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides (By similarity)	05BZH	9.90E-05
Two component, sigma54 specific, transcriptional regulator, Fis family	05C1W	6.70E-04
Glycosyl transferase, family 2	05EYS	3.00E-03
Fad-binding protein	07T3I	3.00E-03
Redoxin domain protein	080EH	3.10E-03
Phenylacetate-CoA ligase	05DZS	3.10E-03
Acyl-coenzyme A 6-aminopenicillanic acid acyl-transferase	07T6G	3.10E-03
Dehydratase	05M49	3.10E-03
Outer membrane lipoprotein carrier protein LolA	05IAS	3.10E-03
Synthase	05YNY	3.10E-03
Synthase	05F6T	3.10E-03
Acyl carrier protein	05VME	3.10E-03
NA	08XF1	3.10E-03
Synthase	05EWP	3.10E-03
Flexirubin-type pigment biosynthesis acyl carrier protein	0636E	3.10E-03
Lipid A biosynthesis acyltransferase	08M6I	3.10E-03
NA	05SQF	3.10E-03
Di-iron-containing protein involved in the repair of iron-sulfur clusters damaged by oxidative and nitrosative stress conditions (By similarity)	05FJH	3.90E-03
Catalase	05CH6	4.90E-03
Thioesterase	05H6H	5.80E-03
NA	05YQ6	6.60E-03
Two component transcriptional regulator (Winged helix family	05XJC	6.60E-03
Transcriptional regulator, BadM Rrf2 family	05GZ5	7.80E-03
Receptor	07T30	1.20E-02
Luciferase family	05DEN	1.20E-02
Short-chain dehydrogenase reductase Sdr	05FKB	1.20E-02
NA	08KH3	1.20E-02

Purine nucleoside phosphorylase DeoD-type	05D3A	1.50E-02
Transcriptional regulator	08X05	1.50E-02
Radical SAM domain protein	05DCH	1.60E-02
Cytochrome C oxidase, cbb3-type, subunit i	05EUH	1.60E-02
Secondary thiamine-phosphate synthase enzyme	08YXI	1.60E-02
Catalyzes the condensation of iminoaspartate with dihydroxyacetone phosphate to form quinolinate (By similarity)	05D0I	1.60E-02
L-aspartate oxidase	08IYU	1.60E-02
Protein of unknown function (DUF2874)	05YI6	1.60E-02
Methyltransferase	08VFE	1.60E-02
Extracellular solute-binding protein family 3	08MAW	1.60E-02
Dehydrogenase	05CPQ	1.60E-02
DsrE/DsrF-like family	05IRV	1.80E-02
Participates in control of cell volume in low-osmolarity conditions (By similarity)	05C3N	1.90E-02
Chloride channel	05CMQ	1.90E-02
NA	05IB6	2.00E-02
Histidine kinase	08JSN	2.10E-02
(Ubiquinol oxidase) subunit I	05C4M	2.40E-02
Uracil-dna glycosylase	08RG0	2.40E-02
NA	0724W	2.50E-02
4Fe-4S ferredoxin, iron-sulfur binding	05Y0W	2.50E-02
NA	05JPM	2.50E-02
NA	05YNB	2.50E-02
NA	05USS	2.50E-02
Glycosyl transferase (Group 1)	05CGN	3.10E-02
Transcriptional regulator, arac family	05FX6	3.10E-02
Gluconolactonase (EC 3.1.1.17)	05E7N	3.10E-02
NA	06CT7	3.30E-02
NA	05QIK	3.40E-02
Nicotinate phosphoribosyltransferase	07QI7	3.40E-02
Reductase	05CHR	3.70E-02
Nitrous-oxide reductase is part of a bacterial respiratory system which is activated under anaerobic conditions in the presence of nitrate or nitrous oxide (By similarity)	05EQJ	3.70E-02

NA	08JKI	3.70E-02
Periplasmic copper-binding protein	05DH4	3.70E-02
Nitrous-oxide metabolic protein nosy	05J39	3.70E-02
Cytochrome C, class I	05TBX	3.70E-02
Histidine ammonia-lyase	05C84	3.70E-02
Catalyzes the condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP (By similarity)	05C0Q	3.70E-02
Transcription factor jumonji	05ECT	3.70E-02
Helix-turn-helix domain protein	07ZHM	3.70E-02
Hydrolase	07XS0	3.70E-02
Conserved membrane protein	06GU5	3.70E-02
NA	08MDD	3.70E-02
Histidine kinase	05DDH	3.70E-02
NA	05Z6A	3.70E-02
NA	08RMG	3.70E-02
K01470 creatinine amidohydrolase EC 3.5.2.10	07UD0	3.70E-02
NA	07W59	3.70E-02
Auxiliary transport protein, membrane fusion protein	07TS9	3.70E-02
ABC-2 type transporter	07X4A	3.70E-02
Nudix hydrolase	05BZX	4.20E-02
NA	07V1J	4.30E-02
2-Amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	05K8U	4.30E-02
Ferritin	090A1	4.70E-02
Rmlc-like cupin family protein	0674Z	4.70E-02
GreA GreB family elongation factor	081QI	4.90E-02
Nitric oxide reductase	05DSQ	4.90E-02
Alcohol dehydrogenase	05CIS	4.90E-02
Cation diffusion facilitator family transporter	05PGM	5.00E-02
Nicotinamidase	08RWI	5.00E-02
NA	07V1J	4.30E-02
2-Amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	05K8U	4.30E-02
Ferritin	090A1	4.70E-02
Rmlc-like cupin family protein	0674Z	4.70E-02

GreA GreB family elongation factor	081QI	4.90E-02
Nitric oxide reductase	05DSQ	4.90E-02
Alcohol dehydrogenase	05CIS	4.90E-02
Cation diffusion facilitator family transporter	05PGM	5.00E-02
Nicotinamidase	08RWI	5.00E-02

Table 3-S2. List of eggNOG orthologue groups with distributions biased to PR+ Flavobacteriia.

Annotation	bactNOG ID	q-value
Rhodopsin	05CSB	0.00E+00
Beta-carotene 15,15'-monooxygenase, Brp Blh family	05FTR	0.00E+00
Uncharacterized protein conserved in bacteria (DUF2237)	08Z4C	3.90E-10
Deoxyribo-dipyrimidine photolyase	05CVP	1.10E-05
Amine oxidase	06AJX	2.00E-05
NA	9078	1.10E-04
NA	084QQ	1.30E-04
Sodium-dependent bicarbonate transporter	05EGC	1.30E-04
NA	06C6F	1.40E-04
NA	05Y5H	1.80E-04
NA	05Z9Q	1.80E-04
Cdp-alcohol phosphatidyltransferase	05DWS	1.20E-03
NA	05VHF	3.00E-03
Response regulator	08V8A	3.10E-03
NA	05ZNW	3.10E-03
NA	05DBJ	3.10E-03
Protein of unknown function (DUF2452)	090U0	4.40E-03
NA	08QRR	4.90E-03
Two component transcriptional regulator (Winged helix family	08PZG	5.60E-03
NADH ubiquinone oxidoreductase complex i intermediate-associated protein 30	05F58	6.60E-03
Transcriptional regulator, lysR family	05DUM	7.60E-03
Glutaredoxin	08DMZ	7.80E-03
Deoxyribodipyrimidine photolyase-related protein	05EXX	8.60E-03
NA	08NE7	9.70E-03
Protein of unknown function (DUF2490)	05IFM	1.00E-02
NA	05EPG	1.20E-02
Protein of unknown function (Porph_ging)	08SQC	1.30E-02
Phytoene	05EWK	1.40E-02
Inherit from NOG: Band 7 protein	07JWQ	1.60E-02
NA	05EZ4	1.80E-02
Spheroidene monooxygenase	08ZIP	2.20E-02
Histidine kinase	08M36	2.40E-02
Protein of unknown function (DUF2805)	05VPW	2.50E-02

Alkane 1-monooxygenase	05EEF	3.10E-02
Outer membrane efflux protein	05D2W	3.70E-02
Pyridoxamine 5'-phosphate oxidase-related, FMN-binding	05Q4V	3.70E-02
NA	066UH	3.70E-02
Protein of unknown function (DUF422)	05KAW	3.70E-02
TM2 domain	05TFM	3.70E-02
Hydrolase family 16	05FBU	3.90E-02
Transporter	05CT4	4.60E-02
NA	07ERR	4.70E-02
Flavin reductase domain protein FMN-binding protein	08UH3	5.00E-02