Genomic differentiation among two strains of the PS1 clade isolated from geographically separated marine habitats

Francy Jimenez-Infante1, David K. Ngugi1, Intikhab Alam2, Mamoon Rashid1, Wail Baalawi2, Allan A. Kamau2, Vladimir B. Bajic2 & Ulrich Stingl1

1Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia; and 2Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Correspondence: Ulrich Stingl, Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. Tel.: +966 (0) 12808 2381; fax: +966 128020152; e-mail: uli.stingl@kaust.edu.sa

Received 11 December 2013; revised 20 March 2014; accepted 12 April 2014. Final version published online 22 May 2014.

DOI: 10.1111/1574-6941.12348

Editor: Gary King

Keywords
Alphaproteobacteria; dilution-to-extinction; OCS116 clade; proteorhodopsin; PS1 clade; Red Sea.

Abstract
Using dilution-to-extinction cultivation, we isolated a strain affiliated with the PS1 clade from surface waters of the Red Sea. Strain RS24 represents the second isolate of this group of marine Alphaproteobacteria after IMCC14465 that was isolated from the East (Japan) Sea. The PS1 clade is a sister group to the OCS116 clade, together forming a putatively novel order closely related to Rhizobiales. While most genomic features and most of the genetic content are conserved between RS24 and IMCC14465, their average nucleotide identity (ANI) is < 81%, suggesting two distinct species of the PS1 clade. Next to encoding two different variants of proteorhodopsin genes, they also harbor several unique genomic islands that contain genes related to degradation of aromatic compounds in IMCC14465 and in polymer degradation in RS24, possibly reflecting the physicochemical differences in the environment they were isolated from. No clear differences in abundance of the genomic content of either strain could be found in fragment recruitment analyses using different metagenomic datasets, in which both genomes were detectable albeit as minor part of the communities. The comparative genomic analysis of both isolates of the PS1 clade and the fragment recruitment analysis provide first insights into the ecology of this group.

Introduction
Ecotypes are genetically cohesive groups of organisms playing distinct ecological roles (Cohan, 2001; Cohan & Perry, 2007). Differences in their ecological and physiological properties often determine their niche specificity (Cohan & Perry, 2007). The global ocean is teeming with examples of prokaryotic lineages where ecotype differentiation has been documented, and some examples for which ecologically distinct niches have been described. The best-studied lineages that harbor ecotypes include mostly the abundant marine prokaryotes, namely Prochlorococcus marinus (Rocap et al., 2003; Johnson et al., 2006), the SAR11 clade (Field et al., 1997; Carlson et al., 2009; Vergin et al., 2013), SAR86 (Dupont et al., 2012), Alteromonas spp. (Lopez-Lopez et al., 2005; Ivars-Martinez et al., 2008), and MGI (Massana et al., 2000). In these organisms, niche partitioning may be driven by several physicochemical gradients present in the water column including light, depth and nutrient availability (West et al., 2001 and references therein), temperature, seasonal effects, and geography (Brown et al., 2012 and references therein). Biological factors such as gain/loss of plasmids, phages, or mobile genetic elements also play an important role in niche adaptation (Cohan & Perry, 2007; Wiedenbeck & Cohan, 2011). The availability of cultures for many species of these abundant free-living marine groups has greatly enhanced our understanding of the role of microbial diversity and ecotype differentiation in the proliferation of different phylotypes in specific marine biomes.

The peculiarities of the Red Sea make it an interesting model to study microbial community structure–function relationships, especially as an environment that resembles the worst-case scenario for global warming in temperate oceans. The Red Sea is a landlocked, extremely oligotrophic marine ecosystem (Berninger & Wickham, 2005),
which is also characterized by a lack of any significant nutrient or precipitation inputs (Weisse, 1989), slow turnover of the deep-seawater mass (Cember, 1988), relatively high temperatures (24–35 °C at the surface and 22 °C from 200 m to the bottom, Ngugi & Stingl, 2012), high salinity (c. 40‰, Acker et al., 2008; Ngugi et al., 2012), and increased UV exposure (Boelen et al., 2002). We therefore postulated that these relatively extreme conditions should have an impact on the genomic repertoire of resident microorganisms, because most taxa would have to adapt to the ‘harsh’ Red Sea environment.

Because the ecotype paradigm is well studied among the abundant prokaryotes, in this study, we were interested in how speciation and adaptations to different environments might play a role among the low-abundance groups, which occupy the rare tail of microbial diversity. Members of the ‘rare biosphere’ usually encompass < 1% of the bacterioplankton communities and are prevalent and extremely diverse in marine environments (Sogin et al., 2006; Galand et al., 2009). They constitute both the dormant and active fractions of latent colonizers, which, along with the abundant groups, play an important role for the functioning of marine ecosystems (Campbell et al., 2011; Hugoni et al., 2013). However, the paucity of cultivated representatives for most of these low-abundance groups has hindered our understanding of their role in the world’s ocean. In our efforts to address this knowledge gap, we successfully isolated one such low-abundant bacterium from the euphotic zone of the Red Sea. This bacterium, designated here as strain RS24, belongs to the PS1 clade, a novel alphaproteobacterial group that was previously clustered together with members of the as-yet uncultivated OCS116 clade (Suzuki et al., 1997), but has recently been resolved as a separate lineage (Iverson et al., 2012). The OCS116 clade abundance and distribution seems to be correlated with phytoplankton blooms (Gonzalez et al., 2000; Morris et al., 2005; Treusch et al., 2009; Morris et al., 2012). Unlike OCS116 members, no relation with phytoplankton blooms has been described for members of the PS1 clade, and its abundance, distribution, and ecological role remains uncertain.

To date, there is only a single published draft genome of the PS1 clade available. Strain IMCC14465 (Yang et al., 2012), which is the first cultured member of this lineage, is an uncharacterized proteorhodopsin-containing strain that was isolated from the coastal waters off the East (Japan) Sea, where temperature (15–23 °C) and salinity (c. 33 psu) are comparatively lower and the concentrations of macronutrients (e.g., nitrate and phosphate) are typically higher than in the Red Sea (Kim et al., 2010, 2012; Thompson et al., 2013). Given the geographic separation, our study was designed to compare the genetic repertoire of PS1 clade members from these environmentally distinct habitats and estimate the abundance of the PS1 clade in different depths of diverse marine metagenomes, as well as in estuarine and limnic samples.

Materials and methods

Water sample collection

Seawater samples were collected from an open ocean site in the central Red Sea area (22°18.26987 N, 38°45.21546 E) at a depth of 19 m (July 19, 2011) using a horizontal water sampler. Collected samples were transported in 20-L carboys, which were prewashed with distilled water and rinsed with seawater. These samples were used for both inoculum and media preparation.

Culturing techniques

In the effort of isolating oligotrophic bacterioplankton from the Red Sea, we obtained a culture containing the PS1 clade cells. A high-throughput culturing (HTC) technique (Connon & Giovannoni, 2002) and the dilution-to-extinction method (Button et al., 1993) using a modified low-nutrient heterotrophic media (LNHM, Connon & Giovannoni, 2002; Stingl et al., 2007) were employed (see Appendix S1, Supporting information). Inoculum density and sterility of the media were checked by flow cytometry (Guava EasyCyte, Millipore) as previously described (Stingl et al., 2007), except that 1 × SYBR Green (Invitrogen) final concentration was used. Media were inoculated with seawater and diluted to a final concentration of c. 1 cell mL⁻¹ (Connon & Giovannoni, 2002). Incubation was performed at the in situ temperature of 28 °C in 96-well Teflon plates as previously described (Stingl et al., 2007; see Appendix S1).

Culture screening, identification, and sequencing

Cultures in each well were monitored and screened after four, eight, and twelve weeks postinoculation via cell counts and PCR (see Appendix S1). PCRs for bacterial 16S rRNA genes were performed with the universal primer pair 27F (Hongoh et al., 2003) and 1492R (Lane, 1991). Amplicons from putatively pure cultures [as confirmed by restriction fragment length polymorphism and microscopy, Appendix S1] were purified with MinElute columns (Qiagen) and bidirectionally Sanger-sequenced with the initial primer pair. Taxonomic identification and phylogenetic characterization of the isolated strains were performed using BLAST (Altschul et al., 1990), the RDP classifier (Cole et al., 2009), and the ARB software package (Ludwig et al., 2004).
Characterization of cultures and DNA extraction

For this study, one isolate affiliated with the low-abundance PS1 clade, designated here as strain RS24, was chosen for further genomic characterization. Cells representing a pure culture of strain RS24 were transferred into polycarbonate flasks for the measurements of growth characteristics ($n = 3$). Further studies on the effect of light or darkness on the growth of strain RS24 were conducted in the presence or absence of light supplied at 50 $\mu$M m$^{-1}$ s$^{-1}$ until cells reached late stationary phase ($n = 3$). Cells were frozen at $-80^\circ$C in 10% glycerol or 5% DMSO for long-term storage.

For DNA extraction, cells were grown in a 20-L polycarbonate carboy (Millipore), continuously sparged with sterile air at room temperature until they reached maximum cell densities ($c. 10^7$ cells mL$^{-1}$; c. 1 month). DNA extraction followed a slightly modified standard phenol–chloroform protocol (Lueders et al., 2004; see Appendix S1).

Transmission Electron Microscopy

For visualization using transmission electron microscopy (TEM), cells grown to exponential phase were fixed in 2.5% glutaraldehyde. A 50-mL volume of the culture was then filtered through a 0.1-μm membrane filter (Pall Life Sciences Supor$^{TM}$-100). Cells attached to the filter were resuspended in 2 mL of the same media, from where an aliquot (2 μL) was taken and distributed onto a 3-mm TEM grid and left to adsorb for 2 min. Subsequently, the grid was negatively stained with 2% uranyl acetate prior to microscopy at KAUST’s Imaging and Characterization Core Lab with a Titan G2 80–300 kV TEM (FEI Company) equipped with a 4 x 4 x 4 k CCD camera (US4000) and an energy filter model GIF Tridiem (Gatan, Inc.).

Genome sequencing, assembly, and annotation

The genome of strain RS24 was sequenced using both 454 (134 243 reads) and Illumina GAII (39 513 532 paired-end reads) sequencing technologies. Prior to assembly, raw Illumina sequence data were quality trimmed and filtered using FastQC (Andrews, 2010), Illumina-based filter (Assaf, 2011), and Trimmomatic (Lohse et al., 2012; see Appendix S1). Newbler (Margulies et al., 2005) with scaffolding option was used for assembly, followed by scaffolding using SSPACE (Boetzer et al., 2011) and gap closing with GapFiller (Boetzer & Pirovano, 2012) and Gap Closer (Luo et al., 2012) using default options. The final assembly consists of five scaffolds (N50 = 736 154 bp) with a total length of 1 957 724 bp. Annotation of this draft genome was carried out using an in-house pipeline (Alam et al., 2013) briefly described in the Appendix S1.

Predicted ORFs were assigned to functional categories using the cluster of orthologous genes (COGs) database (Koonin, 2002; updated on August 13, 2003) using BLASTP (E-value < 10$^{-5}$). For pathway reconstruction, predicted genes were searched against the nonredundant (nr) protein database in GenBank and mapped onto the KEGG pathway using MEGAN4 (Huson et al., 2011). Transporter protein families were annotated by BLAST using the Transporter Classification Database (Saier et al., 2006, 2009). Genes encoding for carbohydrate-active enzymes were automatically annotated using the CAZymes Analysis Toolkit (CAT) web server (Park et al., 2010) and validated using Pfam-based annotation (E-value: 10$^{-3}$; bitscore: 55; rule support level: 50).

Fragment recruitment analysis

To estimate the natural abundance and capture the divergence of the PS1 clade in the water column of the Red Sea and other varied aquatic provinces, we performed fragment recruitment analysis (Rusch et al., 2007) using the genomes of RS24 and IMCC14465 as queries against several metagenomic datasets (Table S1 and Appendix S1). The coverage of the recruited reads in these datasets was compared with the results based on representative genomes of a pelagic bacterial strain abundant in the euphotic zone (Pelagibacter sp. HTCC7211; Grote et al., 2012) and one that is ubiquitous in the marine mesopelagic zone (Nitrosopumilus maritimus SCM1; Walker et al., 2010). To account for length differences in the query genome and variability in the metagenomic sequencing efforts, we calculated coverage of the recruited reads by first normalizing the retrieved hit counts to the length of the smallest genome. The product was then multiplied by the average length of all hits in a dataset and divided by the size (in Mbp) of the respective metagenome. Fragment recruitment was performed using BLASTN with the following parameters: eval $10e^{-4}$, -xdrop_gap 150 -penalty -1 -reward 1 -soft_masking true -lcase_masking, and 75% minimum cutoff of the query length. In all cases, the 23S, 16S, and 5S regions were masked in each reference genome prior to recruitment, and only reads with a length of $\geq$ 200 bp were considered.

Comparative genomics

The relatedness of both PS1 strains was assessed at the nucleotide level through the analysis of ANI – a measure of...
of species relatedness, as implemented in the JSpecies software (Richter & Rossello-Mora, 2009) using the default BLAST parameters. To harmonize the genome annotations, the genome of IMCC14465 was re-annotated using an in-house pipeline (Alam et al., 2013). Similarities and differences as well as paraslogs among the proteomes of strains RS24 and IMCC14465 were identified and evaluated using BLAST score ratio (BSR) analysis with a threshold value of 0.4 that represents c. 30% of amino acid identity (AAI) over c. 30% peptide length (Rasko et al., 2005). A total of 1909 protein-coding genes in strain RS24 were used as query against IMCC14465. Parvibaculum lavamentivorans DS1 (Schleheck et al., 2011), the most closely related sequenced genome to the PS1 clade as revealed by preliminary RAST annotation analysis (Aziz et al., 2008), was included for completeness. Here, also the average AAI of orthologous genes between RS24 and IMCC14465 was obtained.

The conservation of gene order (synten) between RS24 and IMCC14465 was calculated using synteny-based scripts described by Yelton et al. (2011). To identify low-synten loci in these genomes, whole-genome sequence alignments were also performed using ProgressiveMauve (Darling et al., 2010).

Phylogenetic analysis

Full-length 16S rRNA and proteorhodopsin (PR) genes were retrieved from both draft genomes and used to calculate phylogenetic relationships. Phylogenetic trees for 16S rRNA genes were inferred in ARB (Ludwig et al., 2004) using the maximum-likelihood approach (1000 replicates) and the GTR + GAMMA + I substitution model. Phylogenetic trees using neighbor-joining as well as Bayesian algorithms (MrBayes; Huelsenbeck & Ronquist, 2001) were also constructed to validate the tree topology and nodal support (see Appendix S1).

The phylogeny of PR was inferred by a neighbor-joining approach (Saitou & Nei, 1987) with Jukes-Cantor distance correction matrix using 1000 replicates as implemented in EPOS software (Griebel et al., 2008). Maximum parsimony (PAUP 4.0) was also used to test the stability of the PR tree topology (more details are provided in Appendix S1).

To better resolve the evolutionary relationship of the PS1 clade within the Alphaproteobacteria, we further tested the phylogenetic placement of both PS1 strains using a concatenated protein alignment of 14 orthologous genes conserved in the genomes of 98 other alphaproteobacterial strains, and four out-group organisms belonging to Beta-, Gamma-, Delta-, and Zeta-Proteobacteria, respectively (Table S2). The 14 orthologous clusters were generated using the automated phylogenomic pipeline Hal (Robbertse et al., 2011). Core gene sequences from individual genomes were concatenated, and the resulting sequences were aligned using CLUSTALW (Thompson et al., 1997). Problematic regions in this superalignment were removed using GBLOCKS (Castresana, 2000) using the liberal settings to retain the maximum number of characters in the alignment. Prior to phylogenetic inference using RAXML (Stamatakis, 2006), the best model for amino acid substitution (i.e., PROTGAMMAI) was estimated using PROTEST3 (Darriba et al., 2011) based on the liberal superalignment. Bootstrap support was estimated based on 100 replication using the rapid bootstrapping and WAG distance matrix options implemented in RAXML. Additionally, Bayesian inference was calculated with the same parameters used for 16S analysis.

Nucleotide sequence accession number

The draft genome of strain RS24 has been deposited in GenBank under the Accession Number AWXE0000000.

Results and discussion

Phylogenetic placement of the PS1 clade members

16S rRNA gene-based phylogenetic analyses placed strain RS24 (and strain IMCC14465) within the PS1 clade (Fig. 1; Fig. S1), clustering together with several environmental sequences. The PS1 clade forms a monophyletic group within the class Alphaproteobacteria. These environmental sequences were until recently considered as part of the as-yet uncultivated OCS116 clade (Suzuki et al., 1997), in the SILVA database (v. 111, Quast et al., 2012). Based on the high bootstrap support (for both maximum-likelihood and Bayesian analyses; Fig S1) in the topology of these two sister clades, coupled with their interspecific distance (11.6%), it is evident that they represent two separate monophyletic groups. Furthermore, the intraspecific divergence of sequences within these clades ranges from 4.2% (OCS116 clade) to 4.0% (PS1 clade), implying that there is a significant diversity even among members of these lineages. Our BLAST-based searches against clone library sequences from Puget Sound (Lau & Armbrust, 2006; Iversen et al., 2012), where PS1 sequences have been previously reported, indicate that they account for < 1% of the total prokaryotes (using a cutoff of 95% identity to 16S rRNA genes from strains RS24 and IMCC14465).

Furthermore, evidence based on fragment recruitment analysis using either of the PS1 representative genomes against several metagenomic datasets from the Red Sea and other oceans around the globe indicates that the oceanic
The prevalence of this clade is at most one order of magnitude lower than that of a strain of the ubiquitous marine heterotrophic group SAR11 or the ammonia-oxidizing thaumarchaeota (Fig. S2). However, these results reflect not only the low abundance of this clade but also its widespread occurrence in the water marine column. The coverage pattern in metagenomic data within the same water column implies that members of this group have a preference for the upper euphotic zone; in the Red Sea, they were relatively abundant around the deep chlorophyll maximum (50 m, Fig. S2). The relatively low ANI (below 95%) of reads of metagenomic fragments recruited by strain RS24 (also IMCC14465) in datasets from the Red Sea (Fig. S3) indicates that there are other divergent genotypes, which might have (so far unknown) well-defined seasonal or geographical patterns of abundance. Interestingly, RS24 recruited remarkably well using metagenomic samples from the South Pacific subtropical gyre (HOT179) at 25 and 70 m with ANIs of > 90%. Nevertheless, given the depth from which RS24 was isolated (19 m), future studies should include metagenomic data from 10 to 20 m depth to better resolve the biogeography of PS1 clade.

The closest validly described relative of the PS1 clade (and OCS116 clade) at the 16S rRNA gene level is P. lavamentivorans (Fig. 1; Fig. S1) with a sequence divergence of 12%, which means that they belong to different lineages. The environmental sequences affiliated with the PS1 lineage stem from very diverse marine habitats (Fig. S1), such as sponge- and coral-associated microbial biomass, the oligotrophic Sargasso Sea, as well as tropical and temperate regions such as Costa Rica and Panama. They were originally recovered from estuarine water ecosystems in Puget Sound (PS)/Saanich inlet, where the acronym ‘PSI’ was derived from (Iversen et al., 2012).

Whereas the 16S rRNA gene of strain RS24 was virtually identical to that of IMCC14465 (99%), the corresponding proteorhodopsin (PR) gene-based phylogeny was clearly distinct (Fig. 2). Proteorhodopsins are retinal-bound transmembrane proteins broadly found in marine bacterioplankton (Beja et al., 2000; de la Torre et al., 2003; McCarren & DeLong, 2007; Miyake & Stingl, 2011; Slamonovits et al., 2011) and viruses (Yutin & Koonin, 2012; Philosof & Béja, 2013), which absorb light at different absorption maxima [525 nm, green-tuned PR (GPR) or 490 nm, blue-tuned PR (BPR); Man et al., 2003] to potentially generate proton motive force via light-driven proton pumping (Beja et al., 2000, 2001). Based on the spectral tuning trait of characterized PRs, which is determined by the amino acid residue at position 83 of the aligned PRs (Fig. S4), being either a glutamine (Q) or leucine (L; Man et al., 2003), we predict that RS24 harbors a BPR, whereas IMCC14465 carries a GPR, with both sharing an AAI of only 62%. The occurrence of two differentially tuned PR proteins in the PS1 clade is consistent with the previously reported predominance of BPRs in the oligotrophic provinces of the world’s oceans such as the open ocean waters.
of the Red Sea and Sargasso Sea (Sabehi et al., 2005; Rusch et al., 2007; Sabehi et al., 2007) and the prevalence of GPRs to coastal (Rusch et al., 2007) or surface layer (Sabehi et al., 2007) environments. Thus, the spectral properties of PR correlate well with the ambient light regimes at the points of isolation, namely the East Sea (strain IMCC14465) and the Red Sea (strain RS24).

The most closely related sequences to the BPR of RS24 are several pelagic metagenomic fragments of the global ocean survey (GOS, Yooshef et al., 2007) and an alpha-proteobacterial fosmid clone, HFL1_19P19 (70% AA; Fig. 2), recovered from samples at the Hawaii Aloha station (Martinez et al. 2007). In contrast, the GPR of IMCC14465 falls in a distant branch containing PR sequences related to the SAR92 (HTCC2207) and the BD1-7 (HTCC2143) clades, suggesting that it may have been laterally acquired from a gammaproteobacterial donor. This lateral transfer of PR genes from external groups was also observed between *Vibrio* and *Glaciecola* species (Akram et al., 2013), presumably owing to its relatively simple gene arrangement and prospective contribution on energy metabolism (Frigaard et al., 2006). Although the role(s) of these PRs is still an open question, based on *BLAST*-based searches, it appears that the PR cluster affiliated with strain RS24 has no close similarities to those from cultured representatives and may be yet uncharacterized, which is consistent with the previous observations that the Red Sea harbors unique variants of PR genes (Sabehi et al., 2003).

**Morphology, growth, and genome features of strain RS24**

Strain RS24 constitutes the second isolate of the PS1 clade, but is the first strain from the Red Sea. There is no information available on growth characteristics or morphology of strain IMCC14465, which was isolated from coastal waters of the East Sea (Yang et al., 2012). When visualized by TEM, exponentially grown cells of strain RS24 appeared as short curved rods harboring a single polar flagellum, with cells of <1 μm in length and a width of about 0.2 μm (Fig. 3). Cells grown in LNHM prepared from Red Sea water at 28 °C in the dark had growth rates of 0.90 ± 0.028 generations day⁻¹ (n = 3), reaching a maximum density of 2 × 10⁷ cells mL⁻¹.

As strain RS24 possesses PR, also the effects of light on its growth behavior were tested. However, no major differences were found in growth rate or yield under the different light conditions (Fig S5). It is possible that a hypothesized effect was masked by the composition of our media. Still, future starvation experiments of strain RS24 and use of other growth-related parameters (e.g., ATP production and gene expression patterns), rather than shear cell numbers, are necessary to investigate the
functionality of PR in this strain, as has been demonstrated for other marine bacteria such as SAR11 (Steindler et al., 2011) and Vibrio (Gomez-Consarnau et al., 2010).

The assembled draft genome of strain RS24 has a total length of 1,957,724 bp, consisting of 5 scaffolds (N50 = 736,154 bp; min length of 25,260 bp) with a GC content of 45.1%. It encodes for 1909 putative protein-coding genes, 39 tRNA genes, and 1 rRNA operon. An estimation of the completeness of the genome using 66 COGs conserved single-copy genes taken from the ‘bacterial core gene set’ from the Human Microbiome Project Data (Mitreva, 2009) shows that 100% of the core genes on the genome of strain RS24 were recovered.

The overlapping fractions of the RS24 and IMCC14465 genomes corresponded to c. 77.3% of their assembled nucleotides, with an ANI of 80.5%. Based on the previous bacterial ‘species delineation’ analyses, about 95% ANI corresponds to 70% DNA–DNA hybridization for the same species (Goris et al., 2007). This indicates that strain RS24 represents a novel species within the PS1 clade, proposed here as Candidatus Micropelagos thuwalensis RS24. The name reflects the small-sized cells (‘micro’) obtained from the pelagic zone (‘pelagos’) of Red Sea waters around the village of Thuwal (‘thuwalensis’).

Core genome of the PS1 clade

Comparison of the protein-coding genes of strain RS24 (1909 in total) with predicted proteins from IMCC14465 (1870 in total) using BLAST score ratio analysis (Rasko et al., 2005) showed that c. 86% of the proteins from RS24 were homologous to IMCC14465, with an average AAI of 88%. One paralog predicted as elongation factor Tu was present in the core genome and was not taken into account. Complementary analysis with a synteny-based annotation approach (Yelton et al., 2011) showed that 84% of these core genes were syntenic and therefore probably also functionally conserved among these strains (Yelton et al., 2011 and references therein). The core genome consists of 1639 proteins (Fig. 4a), of which 88% could be assigned to COG functional categories (Table 1). Most represented COGs encode for proteins involved in the transport and metabolism of amino acids (8.1%) and lipids (8.3%), translation and biogenesis of ribosomes (8.5%), and general functions (11.6%).

The potential taxonomic profile of protein-coding genes in the core genome generated using BLAST-based searches indicates that the nearest phylogenetic neighbors (closest homologs) in sequenced genomes are mostly within the domain Bacteria (95%) and particularly to the Alphaproteobacteria (77%). Most of the assigned fraction had close homologs to proteins belonging to the orders Rhizobiales (42%), Sphingomonadales (25%), unclassified Alphaproteobacteria (21%), and Rhodobacterales and Rhodospirillales (6% each).

 Concatenated alignments of multiple conserved marker genes generally provide a better phylogenetic resolution than single gene-based phylogenies for species estimation (Szollosi et al., 2012). We accordingly tested the evolutionary relationship of the PS1 clade to sequenced members of the main groups within Alphaproteobacteria based on a set of 14 conserved single-copy genes present in 98 complete and draft genomes (Fig. 5; Table S2). Although the 16S rRNA gene-based phylogeny supports a distant affiliation of the PS1 clade members (strains RS24 and IMCC14465) with the orders Rhizobiales and Rhodobacterales, the concatenated gene tree places the PS1 clade as a deep-branching lineage and a sister group to P. lavamentivorans (Fig. 5), a representative of the novel family Rhodobiaceae (Schleheck et al., 2011). Because of the deep-branching order, the high bootstrap support, and the evolutionary distance to members of the order...
Rhizobiales (c. 30% sequence dissimilarity in alignment of 14 concatenated core protein sequences compared with strain RS24), we argue that the PS1 clade putatively represents a novel bacterial order. This potentially also questions the family-level taxonomic rank of Rhodobiaceae within the order Rhizobiales. A better understanding of the affiliation of the PS1 clade members to the Rhodobiaceae family and the order Rhizobiales will require the inclusion of genomes from the OCS116 clade, which seems to be the closest relatives to the PS1 clade, at least at the 16S rRNA gene level.

The core metabolic inventory of strains RS24 and IMCC14465

The core genome of the PS1 clade encodes genes for a complete set of enzymes involved in the tricarboxylic acid cycle (TCA) and gluconeogenesis (Fig. S6). Although some genes for glycolysis via the Embden–Meyerhof–Parnas (EMP pathway) are present, this pathway is most likely nonoperational in both strains due to the absence of the key enzyme for the utilization of hexoses, 6-phosphofructokinase (pfk). This enzyme catalyzes the conversion of fructose 6-phosphate to fructose 1,6-bisphosphate (Artimo et al., 2012). Given the presence of genes such as the phosphogluconate dehydratase and the key enzyme 2-dehydro-3-deoxy-phosphogluconate aldolase (Artimo et al., 2012), glycolysis is most likely taking place via the Entner–Doudoroff (ED) pathway. The intermediates of this pathway can then potentially be used to fuel the pentose phosphate pathway to generate reducing equivalents [NAD(P)H] or produce 5-carbon sugars (Fig. S6). The presence of these metabolic features coupled with the absence of photosynthetic genes or genes for carbon dioxide fixation suggests an aerobic heterotrophic lifestyle for members of the PS1 clade.

The core genome also harbors an array of carbohydrate-active enzymes (Fig. S7; Table S3). Among these are a number of glycoside hydrolases (GHs), which are involved in polysaccharide hydrolysis (Cantarel et al., 2009), including β-glucosidase (GH3), α-amylase (GH13), endoglucanase (GH16), and GHs acting on peptidoglycan (GH23 and GH103). The presence of three genes that encode for GH3 family proteins, but a lack of GHs encoding for cellulases in both genomes (i.e., GH5, -6, -9, etc.; Cantarel et al., 2009), implies that members of the PS1 clade are not primary cellulose degraders, but probably opportunists (Berlemont & Martiny, 2013) that only use cellobiose and small oligosaccharides, presumably generated from the degradation of longer polymers by other organisms. The degree of usage of these resources seems to be strain-specific because strain RS24 carries multiple (unique) GH3s copies compared with IMCC14465 (Table S3). Moreover, the occurrence of genes that encode for putative sugar transporters, including sodium-dependent glucose co-transporter proteins sglT and dhlC, reinforces the hypothesis of potential usage of small oligosaccharides as a carbon and energy source.

The presence of genes that encode for PR and the retinal biosynthesis machinery (Table S4) further suggests the potential for light-dependent ATP generation (Beja et al., 2000, 2001). This was demonstrated in the abundant oligotrophic marine bacterial group SAR11 and preferentially occurred under carbon-depleted conditions (Steindler et al., 2011). Similar photoheterotrophic mechanisms have been proposed for high-nutrient-utilizing bacteria such as Vibrio (Gomez-Consarnau et al., 2010; Wang et al., 2012) and Flavobacteria (Gomez-Consarnau et al., 2007; Lami et al., 2009).
The utilization of one-carbon (C1) compounds is suggested by the presence of putative genes involved in the conversion of tetrahydrofolate (Fig. S8). The tetrahydrofolate-linked oxidation pathway, which oxidizes reduced C1 units (e.g., formaldehyde) to CO2, yields energy in the form of reduced nucleotides and ATP (Sun et al., 2011).

The core genome of both PS1 clade strains also encodes numerous putative alcohol dehydrogenases (12 ADHs; Table S5; Fig. S9). ADHs are ubiquitous in the three domains of life and facilitate the NAD(P)H-dependent reversible interconversion of alcohols to the corresponding aldehydes or ketones (Timpson et al., 2012 and references therein). In the PS1 clade, five ADHs were predicted to contain zinc-binding domains (Pfam PF00107; Table S5). Nevertheless, further phylogenetic analyses classify them within the quinone oxidoreductase (QOR1) and L4BD/prostaglandin dehydrogenase (PGDH) groups (Fig. S9). These ADHs are part of the medium-chain dehydrogenases/zinc-binding (MDR) superfamily but lack the canonical zinc-binding site (QOR; Thorn et al., 1995) or may even be metal-independent (PFGH; Hassaninasab et al., 2011). Other PS1 clade ADHs (and five unique for RS24) were grouped throughout the ADH families, mostly within the short-chain dehydrogenases, showing a diversification of the PS1 clade ADH profile and potential substrate utilization.

Genes encoding enzymes of the glyoxylate shunt (isocitrate lyase and malate synthase; Kornberg & Krebs, 1957) are also present in the core genome. This anaplerotic pathway enables some organisms to use substrates entering the central carbon metabolism (e.g., short-chain fatty acids, alcohols, and methylated compounds) as sole carbon and energy source and thus, through a broader substrate range, might make members of the PS1 clade flexible in an environment already limited by low-nutrient concentrations. Other genes encoding for enzymes necessary for assimilatory sulfate reduction and sulfate uptake are also present. Unlike in the SAR11 clade (Tripp et al., 2008), no genes for dimethylsulfoniopropionate degradation were found in the genomes of either PS1 clade strain, which suggests that members of this group preferentially use sulfate rather than reduced sulfur from dimethylsulfoniopropionate for growth. In addition, both genomes encode enzymes for the biosynthesis of all 20 essential amino acids.

**Mechanisms for coping with the oligotrophic marine environment**

Marine bacterioplankton plays an important role in oceanic carbon and nutrient cycling through incorporation and redistribution of dissolved organic matter (DOM) and inorganic nutrients (Jiao et al., 2010; Kujawinski, 2011). However, most essential nutrients (phosphorus, iron, and nitrogen) and trace micronutrients (Co, Zn, Cu, and Ni), which are necessary for cell growth and the functionality of different enzymes, occur in extremely low concentrations in most parts of the global ocean.
(Bertrand et al., 2011; Aparicio-Gonzalez et al., 2012). Subtropical oceans in particular are severely depleted in these compounds, thus limiting the productivity of these marine environments (Wu et al., 2000, 2001; Mahowald et al., 2005; Martiny et al., 2011). We found that the genome of PS1 clade strains encodes numerous high-affinity transporter genes, which is a genomic hallmark for cells proliferating in low-nutrient environments (Lauro et al., 2009) such as the Red Sea (all year round) and the East Sea (at the point and season of strain IMCC14465 isolation; see Table S6). These include the high-affinity transport systems for phosphate (pstSCABU), zinc, iron (siderophores), and ammonium (amt) as well as sulfate and magnesium uptake systems.

A large proportion of the genes predicted as transporter proteins in the PS1 clade are assigned as ATP-binding cassette (ABC) transporters (c. 22%; Table S7). These are known to transport a wide variety of substrates including amino acids, oligopeptides, and sugars (Davidson et al., 2008). The presence of ABC-type transporters for phosphate, iron and amino acids (Table S8), TonB-dependent transporters (Table S7; TBDT/OMR family: 5.6%), and high-affinity nutrient acquisition systems (e.g., PTS and TRAP-1.6% each; Table S7) implies that they may be

---

Fig. 5. Core genome phylogeny of the Alphaproteobacteria based on a maximum-likelihood tree of 14 concatenated single-copy genes conserved in 98 alphaproteobacterial genomes and the two PS1 clade strains (in bold). The tree was built using RAxML based on the PROTGAMMAI substitution model and the WAG distance matrix (Alphaproteobacteria substitution). Major orders within the Alphaproteobacteria are denoted. Same out groups as in Fig. 1 (not shown) were used for tree rooting.
essential in their oligotrophic and pelagic environment allowing these microorganisms to the use of a broad range of substrates for growth. While ABC transporters are functionally promiscuous and virtually present in all marine bacteria, TBDT are not only highly diverse, but also very substrate-specific (Tang et al., 2012). TBDT have been found as predominant membrane proteins in Gram-negative marine bacteria (Morris et al., 2010). Using proton motive force (PMF), TBDT transport nutrients across the outer membrane. Thus, the presence of both TBDT and proteorhodopsin genes in diverse bacterial groups may represent a potential light-induced PMF to fuel nutrient uptake (Morris et al., 2010). Consequently, strains RS24 and IMCC14465 may take advantage of the TBDT, in conjunction with photoheterotrophy, to sequestrate DOM and assimilate carbohydrates, amino acids, and other scarce nutrients.

Most marine bacteria and archaea accumulate many different organic osmolytes (e.g., ectoine, glycine betaine, trehalose, and glutamate) in response to high osmotic stress (Burg & Ferraris, 2008). Given the bioenergetic burden of synthesizing these osmoprotectants, many of these microorganisms prefer to obtain their compatible solutes from exogenous sources in their surrounding environments (Oren, 2011). Accordingly, the core genome of PS1 clade members encode for a choline-glycine betaine transporter and an ABC-type proline-glycine betaine transport system. In addition, both strains contain small-conductance mechanosensitive ion channel proteins (Table S7), which help to protect cells from hypoosmotic shock when a cell transitions from a high osmolarity to a low osmolarity environment (Booth et al., 2007). Other genes in the PS1 core genome involved in stress response include those for glutathione-based redox and nonredox reactions that are associated with the COG categories 'energy production and conversion' and 'metabolism and transport of sugars and amino acids' in strain RS24, and 'transcription', 'replication, recombination and repair', and 'inorganic ion transport and metabolism' in the case of strain IMCC14465 (Fig. 4b).

Complementary results based on whole-genome alignment of strain RS24 vs. IMCC14465 with Progressive-Mauve (Darling et al., 2010) indicated that there were several loci along the genome of RS24 carrying highly divergent proteins or those absent in IMCC14465 (Fig. S10a and b). The most striking perturbations of synteny are found in four loci of approximately 52, 25, 21, and 34 kb in length (Table S11). The 52-kb region (Fig. S10b) represents the largest genomic island in strain RS24 as it harbors 21% of the RS24-specific genes previously identified using BSR analysis. Moreover, the genes in this stretch of the genome have closest blast hits from different Alphaproteobacteria (34.5%) followed by Gammaproteobacteria (24.1%) and Firmicutes (10.3%), among others (Table S11), which could possibly have been horizontally transferred. The presence of several genes associated with the metabolism of polysaccharides such as glycosyltransferases, TonB-dependent receptors, major facilitator superfamily transporters, and amidohydrolases in this region and the other three low-coverage regions in strain RS24 further suggests a specialization in the degradation of polymers. There are also a large number of genes involved in lipopolysaccharide biosynthesis and capsular production present in the 25-kb genomic island (Table S11), which might be beneficial in the ocean surface waters by facilitating attachment to marine snow (Decho, 1990) and resistance to phage infection (Nwodo et al., 2012).

Using the CRISPRs Finder tool (Grissa et al., 2007), we also identified a putative CRISPR-associated (cas) protein and a CRISPR repeat sequence element of 828 bp among the unique genes of strain RS24. CRISPR–Cas elements have been identified in most sequenced Archaea (c. 90%) and many Bacteria (c. 40%; Horvath & Barrangou, 2010; Makarova et al., 2011) and provide defense against viruses and plasmids by targeting nucleic acids in a sequence-specific manner. The occurrence of these genes coupled with several genes putatively conferring resistance in the genome (e.g., beta-lactamase domain-containing proteins, glyoxalase/bleomycin protein–dioxygenase, and multidrug resistance proteins; Table S11) may confer a

**Unique metabolic pathways in genomes of members of the PS1 clade**

Although the two strains share a substantial fraction of their genetic repertoire (Fig. 4a), we found that about 14% and 12% of the predicted proteins appear unique to RS24 (267 in total; Table S11) and IMCC14465 (222 in total; Table S12), respectively. While 36% (RS24) and 57% (IMCC14465) of these species-specific proteins have so far unknown functions, most of the striking differences between proteins predicted functional comprise COGs that are associated with the COG categories 'energy production and conversion' and 'metabolism and transport of sugars and amino acids' in strain RS24, and 'transcription', 'replication, recombination and repair', and 'inorganic ion transport and metabolism' in the case of strain IMCC14465 (Fig. 4b).
competitive advantage in the marine euphotic zone, where the organism was isolated from.

In contrast, no CRISPR–Cas elements were found in strain IMCC14465, and the genome appears to harbor a putative bacteriophage of approximately 4.4 kb, which was identified with the Prophage Finder tool (Bose & Barber, 2006). The putative prophage encodes a variety of proteins participating in capsid and spike formation and general bacteriophage assembly (Table S12), all of which are virtually identical to proteins from the bacteriophage PhiX174 (family Microviridae). This Microvirus is very specific to the enterobacteria and seems to be extremely rare in seawater ecosystems (Roux et al., 2012). Based on this, it is highly unusual that this stretch of the assembly actually forms part of the genome of IMCC14465.

Strain IMCC14465 also carries three putative genomic islands (Table S12). Islands one (9.6 kb) and two (17.7 kb) contain genes from the phage integrase family (PF00589) such as tyrosine recombinase XerC-like, site-specific recombinase phage integrase, and an integrase recombinase that may be related to prophage establishment. Island 3, which is the largest one in this strain (23.1 kb), interestingly contains a high representation of genes potentially involved in the degradation of recalcitrant compounds such as phenol, naphthalene, and toluene (hydrolases, hydroxylases, and mono-/dioxygenases). These genes are absent in the Red Sea strain and might represent another indicator for an environmentally driven differentiation between these two strains. The potential ability to use phenol and other petroleum hydrocarbons by IMCC14465 may be related to the fact that these pollutants can occur in high concentrations in certain areas of the East Sea, especially in the northwestern region (Belan, 2003).

Conclusion

The PS1 clade is usually part of the rare tail of abundance in marine bacterioplankton communities. Like the sister OCS116 clade, very little was known about their roles in the global oceans. Our successful cultivation of a second OCS116 clade, very little was known about their roles in marine bacterioplankton communities. Like the sister Rhizobiales. Our robust analysis using core marker genes indicates that the PS1 clade probably represents a novel bacterial order, thus disputing their placement within the as-yet uncultivated OCS116 clade in the SILVA database. The central metabolism of members of the PS1 clade seems not only to be driven by the hydrolysis of oligosaccharides and peptides, but also potentially photoheterotrophy using proteorhodopsin. Even for these low-abundance bacteria, we found many examples in the genomes on how the two species have managed to occupy ecologically different ecological niches. Consequently, this also might indicate that they are able to occupy different niches within the same body of water as indicated in similar coverage of the genetic content of both strains in diverse metagenomic datasets.

Acknowledgements

We would like to thank KAUST Coastal and Marine Resources (CMOR) for facilitating with sample collection, the KAUST Core Lab Genomic facilities for the genome sequencing, and Rachid Sougrat in the Imaging and Characterization Lab for the bacterial TEM micrographs.

References


Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P & Tiedje JM (2007) DNA–DNA hybridization...


Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Fig. S1.** 16S rRNA gene-based phylogenetic tree showing the evolutionary relationship of the PS1 clade (RS24: red; IMCC14465: blue), the OCS116 clade, and other Alphaproteobacteria.

**Fig. S2.** Fragment recruitment comparison of RS24 and IMCC14465 with Pelagibacter sp. HTCC7211 and Nitrosopumilus maritimus SCM1, against metagenomes from the Red Sea, Mediterranean Sea (MED), Hawaii Ocean Time-Series (HOT), Bermuda Atlantic Time-Series (BATS), Eastern Tropical South Pacific (ETSP) in fall, winter, and summer, the Amazon River, Lake Gatun, and GOS data from Chesapeake Bay (CBay) and Delaware Bay (Dbay).

**Fig. S3.** Box plot representing the divergence of the genomic content of strain RS24 against different metagenomic reads from the Red Sea, Mediterranean Sea (MED), Hawaii Ocean Time-Series (HOT), Bermuda Atlantic Time-Series (BATS), Eastern Tropical South Pacific (ETSP) in fall, winter, and summer, the Amazon River, Lake Gatun, and GOS data from Chesapeake Bay (CBay) and Delaware Bay (Dbay).

**Fig. S4.** (Partial) alignment of proteorhodopsin sequences including sequences of RS24 and IMCC14465.

**Fig. S5.** Growth of strain RS24 cells in the presence of light (open circles) or under the dark (filled triangles) using modified LNHM at 28 °C.

**Fig. S6.** Predicted TCA, glycolysis/gluconeogenesis, ED, and pentose phosphate pathways, showing enzymes (highlighted in gray) present in the core genome of the PS1 clade.

**Fig. S7.** Relative abundance of putative carbohydrate-active enzymes predicted from the core genome of strains RS24 and IMCC14465 according to the CAZY database nomenclature (Park et al., 2010).

**Fig. S8.** Predicted tetrahydrofolate-linked pathway for C1 metabolism in the core genome of the PS1 clade.

**Fig. S9.** Unrooted phylogenetic tree of ADHs showing the putative classification of ADH proteins from the PS1 clade core genome (in bold) and unique ADHs in RS24 (in red) relative to several well-characterized ADH sequences.

**Fig. S10.** MAUVE-based whole-genome alignment comparison of strains RS24 and IMCC14465.

**Table S1.** Metagenomic libraries and their associated metadata that were used for fragment recruitment analyses.

**Table S2.** Single-copy orthologous genes conserved among strains RS24, IMCC14465, 98 alpha-proteobacterial genomes, and 4 outgroup genomes.

**Table S3.** Carbohydrate-active enzymes identified between the core and unique protein-coding genes in the genomes of strains RS24 and IMCC14465.

**Table S4.** Proteorhodopsin and retinal biosynthesis genes predicted in the genomes of strains RS24 and IMCC14465.

**Table S5.** ADH genes present in strains RS24 and IMCC14465.

**Table S6.** Environmental conditions in the Red Sea and the East (Japan) Sea.

**Table S7.** Transporters representation in both strains of the PS1 clade based on the Transporter Classification Database.

**Table S8.** Predicted ABC transporters in the PS1 core genome.

**Table S9.** Stress response genes in the core genome of the PS1 clade.

**Table S10.** Motility-associated, chemotaxis and sensor-regulatory genes predicted in the core genome of strains RS24 and IMCC14465.

**Table S11.** Unique genes and genomic islands in strain RS24.

**Table S12.** Unique genes and genomic islands in strain IMCC14465.

**Appendix S1.** Supplementary methods.