OMICS A Journal of Integrative Biology
Volume 20, Number 2, 2016
Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2015.0142

Picoplankton Bloom in Global South? A High Fraction of Aerobic Anoxygenic Phototrophic Bacteria in Metagenomes from a Coastal Bay (Arraial do Cabo—Brazil)

Rafael R. C. Cuadrat,[1,2,5] Isabel Ferrera,[2,4] Hans-Peter Grossart,[2,3] and Alberto M. R. Dávila[1]

Abstract

Marine habitats harbor a great diversity of microorganism from the three domains of life, only a small fraction of which can be cultivated. Metagenomic approaches are increasingly popular for addressing microbial diversity without culture, serving as sensitive and relatively unbiased methods for identifying and cataloging the diversity of nucleic acid sequences derived from organisms in environmental samples. Aerobic anoxygenic phototrophic bacteria (AAP) play important roles in carbon and energy cycling in aquatic systems. In oceans, those bacteria are widely distributed; however, their abundance and importance are still poorly understood. The aim of this study was to estimate abundance and diversity of AAPs in metagenomes from an upwelling affected coastal bay in Arraial do Cabo, Brazil, using *in silico* screening for the anoxygenic photosynthesis core genes. Metagenomes from the Global Ocean Sample Expedition (GOS) were screened for comparative purposes. AAPs were highly abundant in the free-living bacterial fraction from Arraial do Cabo: 23.88% of total bacterial cells, compared with 15% in the GOS dataset. Of the ten most AAP abundant samples from GOS, eight were collected close to the Equator where solar irradiation is high year-round. We were able to assign most retrieved sequences to phylo-groups, with a particularly high abundance of *Roseobacter* in Arraial do Cabo samples. The high abundance of AAP in this tropical bay may be related to the upwelling phenomenon and subsequent picoplankton bloom. These results suggest a link between upwelling and light abundance and demonstrate AAP even in oligotrophic tropical and subtropical environments. Longitudinal studies in the Arraial do Cabo region are warranted to understand the dynamics of AAP at different locations and seasons, and the ecological role of these unique bacteria for biogeochemical and energy cycling in the ocean.

Introduction

Marine environments cover 70% of the Earth's surface. These habitats show great variation in temperature, pressure, and salinity. They harbor a wide range of microorganisms from the three domains of life (Archaea, Bacteria, and Eukarya), which are responsible for ~98% of marine primary production (Kennedy et al., 2010; Sogin et al., 2006). This huge biodiversity has great potential, as its exploration affords discovery of new enzymes for industrial use. However, only from 0.001% to 1% of environmental microorganism can be identified using culture-dependent approaches (Kennedy et al., 2010; Pace, 1997; Tringe et al., 2005). To overcome this limitation, metagenomic studies have been conducted using samples from a variety of aquatic environments: from costal seawater, deep seawater, and open ocean waters; and in freshwater from rivers and lagoons (Ghai et al., 2011; 2012; Konstantinidis et al., 2009).

Nevertheless, the microbial diversity in the marine waters of the Brazilian coast remains poorly characterized. The Brazilian coast extends for 7491 km, and it is influenced by the warm North Brazilian Current in the northern portion, the cold Malvinas/Falklands Current in the southern portion, and to a lesser extent, by river mouths and upwelling regions (Prates et al., 2007).

[1]Computational and Systems Biology Laboratory, Oswaldo Cruz Institute, Fiocruz, Brazil.
[2]Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany.
[3]Potsdam University, Institute for Biochemistry and Biology, Potsdam, Germany.
[4]Institut de Ciències del Mar, CSIC, Barcelona, Spain.
[5]Berlin Center for Genomics in Biodiversity Research, Berlin, Germany.

The Arrial do Cabo (Rio de Janeiro state, Brazil) region is affected by upwelling, a phenomena where the up-flow of cold and nutrient-rich waters disturbs ecosystem dynamics, increasing biomass and primary production. This phenomenon can lead to a picoplankton bloom, when the primary producers (including Aerobic anoxygenic phototrophic bacteria) exploit these nutrients coming up from deep cold waters (Alonso-Gutiérrez et al., 2009; Coelho-Souza et al., 2013; Wemheuer et al., 2014).

Aerobic anoxygenic phototrophic bacteria (AAP) require oxygen and reduced organic compounds to grow (Kolber, 2001). In turn, they produce the pigment bacteriochlorophyll *a* (Bchla) and use it to generate additional ATP. Many studies have demonstrated the great metabolic potential of these bacteria, which includes nitrification, carbon dioxide fixation, carotenoids synthesis, and the use of low-molecular-weight organic carbon as energy source (Denner, 2002; Fuchs et al., 2007; Gich, 2006). Therefore, they can inhabit a wide variety of different environments, ranging from terrestrial to aquatic systems, both marine and freshwater, including extreme environments such as Antarctic lakes (Cottrell et al., 2009; Csotonyi et al., 2010; Labrenz et al., 2005). Regardless of the wide distribution of this group in oceans, (Ferrera et al., 2013; Koblízek, 2011) their abundance and importance for carbon fixation and energy cycling is still poorly understood (Goericke, 2002; Koblízek, 2011).

Many studies have been performed in order to estimate AAP abundance and diversity in marine environments, using different approaches, for example, fluorescence detection of bacteriochlorophyll a (BChla) (Cottrell et al., 2006; Kolber et al., 2000; 2001; Lami et al., 2013), and metagenomics (Béjà et al., 2002; Oz et al., 2005; Waidner and Kirchman, 2005; Yutin et al., 2007). In the study performed by Yutin and colleagues (Yutin et al., 2007), the metagenomes from the Global Ocean Sampling Expedition (GOS) (Rusch et al., 2007; Yooseph et al., 2007) were screened for AAPs using specific marker genes revealing a relative AAP abundance of 1%–10% of total bacteria, which are much lower than the values reported by Lami and colleagues (Lami et al., 2013) from the oligotrophic South Pacific Ocean ($\sim$25%).

Reported abundance of AAPs range between <1% and up to 25% (Cottrell and Kirchman, 2009; Hojerová et al., 2011; Lami et al., 2007; Schwalbach and Fuhrman, 2005) and despite initial reports, they support the hypothesis proposed by Kolber (2001) that these organisms would have an advantage in oligotrophic conditions. Recent studies, however, suggest that AAPs thrive better in more eutrophic environments (Cottrell et al., 2010; Hojerová et al., 2011).

Many environmental characteristics such as association to particles, temperature, light attenuation, nutrient limitation, or vulnerability to predation have been proposed as factors that influence AAP abundance, but the ecological role of AAPs is still not well understood. According to Yutin et al. (2007), the AAPs can be classified into 12 phylogroups (from A to L) through *puf*-operon synteny analysis and *pufM* phylogeny.

The primary aim of the present study was to estimate abundance and diversity of AAPs in a metagenome from an upwelling affected coastal bay in the Southwestern Atlantic Ocean (Arraial do Cabo, Brazil). We developed an *in silico* approach based on Profile Hidden Markov Models (pHMM) to screen for two core genes of anoxygenic photosynthesis (*pufM* and *pufL*), distinguishing them from the oxygenic photosynthesis genes (*psbA*–D1 and *psbD*–D2) in addition to analyzing the chlorophyllide reductase subunit X gene (*bch*X). The *puf* and *bch*X genes have been used as AAP markers in many studies (Ferrera et al., 2013, Waidner and Kirchman, 2015; Yutin et al., 2007). The screening results from the Arraial do Cabo metagenomes were compared to those from the GOS datasets. Our analysis also aimed to reveal a deeper insight into AAP abundance and phylogeny in upwelling affected coastal marine waters.

Materials and Methods

Sample collection and filtration

A total of 300 L of surface (< 2 m) was collected (Jan. 24, 2012, 12:41 PM) from the surface of Praia dos Anjos (Angel's Beach), Arraial do Cabo, Rio de Janeiro, Brazil (–22°58′31.33″, –42°0′46.84″).

pH, temperature, and salinity were measured *in situ,* and 1 L was used for determination of Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), as well as concentrations of total N, nitrate, and ammonium. The COD test was performed by the closed reflux method followed by photometric determination, using a COD reactor (Hach Company, Loveland, CO, USA) and visible spectrophotometer (model DR-2500; Hach Company). BOD5, nitrate, and ammonium were determined using the potentiometric method with selective electrodes Orion 081010MD, Orion 9707BNWP, and Orion 9512HPBNWP, respectively (Hach Company). The methodologies used to assess the physicochemical parameters were consistent with those described in the Standard Methods for Examination of Water and Wastewater (APHA, 1998).

The 300 L sample was filtered first through 0.8 μm membranes (mixed cellulose, 47 mm diameter, Milipore) (aiming to hold eukaryotes and particle-associated prokaryotes—named Sample E) and then through 0.22 μm membranes (mixed cellulose, 47 mm diameter, Milipore) (aiming to hold free living prokaryotes—named sample P) using a vacuum filtration system. Therefore, samples P and E refer to free-living and particle-associated bacteria, respectively. The samples were collected during summer, in the upwelling season. The total filtration time through the membranes used prior to extraction was 4 hours, and the samples were kept in an ice bath during the process.

DNA extraction and quantification

DNA was extracted from membranes using the Meta-G-Nome™ DNA Isolation Kit (Epicentre). In order to obtain 20 μg of DNA, a total of 20 membranes of each sample were pooled (representing the filtrate from approximately 40 liters of water). This large amount of DNA is necessary to avoid potential methodological biases. The extracted DNA samples were verified by agarose gel (1%) electrophoresis (100 V), and quantified using ImageJ software, NanoDrop (Thermo Scientific), and Qubit (Life Technologies).

DNA pyrosequencing and sequence pre-processing

A total of 2 μg DNA from each sample (P and E) was sent to LNCC (Laboratório Nacional de Computação Científica, Petrópolis, Rio de Janeiro, Brazil) for pyrosequencing on a 454 (ROCHE) GS FLX+ sequencer. One 454 plate was used, and DNA of each sample constituted half of the plate.

The SFF files generated were analyzed on Stingray (stingray.biowebdb.org) (Wagner et al., 2014) to generate the clipped FASTA and QUAL files. The CD-HIT-454 (Niu

et al., 2010) program was used to remove artificial duplicates (artifacts) using default parameters, and LUCY v1.20 (Chou et al., 2001) (default parameters) was applied to remove low quality and small sequences (<20 Phred score, <100 base pairs [bp]).

Metagenomic datasets and sequence pre-processing

The two samples from Arraial do Cabo (samples P and E) seawater were used in addition to all 82 samples from the GOS dataset. The samples from the GOS dataset were collected around the world in various environments ranging from inside lagoons to open ocean regions (Rusch et al., 2007; Yooseph et al., 2007). The samples of this dataset were filtered in a 0.8 μm and subsequently in 0.1 μm filters, similar filters used in samples from Arraial do Cabo. Two datasets were generated: (i) All reads were translated into six frames using the TRANSEQ (from EMBOSS 6.1.0 package, default parameters) (Rice et al., 2000); (ii) The reads from Arraial do Cabo and from the 10 GOS samples (with the highest AAP abundance) were individually assembled using CAP3 (default parameters) (Huang and Madan, 1999) and the Open Read Frames (ORFs) were extracted from the contigs and singlets using the METAGENMARK (version 2.8, default parameters) (Zhu et al., 2010).

Estimates of AAP abundance in metagenomes by screening for pufM, pufL, *and* bchX *gene frequencies*

The sequences from orthologs groups of AAP genes (*pufM*, *puf*L, and *bch*X); the homologous of these genes (from oxygenic photosynthesis: *psbA, psbD, bchL*, *bchY*, *bchZ,* and *bchN*), and the constitutive gene *rec*A were obtained (both nucleotide and amino acid sequences in fasta format) from KEGG Orthology (KO): K08929 (*pufM*), K08928 (*pufL*), K11333 (*bchX*), K03553 (*recA*), K02703 (*psbA*–D1 protein), K02706 (*psbD*–D2 protein), K04037 (*bchL*), K11334 (*bchY*), K11335 (*bchZ*), and K04038 (*bchN*). These groups were aligned with MAFFT (v7.029b) (Katoh et al., 2002) and each alignment was converted to Stockholm format using a custom PERL script.

The HMMBUILD (from HMMER 3.0 package) program (Eddy, 2011) was used to build a pHMM from each alignment and each pHMM was used (using the HMMSEARCH from HMMER 3.0 package, e-value cutoff 0.1) to search the metagenomic datasets (translated reads and ORFs). The hits were extracted by using the FASTACMD program (from BLAST 2.2.21 package (Altschul et al., 1990)) and obtained sequences were used as input for HMMSCAN (from HMMER 3.0, e-value cutoff 0.1) and compared against all pHMM (concatenated and submitted to the HMMPRESS). Using the oxygenic photosynthesis homologs, it is possible to avoid the false classification of the environmental sequences, excluding the hits more similar to these profiles than to the anoxygenic gene profiles.

The number of ''read equivalents'' (number of reads used to assemble each ORF) of each environmental HMMER hit was calculated using a slightly modified approach based on Yutin and colleagues (2007) and a script developed in RUBY 1.9.3 and BIORUBY (Goto et al., 2010).

In order to estimate the frequency of each marker gene, the number of their reads (or ''read equivalents'' in ORFs analysis) was normalized dividing them by the number of reads of the housekeeping gene *rec*A (coding a critical DNA repair enzyme). This gene represents a single-copy gene in the genome of all bacteria (as the *puf* genes in the AAP genomes), it has the same mean size as the AAP marker genes and thus can be used to estimate the number of bacterial genomes present in the analyzed metagenomic samples (Howard et al., 2006; Venter et al., 2004; Yutin et al., 2007). The percent fraction of the AAP marker gene was calculated as follows:

$$\text{Percent fraction of the AAP maker gene} = (\text{number of reads or ''read equivalent'' from the marker gene} / \text{number of reads of recA}) * 100$$

Additionally, the mean abundance of the three genes (*pufM*, *pufL*, and *bchX*) was calculated to estimate AAP relative abundance in each analyzed sample, as well as standard deviation.

Confirming sequence annotation, calculating its specificity and sensitivity in our approach

In order to confirm the annotation of the environmental sequences (ORFs), obtained by our newly developed approach, the program BLASTX (Altschul et al., 1990) and the RefSeq database (release 61) (from NCBI) were used. The best hits were manually verified and the percentage of false positives was calculated for each gene. The specificity for each gene was inferred by the mean of false positives:

$$\text{Specificity } (\%) = 100 - \text{mean of percentage from false positives for all marker genes}$$

The sensitivity was estimated running the pipeline against the KEGG Orthology (KO) reference sequences and calculating the percentage of sequences obtained from each pHMM.

Phylogenetic analysis of pufM genes

The environmental *pufM* ORFs with more than 700 nucleotides (nt) were extracted and concatenated with reference sequences (from KO and from NCBI). The software MEGA 5.1 (Tamura et al., 2006) was used to (i) translate the nucleotide sequences to amino acids; (ii) align the amino acid sequences using the MUSCLE (Edgar, 2004) program (default parameters) and reverse translation of the alignment to nucleotide; and (iii) calculate the best evolutionary and substitution model for sequence alignment and subsequent phylogenetic analysis.

The obtained alignment was exported as FASTA format and trimmed using TRIMAL 1.2 (Capella-Gutierrez et al., 2009) to remove columns or alignment positions with more gaps than nucleotides before conversion of the final alignment to the NEXUS format.

The program Mr Bayes 3.2 (Ronquist and Huelsenbeck, 2003) was used on CIPRES GATEWAY (http://www.phylo.org/portal2/) (Miller et al., 2010) together with the Generalized Time-Reversible (GTR) model and gamma distribution, to generate a phylogenetic tree using Bayesian analysis. A total of two analyses were carried out with four parallel chains and 10 million executions.

Later, the phylogenetic tree and the alignment were imported into the ARB 5.5 (Ludwig, 2004) program to generate

Table 1. Number of Hits Obtained for Each AAP Marker Gene and *rec*A Gene in Arraial do Cabo Samples

Samples	pufM	pufL	bchX	recA
Sample P	106	117	136	501
Sample E	24	12	15	197

a local ARB database. Environmental ORFs smaller than 700 nucleotides were added to the custom database using the quick add tool of ARB to construct the phylogenetic tree using the parsimony method.

Results

Sampling characteristics

At the time of sampling, the temperature of the water was 26°C, the pH was 7.5, and the salinity was 33%. The physicochemical analysis of the sample showed a BOD of 1 mg/L, COD of 60 mg/L, and concentrations of nitrate, ammonium, and total nitrogen of 0.9 mg/L (or $\sim$15.51 μM), 0 mg/L ($<$ 0.5 μM), and 0.4 mg/L (or $\sim$28.57 μM), respectively. The sample site was affected by upwelling as demonstrated by measurements of Albuquerque et al. (2014).

Metagenomic datasets and sequence pre-processing

The total number of sequence reads from Arraial do Cabo was 1,064,888 (595,534 for sample P [free-living] and 469,354 for sample E [particle-associated]) and 12,672,518 for the GOS (Rusch et al., 2007) dataset.

All the read sequences were submitted to MG-RAST (https://metagenomics.anl.gov/) under ID: 4539290.3 for sample P and 4539291.3 for sample E.

By assembling the reads using the CAP3 program, a total of 29,074 contigs and 269,587 singlets from sample P were generated. From sample E, 20,792 contigs and 396,371 singlets were generated. Using the METAGENMARK (Zhu et al., 2010) program, the total number of ORFs obtained was 409,111 for sample P and 451,722 for sample E. Supplementary Table S1 (supplementary material is available online at www.liebertpub.com/omi) shows the total number of sequences (reads and ORFs) obtained from our GOS dataset analysis.

AAP abundance in metagenomes estimated by mean of ratio of pufM, pufL, *and* bchX *gene in relation to housekeeping* recA *gene*

The screening of the environmental reads revealed a total of 860 and 248 hits obtained from sample P and E, respectively. Table 1 shows the number of hits for each gene screened in the Arraial do Cabo samples.

The number of hits obtained in the 82 samples from the GOS dataset is given in Supplementary Table S2. Additionally, AAP abundance was calculated using the mean of the ratio of each marker gene (*pufM*, *pufL,* and *bchX*) in relation to the housekeeping *recA* gene. Highest abundance of AAPs was found in sample P (23.88% $\pm$ 3.02 of free-living bacteria).

Figure 1 shows the percentage of AAPs in the 10 GOS samples that had the highest AAP % fraction in addition to our two samples from Arraial do Cabo (Sample P [free-living bacteria] and E [particle-associated bacteria]).

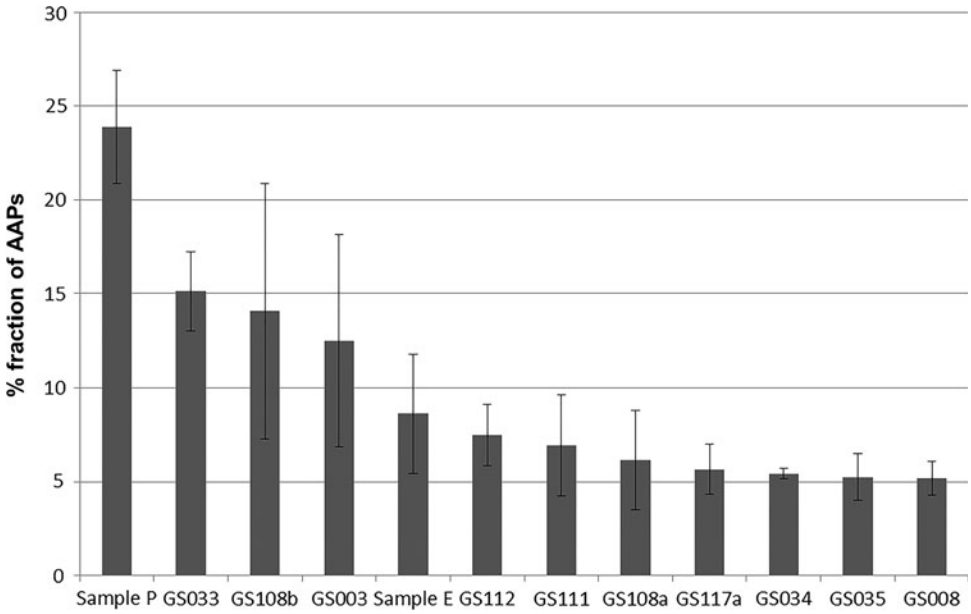

FIG. 1. Percent fraction of AAPs in ten samples of the GOS dataset with the highest AAP frequencies and our two samples from Arraial do Cabo (unassembled samples). Sample P, Arraial do Cabo; GS033, Punta Cormorant, Floreana Island (Hypersaline Lagoon); GS108b, Coccos Keeling, Inside Lagoon (>0.8 μm fraction); GS003, Browns Bank, Gulf of Maine; Sample E, Arraial do Cabo (>0.8 μm fraction); GS112, Indian Ocean; GS111, Indian Ocean; GS108a,Coccos Keeling, Inside Lagoon; GS117a, St. Anne Island, Seychelles; GS034, North Seamore Island (Galapagos); GS035, Wolf Island (Galapagos); and GS008, Newport Harbor, RI.

TABLE 2. NUMBER OF PUTATIVE *pufM*, *pufL*, *bchX*, AND *recA* ORFS OBTAINED FOR EACH ANALYZED ENVIRONMENTAL SAMPLE

Sample	pufM	pufL	bchX	recA
GS111—Indian Ocean	3	2	5	79
GS008—Coccos Keeling, Inside Lagoon	7	5	6	137
GS112—Indian ocean (454 FLX)	32	21	15	666
GS108b—Coccos Keeling, Inside Lagoon (>0.8 μm)	6	5	3	42
GS035—Wolf Islands	6	3	8	174
GS003—Browns Bank, Gulf of Maine	2	5	7	61
GS108a—Coccos Keeling, Inside Lagoon (>0.1 <0.8 μm)	5	2	1	64
GS117a—St. Anne Island, Seychelles	16	12	9	227
GS033—Punta Cormorant, Floreana Island (Hypersaline Lagoon)	61	51	82	294
GS034—North Seamore Island	9	8	6	178
Arraial do Cabo sample P	33	33	39	219
Arraial do Cabo sample E (>0.8 μm)	15	8	19	148
Total	195	155	200	2289

After sequence assembly and ORF extraction from these 10 GOS samples (and our two samples from Arraial do Cabo), a total of 195, 155, and 200 putative *puf*M, *puf*L, and *bch*X ORFs were found, respectively. Table 2 gives the number of ORFs for each gene obtained from the analyzed environmental samples.

The % fraction of AAPs in ORFs was calculated using an approach adapted from a study of Yutin and colleagues (Yutin et al., 2007), calculating the ''read equivalents'' of each gene in the detected ORFs.

As for unassembled reads, in the ORFs, the highest % fraction of AAPs was found in sample P from Arraial do Cabo (22.03% ± 3.6). Figure 2 shows the % fraction of AAPs obtained from all 12 samples used for our analysis. Figure 3 gives the worldwide distribution of the analyzed GOS samples and our two samples from Arraial do Cabo.

Sequence annotation, specificity, and sensitivity of our approach

In order to confirm the function of the ORFs obtained, all sequences were queried against the RefSeq protein database using BLASTX. All hits were manually checked.

Of all the ORFs obtained with the *pufM* gene pHMM, only four sequences showed a similarity with other genes, and three showed no hits to the RefSeq database. Sequences with

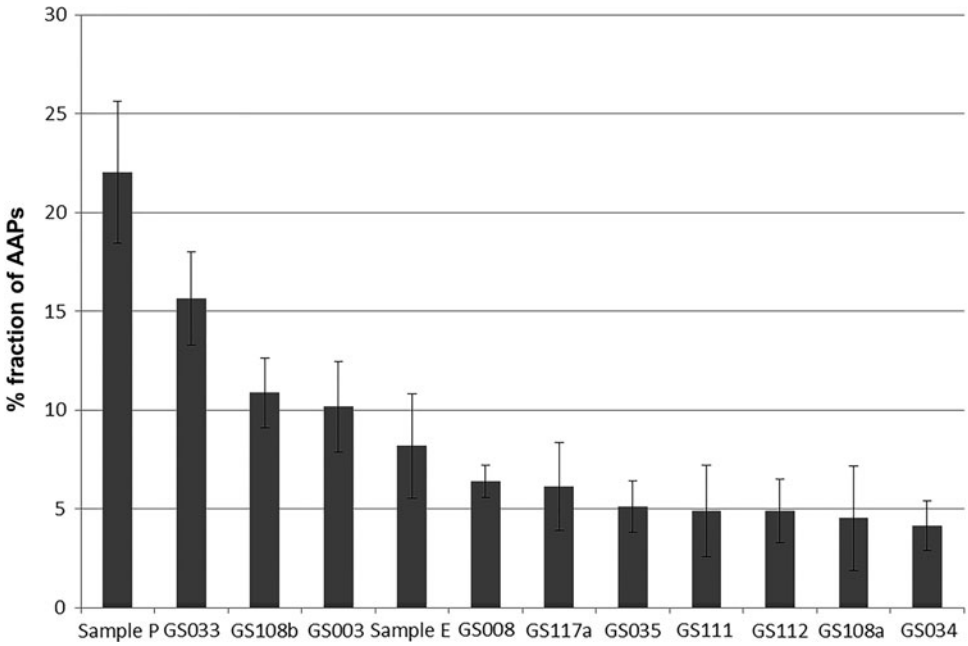

FIG. 2. Percent fraction of AAPs in ORFs (calculated from ''reads equivalents'' of each ORF) from the selected ten samples of the GOS dataset and our two samples from Arraial do Cabo. Sample P (free-living), Arraial do Cabo; GS033, Punta Cormorant, Floreana Island (Hypersaline Lagoon); GS108b, Coccos Keeling, Inside Lagoon (0.8 μm fraction); GS003, Browns Bank, Gulf of Maine; Sample E (particle-associated), Arraial do Cabo (0.8 μm fraction); GS008, Newport Harbor, RI; GS117a, St. Anne Island, Seychelles; GS035, Wolf Island (Galapagos); GS112, Indian Ocean; GS111, Indian Ocean; GS108a, Coccos Keeling, Inside Lagoon; and GS034, North Seamore Island (Galapagos).

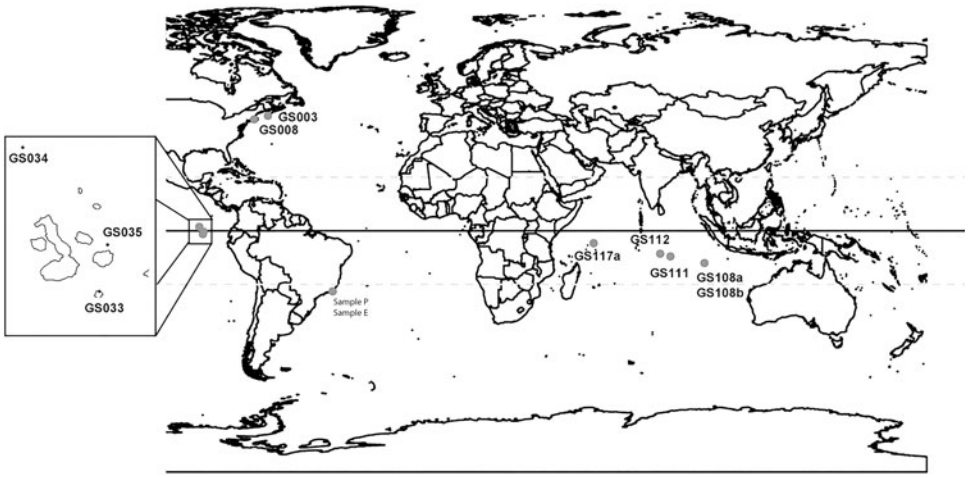

FIG. 3. Worldwide distribution of sample sites of the ten GOS datasets with the highest AAP % fraction and the samples from Arraial do Cabo.

no hit were compared against GenBank using BLASTX, whereby two of the three sequences with no hit in RefSeq, showed a similarity to environmental *pufM* genes. Therefore, a total of five sequences (four similar to different known genes and one without any hit) were classified as possible false positives (5/195—2.56%).

Analyzing all ORFs obtained from *pufL* gene pHMM, only 1.29% (2/155) of sequences showed hits with different genes in the RefSeq and only one (1/155) sequence showed no hit. The sequence with no hit was further compared against GenBank using BLASTX and then revealed hits with hypothetical proteins. Thus, a total of 1.93% (3/155) sequences were potential false positives and consequently removed from our phylogenetic analysis.

From all sequences obtained with *bchX* pHMM, 8.5% (17/200) of the sequences revealed hits with other genes, and only two showed no hit with known sequences when using both databases (RefSeq and NCBI). Thus, a total of 9.5% (19/200) of possible false positive sequences were found.

Finally, the specificity of our approach was calculated (mean of the three genes false positives (4.66%)) and subtracting it from 100, which then resulted in a mean specificity of 95.34%.

Phylogenetic analysis of pufM *genes*

In order to classify our environmental *pufM* sequences, a Bayesian analysis (using Mr Bayes) was performed using the ORFs from our ten analyzed GOS samples and our two samples from Arraial do Cabo, in addition to the 38 retrieved reference sequences from KEGG (KO) and NCBI. Figure 4 shows the phylogenetic tree.

From the six environmental sequences obtained from Arraial do Cabo in our calculated tree, five (83.33%) were assigned to the *Roseobacter* clade (phylo-group G). The unique sequence is affiliated to phylo-group K.

Additionally, all short sequences were added to the Bayesian tree, using the ARB program (''quick add by parsimony'' method) (Ludwig, 2004), in order to relate them to the retrieved phylotypes. Figure 5 shows the relative abundance of each phylo-group in all 12 analyzed metagenome samples.

When adding all short sequences to the tree, only one sequence from sample P (free-living) could be classified as phylo-group A (without representative cultivated organism), two sequences of sample P and one of sample E (particle-associated) were grouped into phylo-group F (*Rhodobacter* clade), 22 sequences of sample P and 10 of sample E were classified as phylo-group G (*Roseobacter* clade), four sequences of sample P and four of sample E were assigned to phylo-group H (uncultured), and just three sequences of sample P were assigned to phylo-group K (gamma-proteobacterial clade).

Discussion

Using metagenomics and bioinformatics approaches, it was possible to infer the abundance of AAPs in an upwelling affected environment, and to compare that with a large public dataset (GOS). The results show that Arraial do Cabo, affected by upwelling phenomena, has a very high abundance of AAPs, outnumbering the 82 samples from GOS dataset (Fig. 1). These results reinforce the hypothesis that after the upwelling and subsequent phytoplankton bloom, the AAPs can grow using the nutrients from the deep cold water and the abundant light present in the surface water, even in oligotrophic tropical and subtropical environments. Other studies also point to a tight linkage between phytoplankton bloom development and abundance of AAPs (Alonso-Gutiérrez et al., 2009; Wemheuer et al., 2014), which may be either related to the inorganic nutrient input or the direct coupling between phytoplankton and AAPs (e.g., via phytoplankton exudates).

Judging from the raw and assembled reads, sample P (free-living bacteria) from Arraial do Cabo had the highest % fraction of AAPs (up to $23.88\% \pm 3.02\%$). This result shows that Arraial do Cabo can be regarded as a marine environment with one of the highest so far known AAP abundance worldwide. Although the GOS and Arraial do Cabo samples were fractionated, we needed to calculate the mean of the size-fractionated samples in order to compare our results with samples from other studies. We found that the total abundance of AAPs in the samples of Arraial do Cabo was

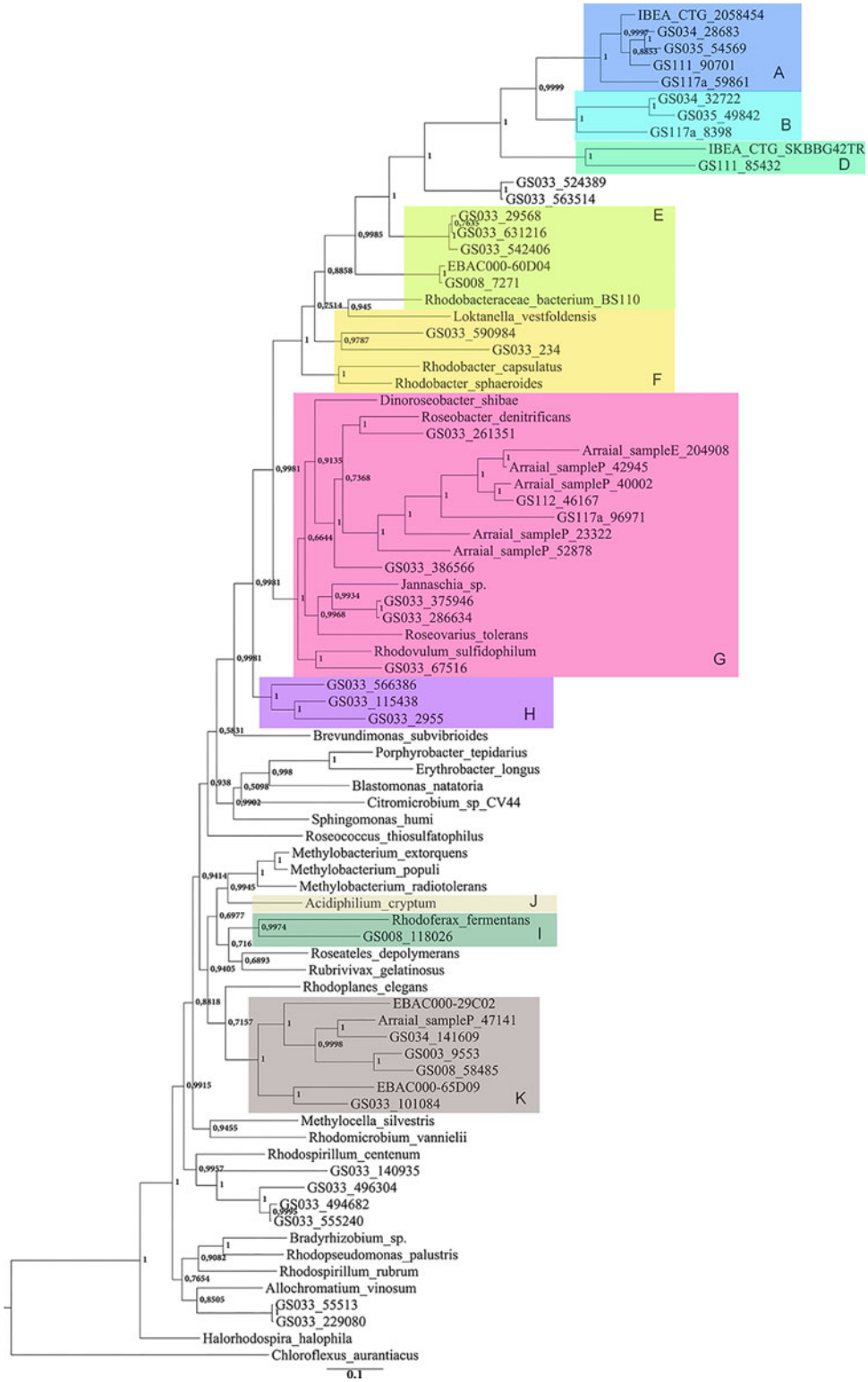

FIG. 4. Phylogenetic tree of *pufM* genes from all ten GOS samples, our two Arraial do Cabo samples, and all reference sequences retrieved from NCBI and KEEG (KO). Only sequences with more than 700 nucleotides were used. The tree was obtained by Bayesian analysis on Mr Bayes 3.2, using the GTR model and gamma distribution. Two executions were carried out with four parallel chains and 10 millions of executions. The *highlighted clades* refer to the different AAP phylo-groups defined by Yutin et al., (2007).

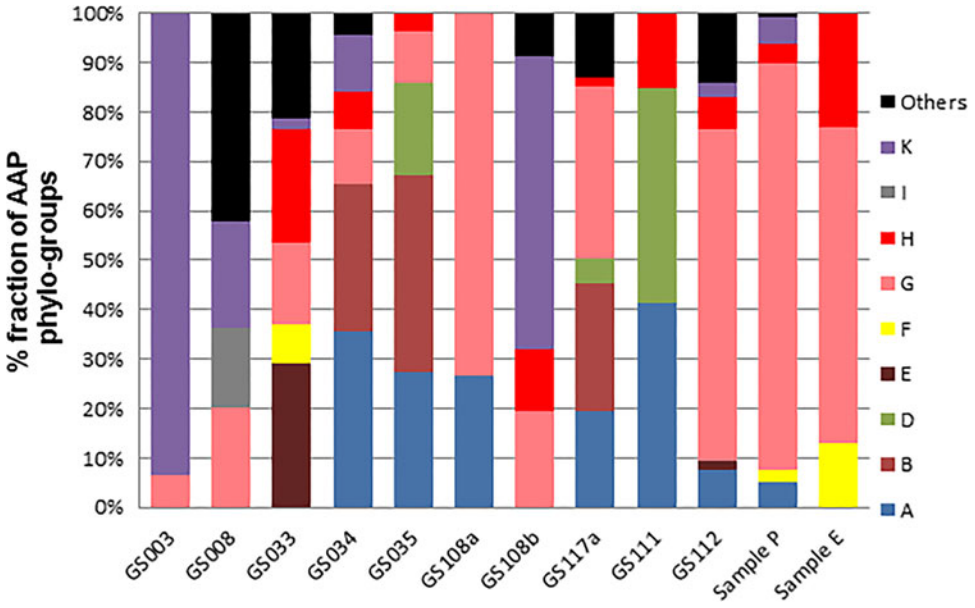

FIG. 5. Relative abundance of each phylo-group retrieved from the different analyzed metagenomic samples. Number of read equivalents for each obtained ORF was counted and percentages were calculated by using the classification of the phylogenetic tree generated by ARB.

16.07%, which is at the higher end of other oceanic studies (Cottrell and Kirchman, 2009; Ritchie and Johnson, 2012).

The relatively high AAP abundance in Arraial do Cabo is comparable to the results of Waidner and Kirchman (2007) from turbid estuary waters (ranging from 12% to 17% of total bacteria), but higher than those of Cottrell and Kirchman (2009) from temperate and polar Artic Ocean (ranging from 5% to 8%), and Ritchie and Johnson (2012) from coastal regions of the Pacific Ocean (1.2%, on average).

The sample with the second highest AAP abundance was GS33 (Browns Bank, Gulf of Maine), an anoxic hyper saline lagoon (63.4 PSU, dissolved oxygen, 0.06 mg/L) with 15.64% ± 2.36 (Rusch et al., 2007). This sample was discussed separately by Yutin et al. (2007), because it is likely that due to the anoxic environment and the detected anoxygenic phototrophic bacteria are anaerobic photoautotrophs and not AAPs. However, our phylogenetic results reveal many *pufM* sequences clustering within the phylogroup G (16.49% of total read equivalents), suggesting the presence of an active AAP community in this environment too.

It is important to note that the comparisons between different samples may be biased by differences in methods used for sampling, filtration, DNA extraction, and sequencing. In addition, the collection timing should be considered since many aquatic systems are characterized by seasonal variances in their AAP community (Cottrell and Kirchman, 2009; Ferrera et al., 2013). Furthermore, Ferrera et al. (2013) showed a high AAP abundance in summer but a low richness compared to the winter situation, corroborating many previous studies (Lamy et al., 2011; Masín et al., 2006; Zhang and Jiao, 2007).

Other studies have revealed many different environmental variables such as light, nutrient availability, temperature, vulnerability to predation, and Chl *a* concentration influencing AAP abundance and diversity (Hojerová et al, 2011; Lamy et al., 2011; Zhang and Jiao, 2007). In addition, Ferrera et al. (2013) showed a tight correlation between day length and AAP abundance, corroborating data from AAP culture studies suggesting that light enhances organic carbon utilization efficiency, energy cycling, and hence growth (Hauruseu and Koblizek, 2012; Spring and Riedel, 2013).

The samples from Arraial do Cabo were collected in summer around noon when light irradiation was highest. Future seasonal studies should be performed to better understand the variation of AAP abundance in the context of light availability because this may be an important factor to explain the extraordinary high abundance of AAP in our study.

In addition, it was showed that 80% of the most abundant samples from GOS were collected close to the Equator (Fig. 3). These results can be explained by the fact that all tropical sites are characterized by a high light availability and hence great water transparency allowing for light harvesting even at greater depth (>100 m) and consequently positive AAP growth throughout an extended part of the water column as compared to other marine sites (Hauruseu and Koblizek, 2012).

Our pipeline enabled us to screen a total of 12,672,518 reads rapidly from 82 GOS samples and 1,064,888 reads from our two Arraial do Cabo metagenomes. The pipeline was very sensitive and highly specific (95.34% of specificity and 100% of sensitivity). Even with such a large-scale study including many metagenomic samples, it proved possible to run it on a simple desktop computer. The abundance of AAPs was evaluated on unassembled data (raw reads) allowing for determination of environments with the highest AAP fraction for targeted assembly selection, ORF extraction, and AAP screening (using a similar approach of the unassembled screening).

To our knowledge, this is the first study estimating AAP abundance on unassembled metagenomic samples, since the

study of Yutin et al. (2007) was performed for assembled samples. Moreover, Yutin and colleagues used the cross assembly (contigs were obtained from all concatenated samples), which significantly increases the likelihood to generate chimeric sequences (Wommack et al., 2008). Major advantages of using raw reads are (i) preventing the assembly step for all samples, which is slow because it is computationally expensive, and (ii) avoiding chimeric sequences obtained by metagenomic assembly (Pignatelli and Moya, 2011). However, due to the limitation of read size, the assembly step is required for both the phylogenetic and *puf* operon synteny analyses.

The present study uses the newest version of the GOS project (Rusch et al., 2007), meaning an additional 38 new samples were analyzed and compared to the original study by Yutin (that used only 44 samples) (Yutin et al., 2007) totaling 82 GOS samples, plus our two samples from Arraial do Cabo. The main reason to study AAPs in the upwelling affected Arraial do Cabo samples is the high abundance of the genus *Roseobacter,* which accounts for 15% of identified genera on sample P (free-living bacteria). In addition, a number of other known AAP genera (e.g., *Jannaschia* and *Dinoroseobacter*) were found in our previous exploratory work to characterize these samples (unpublished data).

To minimize the chimeric sequence formation, all samples were assembled individually (our two samples from Arraial do Cabo and the 10 GOS samples with the highest number of AAP reads), using the CAP3 program with specific default parameters (Huang and Madan, 1999). Results of the AAP screening of unassembled samples were compared to those of assembled samples, showing a good consistency between the obtained results. Another advantage of the individual sample assembly is the possibility for further phylogeographic sequence analysis.

Our AAP markers were the *pufM*, *pufL,* and *bchX* genes, those have been used in many previous studies in PCR, qPCR (mainly *pufM*) (Waidner and Kirchman, 2008) or *in silico* analyses (Béjà et al., 2002; Oz et al., 2005; Waidner and Kirchman, 2005; Yutin et al., 2007). The *pufM* environmental sequences obtained in all of our 12 samples (>700 nucleotides) were used together with reference sequences in a Bayesian analysis (Fig. 4). Small sequences were added by the ARB parsimony method to the resulting phylogenetic tree. The topology of the resulting tree corroborates previous studies [e.g., Yutin et al. (2007) and Lehours et al. (2010)], and as expected the reference sequences used clustered into specific phylo-groups. Further, these results were confirmed by analyzing the distribution and phylogenetic relatedness of the *puf* operon, as discussed by Yutin et al. (2007).

Our results show the predominance of phylo-group G (*Roseobacter* clade) in both Arraial do Cabo samples, with 82.36% (sample P [free-living bacteria]) and 64.05% (sample E [particle-associated bacteria]) of total AAP in this environment. These results suggest a possible link between the phytoplankton bloom induced by upwelling and the high abundance of phylo-group G, as has been earlier shown by González et al., (2000) and Suzuki et al., (2001).

In addition, the study by Ferrera et al. (2013) in the coastal Mediterranean showed that the alpha-proteobacterial groups E, F, and G only outnumber the gamma-proteobacterial groups in the high-nutrient season, reinforcing the correlation of these groups with high nutrient concentration and phytoplankton bloom development.

In our samples, the *Roseobacter* clade was the most widespread, present in 11 of the 12 samples analyzed, corroborating the results of other studies (Buchan et al., 2005), including the previous GOS study by Yutin et al. (2007). However, in the GOS samples analyzed, we detected a higher abundance of the phylo-group G (*Roseobacter* clade) in samples from the Indian Ocean (and less in the GS111 sample) when comparing them with samples from the Eastern Pacific Ocean (Galapagos) or the Atlantic West Coast (USA).

AAP lifestyles and phylo-groups

The relative abundance of free-living AAPs (samples P and GS108a) was higher than particle-associated AAPs (samples E and GS108b), suggesting that the phylo-group G refers mainly to organisms with a free-living lifestyle.

In addition, phylo-group A may also comprise organisms with a predominantly free-living lifestyle since this group is absent in samples E and GS108b (>0.8 μm). Interestingly, this group is also absent in the anoxic sample (GS033), suggesting a dependency of this group on oxygen availability.

In contrast, phylo-group H is more abundant in the >0.8 μm size fraction (samples E and GS108b) and in the anoxic GS033 sample, whereas phylo-group E was exclusively found in the anoxic GS033 sample and at very low abundance in the GS112 sample (0.1 μm, Indian Ocean). The phylo-group F (*Rhodobacter* clade) was found in both size fractions of Arraial do Cabo (although with a higher abundance in sample E), but in the GOS samples it was exclusively present in the anoxic GS033 sample.

The correlation between AAP abundance of the assigned groups in the anoxic GS033 sample, but also in the >0.8 μm fraction of samples E and GS108b can be explained by the formation of potentially anoxic microenvironments (e.g., on macroscopic organic aggregates even in an oxygenated water column). Such aggregates are normally trapped on the 0.8 μm membranes. The specific AAP groups are abundant in these samples and seem to be well adapted to harvest light on the organic matter-rich particles, which also provide an excellent organic substrate for these photoheterotrophic bacteria (Cottrell et al., 2010; Mašin et al., 2012; Salka et al., 2011).

In the study performed by Yutin and colleagues (2007), *Rhodoplanes* (alfa-proteobacteria) and *Rosealetes* (beta-proteobacteria) genera clustered together. However, in our work, the *Rosealetes* clustered with other beta-proteobacteria genera: *Rubrivivax* and *Rhodoferax*, separating them from alpha-proteobacteria. In our study, just a single environmental sequence was affiliated to Rhodoferax clade (clade I) (GS008_118026).

Alpha-proteobacteria of the genus *Rhodoplanes* clustered together with the phylo-group K (gamma-proteobacteria). This fact may be explained by a possible horizontal gene transfer (HGT) of photosynthetic apparatus, as proposed by several previous studies (Alonso-Gutiérrez et al., 2009; Cottrell and Kirchman, 2009; Ferrera et al., 2013). Some AAP strains (e.g., *Roseobacter litoralis* Och 149) contain a plasmid with all genes from the anoxygenic photosynthesis (Kalhoefer et al., 2011; Petersen et al., 2009) and the presence of phage DNA that is directly associated with the photosynthesis operons (Jiao et al., 2010; Yurkov et al., 2013). This may corroborate the hypothesis of HGT among several AAPs.

It is noteworthy that, similarly to the study of Yutin et al. (2007), no α-4 subclass AAP was detected in our extensive phylogenetic analysis. This group, normally present in diverse marine and fresh water environments (Yurkov and Csotonyi, 2009), forms a separate clade (*Erythrobacter*, *Blastomonas, Sphingomonas,* and *Porphyrobacter*), without any known environmental sequence.

Conclusions

This study presents an approach for the fast screening of specific markers such as anoxygenic photosynthesis genes and to evaluate their abundance and diversity in environmental samples (raw and/or assembled reads).

Our results obtained from 84 unassembled and 12 assembled metagenome samples reveal that our newly developed approach leads to consistent and reliable results for both types of datasets. When using the unassembled samples, it was possible to screen relatively large datasets and to select samples with the highest AAP abundance for further in depth phylogenetic analysis. Free-living bacteria (sample P) from Arraial do Cabo showed an extremely high AAP abundance, which was higher than of any other GOS samples analyzed.

The phylogenetic analysis of *pufM* ORFs enabled us to classify specific phylo-groups of AAPs present in these environments, showing that the *Roseobacter* clade (phylogroup G) is the predominant AAP group in the Arraial do Cabo environment (as expected for an upwelling affected site) and the most ubiquitous AAP group of all 12 assembled metagenome samples.

These promising results encourage us to perform a larger and more detailed longitudinal study in the Arraial do Cabo region in the near future, to investigate the dynamics of AAP at different locations and seasons, and to better understand the ecological role of these unique bacteria for biogeochemical and energy cycling in the ocean.

Acknowledgments

We thank the sequencing group of LNCC for performing the DNA pyrosequencing. Dr. Adriana M. Froes, Dr. Aline Dumaresq, Dr. Gisele Lopes Nunes, Fábio Bernardo da Silva, and Bruno Manoel da Silva are acknowledged for their help in collecting and pre-processing the samples; Dr. Kevin Tyler (UEA) for valuable help with the MS English; Dr. Yara Traub-Cseko for her willingness and kindness allowing us to use the Molecular Biology of Parasites and Vectors Lab's facilities. This study was supported by the Science without Borders Program (Ciência Sem Fronteiras), CNPq, FIOCRUZ, and CAPES.

Author Disclosure Statement

Conceived and designed the experiments: RRCC, IF, HPG, and AMRD. Performed the experiments: RRCC. Analyzed the data: RRCC, IF, and HPG. Contributed reagents/materials/analysis tools: RRC and AMRD. All authors wrote the manuscript and revised it for significant intellectual content.

References

Albuquerque ALS, Belem AL, Zuluaga FJB, et al. (2014). Particle fluxes and bulk geochemical characterization of the Cabo Frio upwelling system in Southeastern Brazil: Sediment trap experiments between Spring 2010 and Summer 2012. An Acad Bras Cienc 86, no. 2.

Alonso-Gutiérrez J, Lekunberri I, Teira E, Gasol JM, Figueras A, and Novoa B. (2009). Bacterioplankton composition of the coastal upwelling system of ''Ría de Vigo'', NW Spain. FEMS Microbiol Ecol 70, 493–505.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Béjà O, Suzuki MT, Heidelberg JF, et al. (2002). Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature 415, 630–633.

Buchan A, Gonzalez JM, and Moran MA. (2005). Overview of the marine Roseobacter lineage. Appl Environ Microbiol 71, 5665–5677.

Capella-Gutierrez S, Silla-Martinez JM, and Gabaldon T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Chou HH, and Holmes MH. (2001). DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093–1104.

Coelho-Souza SA, Pereira GC, Coutinho R, and Guimarães JR. (2013). Yearly variation of bacterial production in the Arraial do Cabo protection area (Cabo Frio upwelling region): An evidence of anthropogenic pressure. Braz J Microbiol 44, 1349–1357.

Cottrell MT, and Kirchman DL. (2009). Photoheterotrophic microbes in the Arctic Ocean in summer and winter. Appl Environ Microbiol 75, 4958–4966.

Cottrell MT, Mannino A, and Kirchman DL. (2006). Aerobic anoxygenic phototrophic bacteria in the Mid-Atlantic Bight and the North Pacific Gyre. Appl Environ Microbiol 72, 557–564.

Cottrell MT, Ras J, and Kirchman DL. (2010). Bacteriochlorophyll and community structure of aerobic anoxygenic phototrophic bacteria in a particle-rich estuary. ISME Journal 4, 945–954.

Csotonyi JT, Swiderski J, Stackebrandt E, and Yurkov V. (2010). A new extreme environment for aerobic anoxygenic phototrophs: Biological soil crusts. Adv Exp Med Biol 675, 3–14.

Denner E. (2002). Erythrobacter citreus sp. nov., a yellow-pigmented bacterium that lacks bacteriochlorophyll a, isolated from the western Mediterranean Sea. Int J Syst Evol Microbiol 52, 1655–1661.

Eddy SR. (2011). Accelerated profile HMM searches. PLoS Comput Biol **7**, e1002195.

Edgar RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792–1797.

Ferrera I, Borrego CM, Salazar G, and Gasol JM. (2013). Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of *pufM* gene: Marine AAP dynamics in coastal sea. Environ Microbiol 16, 2953–2965.

Fuchs BM, Spring S, Teeling H, et al. (2007). Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. Proc Natl Acad Sci 104, 2891–2896.

Ghai R, Hernandez CM, Picazo A, et al. (2012). Metagenomes of Mediterranean coastal lagoons. Sci Rep 2, 490.

Ghai R, Rodríguez-Valera F, McMahon KD, et al. (2011). Metagenomics of the water column in the pristine upper course of the Amazon River. PLoS ONE 6, e23785.

Gich F. (2006). Sandarakinorhabdus limnophila gen. nov., sp. nov., a novel bacteriochlorophyll a-containing, obligately aerobic bacterium isolated from freshwater lakes. Int J Syst Evol Microbiol 56, 847–854.

Goericke R. (2002). Bacteriochlorophyll a in the ocean: Is anoxygenic bacterial photosynthesis important? Limnol Oceanogr 47, 290–295.

Gonzalez JM, Simo R, Massana R, Covert JS, Casamayor EO, Pedros-Alio C, and Moran MA. (2000). Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Appl Environ Microbiol 66, 4237–4246.

Goto N, Prins P, Nakao M, Bonnal R, Aerts J, and Katayama T. (2010). BioRuby: Bioinformatics software for the Ruby programming language. Bioinformatics 26, 2617–2619.

Hauruseu D, and Koblizek M. (2012). Influence of light on carbon utilization in aerobic anoxygenic phototrophs. Appl Environ Microbiol 78, 7414–7419.

Hojerová E, Mašín M, Brunet C, Ferrera I, Gasol JM, and Koblížek M. (2011). Distribution and growth of aerobic anoxygenic phototrophs in the Mediterranean Sea: AAP bacteria in the Mediterranean Sea. Environ Microbiol 13, 2717–2725.

Howard EC, Henriksen JR, Buchan A, et al. (2006). Bacterial taxa that limit sulfur flux from the ocean. Science 314, 649–652.

Huang X, and Madan A. (1999). CAP3: A DNA sequence assembly program. Genome Res 9, 868–877.

Jiao N, Zhang R, and Zheng Q. (2010). Coexistence of two different photosynthetic operons in Citromicrobium bathyomarinum JL354 as revealed by whole-genome sequencing. J Bacteriol 192,1169–1170.

Kalhoefer D, Thole S, Voget S, et al. (2011). Comparative genome analysis and genome-guided physiological analysis of Roseobacter litoralis. BMC Genomics 12, 324.

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

Kennedy J, Flemer B, Jackson SA, et al. (2010). Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. Marine Drugs 8, 608–628.

Koblízek M. (2011). *Role of Photoheterotrophic Bacteria in the Marine Carbon Cycle. Microbial Carbon Pump in the Ocean.* Science/AAAS, pp. 49-51.

Kolber ZS. (2001). Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. Science 292, 2492–2495.

Kolber ZS, Van Dover CL, Niederman RA, and Falkowski PG. (2000). Bacterial photosynthesis in surface waters of the open ocean. Nature 407, 177–179.

Konstantinidis KT, Braff J, Karl DM, and DeLong EF. (2009). Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at Station ALOHA in the North Pacific Subtropical Gyre. Appl Environ Microbiol 75, 5345–5355.

Labrenz M, Lawson PA, Tindall BJ, Collins MD, and Hirsch P. (2005). Roseisalinus antarcticus gen. nov., sp. nov., a novel aerobic bacteriochlorophyll a-producing alpha-proteobacterium isolated from hypersaline Ekho Lake, Antarctica. Int J Syst Evol Microbiol 55, 41–47.

Lami R, Cottrell MT, Ras J, et al. (2007). High abundances of aerobic anoxygenic photosynthetic bacteria in the South Pacific Ocean. Appl Environ Microbiol 73, 4198–4205.

Lamy D, De Carvalho-Maalouf P, Cottrell MT, et al. (2011). Seasonal dynamics of aerobic anoxygenic phototrophs in a Mediterranean coastal lagoon. Aquat Microb Ecol 62, 153–163.

Lehours A-C, Cottrell MT, Dahan O, Kirchman DL, and Jeanthon C. (2010). Summer distribution and diversity of aerobic anoxygenic phototrophic bacteria in the Mediterranean Sea in relation to environmental variables. FEMS Microbiol Ecol 74, 397–409.

Ludwig W. (2004). ARB: A software environment for sequence data. Nucleic Acids Res 32, 1363–1371.

Mašin M, Čuperová Z, Hojerová E, Salka I, Grossart HP, and Kobížzek M. (2012). Distribution of aerobic anoxygenic phototrophic bacteria in glacial lakes of northern Europe. Aquat Microb Ecol 66, 77–86.

Masín M, Zdun A, Ston-Egiert J, et al. (2006). Seasonal changes and diversity of aerobic anoxygenic phototrophs in the Baltic Sea. Aquat Microb Ecol 45, 247–254.

Miller MA, Pfeiffer W, and Schwartz T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees, p. 1–8. In: *Gateway Computing Environments Workshop (GCE)*, 2010. IEEE.

Niu B, Fu L, Sun S, and Li W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinform 11, 187.

Oz A, Sabehi G, Koblizek M, Massana R, and Beja O. (2005). Roseobacter-like bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. Appl Environ Microbiol 71, 344–353.

Pace NR. (1997). A molecular view of microbial diversity and the biosphere. Science 276, 734–740.

Petersen J, Brinkmann H, and Pradella S. (2009). Diversity and evolution of repABC type plasmids in Rhodobacterales. Environ Microbiol 11, 2627–2638.

Pignatelli M, and Moya A. (2011). Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS One 6, e19984.

Prates AP, Henrique De Lima L, and Chatwin A. (2007). Coastal and marine conservation priorities in Brazil. In: Chatwin A, editor. *Priorities for Coastal and Marine Conservation in South America.* Arlington, Virginia. USA: The Nature Conservancy. pp. 15–23.

Rice P, Longden I, and Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16, 276–277.

Ritchie AE, and Johnson ZI. (2012). Abundance and genetic diversity of aerobic anoxygenic phototrophic bacteria of coastal regions of the Pacific Ocean. Appl Environ Microbiol 78, 2858–2866.

Ronquist F, and Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Rusch DB, Halpern AL, Sutton G, et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5, e77.

Salka I, Cuperova Z, Masin M, Koblizek M, and Grossart HP. (2011). Rhodoferax-related pufM gene cluster dominates the aerobic anoxygenic phototrophic communities in German freshwater lakes. Environ Microbiol 13, 2865–2875.

Schwalbach MS, and Fuhrman JA. (2005). Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. Limnol Oceanog 50, 620–628.

Spring S, and Riedel T. (2013). Mixotrophic growth of bacteriochlorophyll a-containing members of the OM60/NOR5

clade of marine gammaproteobacteria is carbon-starvation independent and correlates with the type of carbon source and oxygen availability. BMC Microbiol 13, 117.

Sogin ML, Morrison HG, Huber JA, et al. (2006). Microbial diversity in the deep sea and the underexplored ''rare biosphere.'' Proc Natl Acad Sci 103, 12115–12120.

Suzuki MT, Preston CM, Chavez FP, and DeLong EF. (2001). Quantitative mapping of bacterioplankton populations in seawater: Field tests across an upwelling plume in Monterey Bay. Aquat Microb Ecol 24, 117–127.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28, 2731–2739.

Tringe SG, and Rubin EM. (2005). Metagenomics: DNA sequencing of environmental samples. Nature Rev Genet 6, 805–814.

Venter JC, Remington K, Heidelberg JF, et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74.

Wagner G, Jardim R, Tschoeke DA, et al. (2014). Stingray: System for integrated genomic resources and analysis. BMC Res Notes 7, 132.

Waidner LA, and Kirchman DL. (2005). Aerobic anoxygenic photosynthesis genes and operons in uncultured bacteria in the Delaware River. Environ Microbiol 7, 1896–1908.

Waidner LA, and Kirchman DL. (2007). Aerobic anoxygenic phototrophic bacteria attached to particles in turbid waters of the Delaware and Chesapeake estuaries. Appl Environ Microbiol 73, 3936–3944.

Waidner LA, and Kirchman DL. (2008). Diversity and distribution of ecotypes of the aerobic anoxygenic phototrophy gene pufM in the Delaware estuary. Appl Environ Microbiol 74, 4012–4021.

Wemheuer B, Gullert S, Billerbeck S, et al. (2014). Impact of a phytoplankton bloom on the diversity of the active bacterial community in the southern North Sea as revealed by metatranscriptomic approaches. FEMS Microbiol Ecol 87, 378–389.

Wommack KE, Bhavsar J, and Ravel J. (2008). Metagenomics: Read length matters. Appl Environ Microbiol 74, 1453–1463.

Yooseph S, Sutton G, Rusch DB, et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Expanding the universe of protein families. PLoS Biol 5, e16.

Yurkov V, and Csotonyi J. (2009). New lght on aerobic anoxygenic phototrophs. In: Hunter CN, Daldal F, Thurnauer M, Beatty JT (eds.). *The Purple Phototrophic Bacteria.* Springer Netherlands; pp. 31–55.

Yurkov V, and Hughes E. (2013). Chapter Eleven—Genes associated with the peculiar phenotypes of the aerobic anoxygenic phototrophs. In: J. Thomas Beatty (ed.), *Advances in Botanical Research.* Academic Press; pp. 327–358.

Yutin N, Suzuki MT, Teeling H, et al. (2007). Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environ Microbiol 9, 1464–1475.

Zhang Y, and Jiao N. (2007). Dynamics of aerobic anoxygenic phototrophic bacteria in the East China Sea: AAPB in the East China Sea. FEMS Microbiol Ecol 61, 459–469.

Zhu W, Lomsadze A, and Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38, e132–e132.

Address correspondence to:
Dr. Alberto Dávila
Computational and Systems Biology Laboratory
IOC, FIOCRUZ
Av. Brasil 4365
Rio de Janeiro, RJ 21040-360
Brazil

E-mail: davila@fiocruz.br