ORIGINAL ARTICLE

High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton

Manuel Martinez-Garcia, Brandon K Swan, Nicole J Poulton, Monica Lluesma Gomez, Dashiell Masland, Michael E Sieracki and Ramunas Stepanauskas
Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, ME, USA

Recent discoveries suggest that photoheterotrophs (rhodopsin-containing bacteria (RBs) and aerobic anoxygenic phototrophs (AAPs)) and chemoautotrophs may be significant for marine and freshwater ecosystem productivity. However, their abundance and taxonomic identities remain largely unknown. We used a combination of single-cell and metagenomic DNA sequencing to study the predominant photoheterotrophs and chemoautotrophs inhabiting the euphotic zone of temperate, physicochemically diverse freshwater lakes. Multi-locus sequencing of 712 single amplified genomes, generated by fluorescence-activated cell sorting and whole genome multiple displacement amplification, showed that most of the cosmopolitan freshwater clusters contain photoheterotrophs. These comprised at least 10–23% of bacterioplankton, and RBs were the dominant fraction. Our data demonstrate that *Actinobacteria*, including clusters *acl*, *Luna* and *acSTL*, are the predominant freshwater RBs. We significantly broaden the known taxonomic range of freshwater RBs, to include *Alpha*-, *Beta*-, *Gamma*- and *Deltaproteobacteria*, *Verrucomicrobia* and *Sphingobacteria*. By sequencing single cells, we found evidence for inter-phyla horizontal gene transfer and recombination of rhodopsin genes and identified specific taxonomic groups involved in these evolutionary processes. Our data suggest that members of the ubiquitous betaproteobacteria *Polynucleobacter* spp. are the dominant AAPs in temperate freshwater lakes. Furthermore, the *RuBisCO* (ribulose 1,5-bisphosphate carboxylase/oxygenase) gene was found in several single cells of *Betaproteobacteria*, *Bacteroidetes* and *Gammaproteobacteria*, suggesting that chemoautotrophs may be more prevalent among aerobic bacterioplankton than previously thought. This study demonstrates the power of single-cell DNA sequencing addressing previously unresolved questions about the metabolic potential and evolutionary histories of uncultured microorganisms, which dominate most natural environments.

The ISME Journal (2012) 6, 113–123; doi:10.1038/ismej.2011.84; published online 30 June 2011

Subject Category: integrated genomics and post-genomics approaches in microbial ecology

Keywords: photoheterotrophs; rhodopsin; *pufM*; single cell; *RuBisCO*; bacterioplankton

Introduction

Photosynthetic reactions performed by algae and cyanobacteria are the primary autochthonous sources of energy and organic carbon in most aquatic ecosystems. However, recent studies demonstrate that some heterotrophic, planktonic bacteria harness solar energy to produce ATP, in this way supplementing their energy requirements but not fixing inorganic carbon (Zubkov, 2009). Such photoheterotrophs include rhodopsin-containing bacteria (RBs) (Beja *et al.*, 2002) and aerobic anoxygenic

phototrophs (AAPs) (Yurkov and Beatty, 1998; Beja *et al.*, 2002). Both RBs and AAPs are abundant in the ocean (Béjà *et al.*, 2000; de la Torre *et al.*, 2003; Sabeji *et al.*, 2005; Jiao *et al.*, 2007, 2010; DeLong and Béjà, 2010), potentially contributing significantly to the ecosystem productivity. In marine ecosystems, members of *Proteobacteria*, *Flavobacteria*, *Planctomycetes* and *Euryarchaea* have been found to contain rhodopsins (DeLong and Béjà, 2010), whereas AAPs have been identified among *Alpha*- and *Gammaproteobacteria* (Allgaier *et al.*, 2003; Cho *et al.*, 2007). In contrast to the marine environments, only a handful of studies on photoheterotrophy have been conducted in freshwater ecosystems (Waidner and Kirchman, 2005; Atamna-Ismaeel *et al.*, 2008; Mašin *et al.*, 2008; Sharma *et al.*, 2008, 2009; Eiler *et al.*, 2009). So far, only *Actinobacteria* have been found to possess rhodopsins in freshwater ecosystems, as a result of

Correspondence: R Stepanauskas, Single Cell Genomics Center, Bigelow Laboratory for Ocean Sciences, PO Box 475, 180 McKown Point Road, West Boothbay Harbor, ME 04575-0475, USA.

E-mail: rstepanauskas@bigelow.org

Received 10 March 2011; revised 3 May 2011; accepted 4 May 2011; published online 30 June 2011

metagenomic- and cultivation-based studies (Sharma *et al.*, 2009). In the case of freshwater AAPs, several alpha- and betaproteobacteria strains have been isolated (Yurkov and Beatty, 1998; Suyama *et al.*, 1999; Page *et al.*, 2004; Gich and Overmann, 2006; Wagner-Döbler and Biebl, 2006) and a few surveys have been conducted to study the diversity and distribution of genes involved in aerobic AAP, such as *pufM* and *BchlY* (Waidner and Kirchman, 2005; Yutin *et al.*, 2005; Mašín *et al.*, 2008). Thus, the existing data suggest that RBs and AAPs are present and diverse in freshwater environments, but their abundance and taxonomic identities remain largely unknown.

Chemoautotrophs constitute another potentially underappreciated functional group of freshwater bacterioplankton. Several recent studies demonstrate significant CO₂ fixation in the dark in both anoxic and oxygenated water column, and some of this CO₂ fixation appears to be driven by non-pigmented prokaryotes of unknown taxonomic affiliation (García-Cantizano *et al.*, 2005; Casamayor *et al.*, 2008). In contrast, the only molecular survey on the diversity of the ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) gene in freshwater bacterioplankton that we are aware of (Tabita *et al.*, 2008) concluded (based on RuBisCO phylogeny) that photosynthetic organisms were the only autotrophs in the epilimnion. The paucity of information about the taxonomic identities of important photoheterotrophs and chemoautotrophs is primarily the result of methodological limitations. On the one hand, it is well known that current cultivation techniques do not recover the vast majority of the environmental microbial diversity (Rappé and Giovannoni, 2003). On the other hand, most culture-independent research tools, such as environmental polymerase chain reaction (PCR)-based gene surveys or metagenomic shotgun sequencing, are poorly suited to link metabolic genes to taxonomic markers at the organism level (Rusch *et al.*, 2007).

To circumvent these methodological limitations and to significantly expand our knowledge of photoheterotroph and chemoautotroph diversity in freshwater ecosystems, we employed multi-locus DNA sequencing from individual microbial cells (Raghunathan *et al.*, 2005; Zhang *et al.*, 2006), which has been proven suitable for the study of uncultured microorganisms (Kvist *et al.*, 2007; Marcy *et al.*, 2007; Stepanauskas and Sieracki, 2007; Woyke *et al.*, 2009). Our approach enabled us to link photoheterotrophic and chemoautotrophic gene markers to specific taxonomic groups of bacterioplankton inhabiting the epilimnia of temperate freshwater lakes.

Methods

Sample collection

Water samples were collected from 0.5 to 1 m depth of the temperate freshwater lakes Mendota, Damariscotta, Sparkling and Trout Bog (Supplementary

Table S1). The same samples were used for metagenomic shotgun sequencing and for single-cell analyses. For metagenomics, microbial biomass from 0.5 to 4.5 l was collected on 0.2-µm pore size membranes (Supor PES filters, Pall Corporation, NY, USA) and stored at -80 °C until DNA extraction. For single-cell analyses, replicate, 1-ml aliquots of environmental samples were cryopreserved with 6% glycine betaine (Sigma, St Louis, MO, USA) at -80 °C until used (Cleland *et al.*, 2004).

Metagenomic analyses

DNA extractions from Damariscotta samples were carried out using the PowerWater kit (MoBio Laboratories Inc., Carlsbad, CA, USA) following the manufacturer's protocol. The obtained DNA was concentrated with Microcon YM-10 columns (Millipore, Bedford, MA, USA) until reaching the desired concentration for 454 pyrosequencing. For Mendota, Sparkling and Trout Bog samples, DNA extractions were performed using the xanthogenate-sodium dodecyl sulfate protocol slightly modified from that described by Tillett and Neilan (2000). Briefly, each filter was incubated with 250 µl of TE buffer (10 mM Tris-Cl, pH 7.5 and 1 mM EDTA) and 25 µl of lysozyme (10 mg/ml) for 10 min at room temperature. Then, 50 µl of 10 mg/ml proteinase K was added and incubated at 55 °C for 1 h, and then proceeded as described (Tillett and Neilan, 2000). A total of eight shotgun libraries (two per lake: spring and summer) were constructed at the KTH Genome Center (Stockholm, Sweden) and the Institute for Genome Sciences (Baltimore, MD, USA) using the 454 GS FLX Titanium Sequencing Platform (Roche, Branford, CT, USA) according to the manufacturer's instructions (Supplementary Table S2). The obtained 454 reads were quality-trimmed and the redundant reads were removed from the data set (Gomez-Alvarez *et al.*, 2009). The individual 454 reads were then annotated using both the RAMMCP pipeline (Li, 2009) implemented in the CAMERA (<https://portal.camera.calit2.net>) and the MG-RAST server (Meyer *et al.*, 2008). In addition, we built a local database with our metagenomic reads and used it for BLASTx similarity searches to detect the rhodopsin, *pufLM*, *BchlY*, RuBisCO and *recA/radA* sequences present in the metagenomes. We used the standalone BLAST 2.2.22+ package (<ftp://ftp.ncbi.nlm.nih.gov/blast/>) with an E-value cutoff of 10⁻⁵. Sequences detected by the standalone BLAST, CAMERA and MG-RAST were used to design new or to improve existing primers using Primer 3 (Untergasser *et al.*, 2007).

Single-cell sorting, whole genome amplification, and PCR screening of SAG libraries

Before cell sorting, environmental samples with prokaryote cell abundances above 5 × 10⁵ ml⁻¹ were diluted 10 × with sterile-filtered lake water and

pre-screened through a 70- μ m mesh-size cell strainer (Becton Dickinson, Franklin Lakes, NJ, USA). For prokaryote detection, diluted subsamples (1–3 ml) were incubated for 10–120 min with SYTO-9 DNA stain (5 μ M final concentration; Invitrogen, Carlsbad, CA, USA). The high nucleic acid (HNA) and low nucleic acid (LNA) cell fractions were sorted separately. Cell sorting was performed with a MoFlo (Beckman Coulter, Danvers, MA, USA) flow cytometer using a 488 nm argon laser for excitation, a 70 μ m nozzle orifice and a CyClone robotic arm for droplet deposition into microplates. The cytometer was triggered on side scatter. The ‘single 1 drop’ mode was used for maximal sort purity, which ensures the absence of non-target particles within the target cell drop and the drops immediately surrounding the cell. The accuracy of 10 μ m fluorescent bead deposition into the 384-well plates was verified by microscopically examining the presence of beads in the plate wells. Of the 2–3 plates examined each sort day, <2% wells were found to not contain a bead and only <0.5% wells were found to contain more than one bead, indicating thus very high purity of single cells. The latter is most likely caused by co-deposition of two beads attached to each other, which at certain orientation may have similar optical properties to a single bead. Cells were deposited into 384-well plates containing 0.6 μ l per well of either (a) 1 $\times$ TE buffer or (2) prepGEM Bacteria (Zygem, Solana Beach, CA, USA) reaction mix, and stored at –80 °C until further processing. Of the 384 wells, 315 were dedicated for single cells, 66 were used as negative controls (no droplet deposition) and 3 received 10 cells each (positive controls).

The cells that were sorted into TE buffer (most of the single amplified genomes (SAGs)) were lysed and their DNA was denatured using cold KOH (Raghunathan *et al.*, 2005). The cells that were sorted into the prepGEM Bacteria reaction mix (SAG names starting with AAA041) were first lysed following Zygem instructions and then exposed to KOH treatment as above. There was no statistically significant difference ($P < 0.05$) between the composition of SAGs obtained using a KOH lysis treatment or a combination of prepGEM enzymes and KOH for cell lysis. Genomic DNA from the lysed cells was amplified using multiple displacement amplification (MDA) to generate enough template for subsequent, multiple PCR-based or genomic sequencing (Dean *et al.*, 2002; Raghunathan *et al.*, 2005) in 10 μ l final volume. The MDA reactions contained 2 U/ μ l Replphi polymerase (Epicentre, Madison, WI, USA), 1 $\times$ reaction buffer (Epicentre), 0.4 mM each dNTP (Epicentre), 2 mM dithiothreitol (Epicentre), 50 mM phosphorylated random hexamers (IDT) and 1 μ M SYTO-9 (Invitrogen) (all final concentration). The MDA reactions were run at 30 °C for 12–16 h, and then inactivated by 15 min incubation at 65 °C. The amplified genomic DNA was stored at –80 °C until further processing. We refer to the MDA products originating from individual cells as SAGs.

The instruments and the reagents were decontaminated for DNA before sorting and MDA setup, as described previously (Stepanaukas and Sieracki, 2007). DNA contaminants in MDA reagents were crosslinked by an UV treatment in Stratilinker (Stratagene, Santa Clara, CA, USA) for 40–90 min, rendering them unamplifiable by MDA. During UV treatment, reagents were placed on ice to avoid overheating. An empirical optimization of the UV exposure was performed to ensure the removal of amplifiable contaminants without inactivating MDA. Cell sorting and MDA setup were performed in a high-efficiency particulate air-filtered environment. As a quality control, the kinetics of all MDA reactions were monitored by measuring the SYTO-9 fluorescence using either LightCycler 480 (Roche) or FLUOstar Omega (BMG, Cary, NC, USA). The critical point (Cp) was determined for each MDA reaction as the time required to produce half of the maximal fluorescence. The Cp is inversely correlated to the amount of DNA template (Zhang *et al.*, 2006). Only microplates in which Cp values were significantly lower in 1-cell wells compared to 0-cell wells ($P < 0.05$; Wilcoxon’s two-sample test) were used in further analysis. Humic Lake Trout Bog SAG libraries had very low MDA success rate, probably due to the high concentration of humic acids. Thus, that particular sample was not considered in our estimates of photoheterotroph abundance.

The MDA products were diluted 50-fold in sterile TE buffer. Then, 0.5 μ l aliquots of the dilute MDA products served as templates in 5 μ l real-time PCR screens. The SSU rRNA, *pufM*, BchlY, rhodopsin and RuBisCO genes were targeted in these PCR using primers and thermal cycling conditions specified in Supplementary Table S3. Forward (5′-GTAAAACGACGGCCAGT-3′) or reverse (5′-CAGGAAACAGCTATGACC-3′) M13 sequencing primer was appended to the 5′ end of each PCR primer to aid direct sequencing of the PCR products. All PCRs were performed using LightCycler 480 SYBR Green I Master mix (Roche) in a LightCycler 480 II real-time thermal cycler (Roche). The real-time PCR kinetics and the amplicon melting curves served as proxies detecting SAG target genes. New 20 μ l PCR reactions were set up for the PCR-positive SAGs and the amplicons were sequenced from both ends using M13 targets and Sanger technology by the Beckman Coulter Genomics.

Single-cell sorting, whole genome amplification and real-time PCR screens were performed at the Bigelow Laboratory Single Cell Genomics Center (www.bigelow.org/scgc). Our previous studies and other recent publications using our single-cell sequencing techniques demonstrate the reliability of our methodology with insignificant levels of DNA contamination (Stepanaukas and Sieracki, 2007; Woyke *et al.*, 2009; Fleming *et al.*, 2011; Hess *et al.*, 2011; Heywood *et al.*, 2011).

Table 1 Summary of SAGs analyzed and genes obtained from single cells

Lake	SAGs	16S rRNA ^a	Rhodopsin	pufM	BchlY	Photoheterotrophs ^b (%)	RuBisCO
Damariscotta (spring)	945	200 (21%)	40	4	2	23	1
Damariscotta (summer)	945	179 (19%)	34	1	1	20	0
Mendota (spring)	630	188 (30%)	33	3	0	20	1
Sparkling (spring)	630	145 (23%)	12	5	0	10	6
Total	3150	712 (23%)	119	13	3	133	8

Abbreviations: AAP, aerobic anoxygenic phototroph; RB, rhodopsin-containing bacteria; RuBisCO, ribulose 1,5-bisphosphate carboxylase/oxygenase; SAG, single amplified genome.

^aNumber of SAGs yielding 16S rRNA gene sequences (success rate, %).

^bA total of 133 photoheterotrophic cells were identified. Frequency of photoheterotrophs (RBs and AAPs) among the SAGs that yielded 16S rRNA gene.

Phylogenetic analysis

The 16S rRNA gene sequences obtained from SAGs were aligned using the SILVA aligner (Pruesse *et al.*, 2007). Only sequences displaying $\geq 80\%$ of the alignment quality score in the SILVA aligner were considered for the analysis (<http://www.arb-silva.de/>). Phylogenetic analysis based on maximum likelihood (1000 bootstrap replications) was performed with RAxML version 7.0.3 (Stamatakis, 2006) implemented in ARB package (Ludwig *et al.*, 2004) using the reference ARB database 102 containing 460 783 high-quality 16S rRNA sequences (<http://www.arb-silva.de/>). The core tree was calculated with the closest reference sequences and then partial sequences from SAGs (360–833 nucleotide positions) were added using the ARB parsimony tool. Sequences of *pufM*, *BchlY*, rhodopsin and RuBisCO from SAGs were translated to amino acids, aligned with ClustalW and manually revised. The resulting protein alignment was used as a scaffold for constructing the corresponding nucleotide alignment using RevTrans 1.4 Server (Wernersson and Pedersen, 2003). Both the protein and the nucleotide alignments were used to infer the evolutionary history of the studied genes based on maximum likelihood (1000 bootstrap replications) using RAxML version 7.0.4. Recombination detection was performed following the model of dual multiple change-point on aligned nucleotide sequences with the program DualBrothers (Minin *et al.*, 2005). GenBank accession numbers: 16S rRNA (HQ662961–HQ663702), rhodopsin (HQ663727–HQ663845), *pufM* (HQ663703–HQ663715), *BchlY* (HQ663724–HQ663726) and RuBisCO (HQ663716–HQ663723).

Results and discussion

Taxonomic composition of SAGs

Water samples to build the SAG libraries were collected from the euphotic zone of the temperate freshwater lakes Mendota, Damariscotta, Sparkling and Trout Bog (see Supplementary Table S1 for lake characteristics). A total of 3150 SAGs of randomly sorted freshwater planktonic prokaryotes were generated and PCR-screened for the 16S rRNA gene (Table 1). We successfully sequenced the 16S rRNA

gene from 712 SAGs, yielding 5–30% success rate, depending on the lake and season (Table 1). Combined, *Actinobacteria*, *Betaproteobacteria* and *Gammaproteobacteria* comprised 61–97% of SAGs from the studied lakes (Supplementary Figure S1). Each of these groups were dominated by clusters that were previously found to be abundant in freshwater environments using other methods, such as the *Actinobacteria* group *acI* (Supplementary Figure S2a) and the *betaproteobacteria* *Polynucleobacter* spp. (Supplementary Figure S2b) (Warnecke *et al.*, 2005; Allgaier and Grossart, 2006; Rusch *et al.*, 2007; Jezberová *et al.*, 2010). Other ubiquitous albeit less abundant freshwater clusters represented in the SAG libraries included the *Alphaproteobacteria* LD12 clade (Zwart *et al.*, 2002), *Bacteroidetes*, *Deltaproteobacteria* and *Verrucomicrobia* (Supplementary Figures S2d–g). No archaeal 16S rRNA sequences were detected in the studied SAG libraries or in the metagenomic shotgun libraries of the lakes annotated with CAMERA and MG-RAST pipelines. Overall, bacterial diversity data obtained by metagenomics and single-cell sequencing showed similar taxonomic composition, with 59–83% of 16S rRNA gene sequences retrieved using the two techniques displayed $> 97\%$ similarity (Supplementary Figure S3). Furthermore, similar relative abundances were obtained in the 454 shotgun and SAG libraries for the predominant freshwater groups such as *Actinobacteria* and *Betaproteobacteria*, which together comprised 45–60% of total SAGs in the studied lakes. Therefore, the diversity and relative abundance of the obtained SAGs are consistent with metagenomic data and previous studies of similar freshwater environments (Zwart *et al.*, 2002; Warnecke *et al.*, 2005; Allgaier and Grossart, 2006; Rusch *et al.*, 2007; Jezberová *et al.*, 2010), indicating that our single-cell sequencing techniques were suitable to represent the full spectrum of the most abundant epilimnetic bacterioplankton groups in temperate freshwater lakes.

In natural aquatic environments, bacteria with HNA and LNA content are commonly observed with flow cytometry after cell staining with nucleic acid-specific fluorescent dyes (Gasol and Del Giorgio, 2008). We compared the taxonomic composition of SAGs generated from the HNA and LNA bacterioplankton fractions (Figure 1). Results show

Figure 1 Principal coordinates analysis of weighted UniFrac pairwise distances between 16S rRNA gene sequences from the studied environmental samples and the HNA and LNA cell fractions. A neighbor-joining tree (Jukes–Cantor substitution model) including all 16S rRNA gene sequences from SAGs served as the input data for the Fast UniFrac analysis. The archaeon *Nitrosopumilus maritimus* (CP000866) was used as an outgroup.

that the taxonomic composition of LNA cells differed from HNA cells in Damariscotta spring and Mendota spring samples, whereas no significant differences were observed in Damariscotta summer and Sparkling spring samples. This suggests that taxonomic differences between HNA and LNA cells are subject to spatial and temporal variation. Some marine studies suggest that only HNA cells are metabolically active (Gasol and Del Giorgio, 2008), whereas other reports contradict this simple dichotomy (Jochem *et al.*, 2004; Zubkov *et al.*, 2004; Longnecker *et al.*, 2005, 2006; Bouvier *et al.*, 2007). Recently, Wang *et al.* (2009) have shown in freshwater that LNA bacteria affiliated to the *Polynucleobacter* cluster utilize natural assimilable organic carbon and show high growth rates. Interestingly, 95% of the *Polynucleobacter* spp. SAGs originated from HNA cells in our study, whereas there was no such HNA/LNA separation among *Actinobacteria* acI (51% HNA) and *Alphaproteobacteria* LD12 (45% HNA) SAGs. This contrasts with findings from marine systems, where SAR11, the sister group of LD12, is predominantly found in the LNA fraction (Hill *et al.*, 2010).

Abundance of photoheterotrophs

Sequences of rhodopsin, *pufM* and BchlY genes recovered from metagenomic shotgun sequencing of the studied freshwater samples (Supplementary Tables S4) were used to design multiple, optimized primers to PCR amplify and sequence these genes from individual SAGs (Supplementary Table S3). Owing to cost constraints, only two pairs of rhodopsin primers, representing the most abundant metagenomic sequences, were used in the SAG screening (Supplementary Table S3); these primers covered 50–100% of the forward targets and 78–100% of the reverse targets found in the four metagenomic data sets (Supplementary Table S4A). The *pufM* and BchlY primers used in the SAG analysis covered 100% of the diversity of these

Figure 2 The relative frequency of photoheterotrophs in freshwater bacterioplankton, as determined by single-cell approach and metagenomic sequencing. The frequency of photoheterotrophs among SAGs was determined as the fraction of 16S rRNA-positive SAGs from which rhodopsin or *pufM* gene was recovered. The frequency of photoheterotrophs in metagenomes was determined as the ratio of either rhodopsin or *pufM* to *recA* + *radA*. M, metagenomics; SCG, single-cell genomics; Dam, Damariscotta Lake.

genes found in the studied metagenomes (Supplementary Table S4B). In total, we PCR-amplified and sequenced 119 rhodopsin, 13 *pufM* and 3 BchlY genes from SAGs. As the 16S rRNA genes were also sequenced from the same SAGs, this multi-locus sequencing analysis of individual cells provided cultivation-unbiased taxonomic identity of 133 photoheterotrophic freshwater bacteria (Table 1).

Among the studied environmental samples, rhodopsin genes were detected in 8–20% of the SAGs and either *pufM* or BchlY or both were detected in 2–3% of the SAGs (Table 1). This should be considered a conservative estimate of photoheterotrophic bacterioplankton abundance, due to PCR limitations, such as primer–target mismatches (discussed above) and template secondary structures (Potvin and Lovejoy, 2009). The uneven genome amplification by MDA (Zhang *et al.*, 2006; Woyke *et al.*, 2009) may also lead to some PCR reactions to fail. However, the range of AAP abundance detected here is within the published range for temperate freshwater systems obtained by infrared epifluorescence microscopy (<1–20% of total bacteria) (Mašín *et al.*, 2008). Contrary to AAPs, RB abundances cannot be estimated by microscopy. Thus, single-cell sequencing circumvents current methodological limitation to study rhodopsin abundances in microbial communities (DeLong and Béjà, 2010). As an alternative way to determine photoheterotroph abundance, we calculated the ratios of rhodopsin and *pufM* genes to the conserved single copy gene *recA* in the metagenomic data sets obtained from the same lake water samples. Assuming that no more than one copy of these genes occurs in each cell, rhodopsin and *pufM* genes were present in 37–56% and 3–37% of the studied freshwater bacterioplankton samples, respectively (Figure 2).

These metagenomics-based estimates likely better reflect the true frequencies of phototrophs in the studied samples, albeit they do not provide information on photoheterotroph identities, which is a major advantage of single-cell sequencing. Furthermore, 454 shotgun sequencing is also prone to some biases that may distort gene frequency information (Morgan *et al.*, 2010). Despite the existing methodological limitations for photoheterotroph quantification, it is clear that AAPs and RBs constitute a major fraction of freshwater bacterioplankton, at least 8–23% across various types of lakes.

Identity of photoheterotrophs

Rhodopsin genes obtained from SAGs formed two major clusters (Figure 3a). The first cluster was composed of 95 actinobacteria and 5 gammaproteobacteria SAGs and grouped together with the previously published rhodopsin sequences from clones and cultures named as ‘actinorhodopsins’ (Atamna-Ismaeel *et al.*, 2008; Sharma *et al.*, 2008, 2009) (Figure 3b). The second cluster was composed of rhodopsin sequences from 15 SAGs (Figure 3c) belonging to *Alpha*-, *Beta*-, *Gamma*- and *Deltaproteobacteria*, *Verrucomicrobia*, *Sphingobacteria* and *Actinobacteria*. The latter SAG sequences grouped together with previously published sequences designated as ‘proteorhodopsins’ (Béjà *et al.*, 2000) (detailed phylogenetic information from the 16S rRNA gene analysis is provided in Supplementary Figure S2). Three rhodopsin sequences from the SAGs AAA041-G17 (*Betaproteobacteria*), AAA278-C16 (*Deltaproteobacteria*) and AAA487-P23 (*Verrucomicrobia*) were phylogenetically positioned in-between these two major rhodopsin clusters. In addition, the alphaproteobacterium SAG AAA024-J18 carried a rhodopsin that was phylogenetically related to xanthorhodopsins (Figure 3a), which are proton pumps that are abundant in hypersaline environments (Balashov *et al.*, 2005). Interestingly, 45% of SAGs that carried the rhodopsin gene were HNA bacteria, indicating that rhodopsin genes were equally abundant among HNA and LNA bacterioplankton.

Before our study, only *Actinobacteria* have been found to contain rhodopsins in freshwater environments (Sharma *et al.*, 2009). Our data demonstrates *Actinobacteria*, including clusters acI, Luna and acSTL, as the predominant phylum containing rhodopsin genes in temperate freshwater lakes. In addition, we significantly broadened the known taxonomic range of rhodopsin-containing freshwater bacterioplankton, to include *Alpha*-, *Beta*-, *Gamma*- and *Deltaproteobacteria*, *Verrucomicrobia* and *Sphingobacteria*. In fact, rhodopsin-containing *Deltaproteobacteria*, *Verrucomicrobia* and *Sphingobacteria* have never been previously reported from any type of environment.

In most cases, phylogenies of the 16S rRNA and genes involved in photoheterotrophy were

congruent (Figure 3). For instance, we demonstrated that freshwater rhodopsins, related to the marine SAR11 clade, belong to the SAR11 sister group LD12 (Figure 3c and Supplementary Figure S2c). However, this congruency has exceptions. In one case, five gammaproteobacteria SAGs, originating from multiple environmental samples, had rhodopsin sequences clustering with *Actinobacteria*, suggesting their origin through horizontal gene transfer (HGT). In another case, an actinobacteria SAG AAA278-O22 contained two rhodopsins, one typical for this phylum and another closely related to sequences from *Betaproteobacteria*. This implies that ~5% of the observed rhodopsins may have evolved from HGT events. Earlier findings of rhodopsin HGT among phylogenetically distant microbes is consistent with our results (McCarren and DeLong, 2007). Besides HGT, our study provides the first evidence for recombination between actinorhodopsin- and proteorhodopsin-like genes (Figure 3d), resulting in composite rhodopsins, such as those found in the SAGs AAA278-C16 (*Deltaproteobacteria*) and AAA041-G17 (*Betaproteobacteria*) that are phylogenetically positioned in-between the two major rhodopsin clusters (Figure 3a).

We detected either *pufM* or *BchlY* or both in 15 SAGs (Table 1). Unexpectedly, most of these putative AAPs (53%) were *Betaproteobacteria*, primarily members of the *Polynucleobacter* cluster (Figure 4). Two alphaproteobacteria and three gammaproteobacteria SAGs related to *Roseomonas* and *Pseudomonas* spp. also had *pufM* genes. Thus, members of the ubiquitous *Polynucleobacter* cluster (Jezberová *et al.*, 2010) may be among the predominant freshwater AAPs. This is contrary to earlier data obtained from cultures, where freshwater AAPs have been primarily detected among *Alphaproteobacteria* (Yurkov and Beatty, 1998; Suyama *et al.*, 1999; Page *et al.*, 2004; Gich and Overmann, 2006; Wagner-Döbler and Biebl, 2006). Interestingly, genes involved in aerobic AAP are absent in *Polynucleobacter* strain sp. QLW-P1DMWA-1, the only planktonic freshwater bacterium with whole genome information available. Our study demonstrates how single-cell sequencing can provide more reliable and extensive information about the metabolic potential of specific microbial assemblage members compared with other available methods.

Potential planktonic chemoautotrophs

We detected RuBisCO genes in several SAGs of *Beta*- and *Gammaproteobacteria* and *Bacteroidetes*, raising the possibility that they fix inorganic carbon (Figure 4). These SAGs appear to represent aerobic, planktonic organisms, as closely related 16S rRNA gene clones have been obtained from the euphotic and well oxygenated waters of different lakes and continents (Supplementary Figure S2). Moreover, none of these SAGs are related to known

chemolithoautotrophic or photoautotrophic anaerobic bacteria, which would indicate resuspension from sediments or hypolimnion. All the RuBisCO

sequences obtained from these SAGs were from IA or IB, which are typically found in aerobic environments (Tabita *et al.*, 2008). Significantly, PCR

Figure 3 Maximum-likelihood tree of 119 rhodopsin proteins from single cells: (a) general tree; (b) subtree of actinorhodopsin-like proteins and (c) subtree of proteorhodopsin-like proteins. In all, 120 amino-acid positions were used in the tree construction. Bootstrap values ≥ 50 are displayed. SAGs obtained from the HNA and LNA cells are indicated in regular and italic fonts, respectively. The taxonomic identity of rhodopsin-containing SAGs, based on their 16S rRNA gene phylogeny, is provided next to the SAG name (for detailed phylogeny of the 16S rRNA genes, see Supplementary Figure S2). (d) Recombination analysis of rhodopsin genes. The dual multiple change-point model that considers the spatial variation of tree topologies and the substitution process parameters was applied in a Bayesian framework using reversible jump Markov chain Monte Carlo sampling to approximate the joint posterior distribution of all model parameters. Parameters of transition:transversion (κ) and expected divergence (μ) and spatial variation of tree topologies are indicated. Each one of the breakpoints shown in the tree topologies together with κ and μ parameters indicate a putative recombination event. Recombination was not detected within the actinorhodopsin and proteorhodopsin clusters.

Figure 4 Maximum-likelihood phylogenetic analysis of *pufM* and RuBisCO genes and the corresponding 16S rRNA sequences from single cells (limited to *Betaproteobacteria* owing to space constraints). The taxonomic identity of the *pufM*- and RuBisCO-containing SAGs is indicated next to the SAG name. BchlY-containing SAGs are indicated in the phylogenetic tree of 16S rRNA gene (for detailed phylogeny of the 16S rRNA genes, see Supplementary Figure S2). Bootstrap (1000 replicate) values ≥ 50 are displayed. In the case of *pufM* gene, the analysis was conducted on nucleotide sequence alignment (250 nucleotide positions). For the RuBisCO gene, the analysis was based on protein sequence alignment (amino-acid positions 100–262).

primers used in our SAG screens (Supplementary Table S3) displayed 6–9 nucleotide mismatches with most RuBisCO genes detected by metagenomics and failed to amplify RuBisCO genes even from some cyanobacterial SAGs. Unfortunately, our attempts to design broader-range RuBisCO primers have failed so far, due to the high diversity of the typical conserved regions of this gene in our metagenomic data sets. The ratio of RuBisCO versus *recA* genes ranged 6–77% in the metagenomes (Supplementary Figure S4). Both the SAG analysis and the metagenomic sequencing suggested that the highest frequency of cells with RuBisCO genes were in the oligotrophic Sparkling Lake. It is important to note that in this study we did not target other autotrophic carbon fixation pathways than the Calvin–Benson–Bassham cycle. Thus, our analysis may have significantly underestimated the fraction of potential planktonic chemoautotrophs.

In addition to CO₂ fixation, RuBisCO may also be involved in the central redox cofactor recycling

in AAP bacteria inhabiting reducing environments, such as soils and sediments, sometimes thriving on organic substrates more reduced than biomass (McKinlay and Harwood, 2010). It remains to be determined whether similar mechanisms can be significant among bacterioplankton inhabiting oxygenated water column. Chemoautotrophy rather than the recycling of redox cofactors appears a more likely role of the detected SAG RuBisCO genes for the following reasons: (1) we did not find phototrophy genes *pufM* or BchlY in any of the SAGs that contained RuBisCO; and (2) RuBisCO was most abundant in SAGs from the oligotrophic Sparkling Lake, which is the lake containing the lowest concentrations of organic substrates (see Supplementary Table S1). Thus, our study provides taxonomic identities of potential freshwater chemoautotrophs that may be involved in the aerobic CO₂ fixation, a metabolic process that requires further attention to fully understand carbon cycling in freshwater environments.

Concluding remarks

Using a combination of single-cell sequencing and metagenomics, we vastly expanded the knowledge of the predominant photoheterotrophs and potential chemoautotrophs inhabiting the euphotic zone of temperate freshwater lakes. We found that all of the ubiquitous freshwater bacterioplankton clusters, such as *Actinobacteria* acI, Luna and acSTL, *Poly-nucleobacter* spp. (*Betaproteobacteria*) and LD12 (*Alphaproteobacteria*) contain photoheterotrophs, suggesting that photoheterotrophy is an important competitive strategy for freshwater bacterioplankton. Our approach enabled us to perform a high-throughput, cost-effective study while circumventing many analytical limitations inherent to earlier techniques, such as cultivation and metagenomics. For example, more than a decade of studies of marine proteorhodopsins have resulted in direct identification of only 37 bacteria containing rhodopsins, most of which belong to a few taxonomic groups (DeLong and Béjà, 2010). Here, in a 1-year study we identified 118 predominant freshwater bacteria containing rhodopsins, with no apparent taxonomic biases. Furthermore, our single-cell sequencing results indicate HGT and recombination of rhodopsin genes in freshwater bacterioplankton and link the gene's evolutionary history to the taxonomic identities of specific microbial groups that are involved in these evolutionary processes. Finally, SAGs generated here represent the largest and cultivation-unbiased genomic DNA library of freshwater bacterioplankton, opening unprecedented opportunities for additional analyses of specific loci and whole genome sequencing that will provide further insights into the metabolic potential and evolutionary histories of freshwater bacterioplankton.

Acknowledgements

We thank Wendy Korjef-Bellows and Jane Heywood for fieldwork assistance, Stefan Bertilsson and Siv Andersson for access to their metagenomic data, Katherine McMahon and Todd Miller for Wisconsin lake sampling and DNA extraction, and Vladimir M Minin for his advice on recombination event analysis. This research was supported by the NSF Grants DEB-841933 and OCE-821374 to RS and by a Maine Technology Institute research infrastructure grant to the Bigelow Laboratory.

References

Allgaier M, Grossart H-P. (2006). Diversity and seasonal dynamics of actinobacteria populations in four Lakes in Northeastern Germany. *Appl Environ Microbiol* **72**: 3489–3497.

Allgaier M, Uphoff H, Felske A, Wagner-Dobler I. (2003). Aerobic anoxygenic photosynthesis in *Roseobacter*

clade bacteria from diverse marine habitats. *Appl Environ Microbiol* **69**: 5051–5059.

Atamna-Ismaeel N, Sabehi G, Sharon I, Witzel K-P, Labrenz M, Jurgens K *et al.* (2008). Widespread distribution of proteorhodopsins in freshwater and brackish ecosystems. *ISME J* **2**: 656–662.

Balashov SP, Imasheva ES, Boichenko VA, Anton J, Wang JM, Lanyi JK. (2005). Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science* **309**: 2061–2064.

Beja O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T *et al.* (2002). Unsuspected diversity among marine aerobic anoxygenic photo-trophs. *Nature* **415**: 630–633.

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

Bouvier T, Del Giorgio PA, Gasol JM. (2007). A comparative study of the cytometric characteristics of high and low nucleic-acid bacterioplankton cells from different aquatic ecosystems. *Environ Microbiol* **9**: 2050–2066.

Casamayor EO, García-Cantizano J, Pedrós-Alió C. (2008). Carbon dioxide fixation in the dark by photosynthetic bacteria in sulfide-rich stratified lakes with oxic-anoxic interfaces. *Limnol Oceanogr* **53**: 1193–1203.

Cho J-C, Stapels MD, Morris RM, Vergin KL, Schwalbach MS, Givan SA *et al.* (2007). Polyphyletic photosynthetic reaction centre genes in oligotrophic marine Gammaproteobacteria. *Environ Microbiol* **9**: 1456–1463.

Cleland D, Krader P, McCree C, Tang J, Emerson D. (2004). Glycine betaine as a cryoprotectant for prokaryotes. *J Microbiol Methods* **58**: 31–38.

de la Torre JR, Christianson LM, Béjà O, Suzuki MT, Karl DM, Heidelberg J *et al.* (2003). Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci USA* **100**: 12830–12835.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P *et al.* (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99**: 5261–5266.

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

Eiler A, Beier S, Sawstrom C, Karlsson J, Bertilsson S. (2009). High ratio of bacteriochlorophyll biosynthesis genes to chlorophyll biosynthesis genes in bacteria of humic lakes. *Appl Environ Microbiol* **75**: 7221–7228.

Fleming EJ, Langdon AE, Martínez-García M, Stepanauskas R, Poulton N, Masland D *et al.* (2011). What's new is old: resolving the identity of *Leptothrix ochracea* using single cell genomics, pyrosequencing and FISH. *PLoS One* **6**: e17769.

García-Cantizano J, Casamayor E, Gasol J, Guerrero R, Pedrós-Alió C. (2005). Partitioning of CO₂; incorporation among planktonic microbial guilds and estimation of *in situ* specific growth rates. *Microb Ecol* **50**: 230–241.

Gasol JM, Del Giorgio PA. (2008). Physiological structure and single-cell activity in marine bacterioplankton. In: Kirchman DL (ed). *Microbial Ecology of the Oceans*. John Wiley & Sons Inc: Hoboken, NJ, pp 243–298.

Gich F, Overmann J. (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.
- Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.
- Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G *et al.* (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463–467.
- Heywood JL, Sieracki ME, Bellows W, Poulton NJ, Stepanauskas R. (2011). Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J* **5**: 674–684.
- Hill PG, Zubkov MV, Purdie DA. (2010). Differential responses of *Prochlorococcus* and SAR11-dominated bacterioplankton groups to atmospheric dust inputs in the tropical Northeast Atlantic Ocean. *FEMS Microbiol Lett* **306**: 82–89.
- Jezberová J, Jezbera J, Brandt U, Lindström ES, Langenheder S, Hahn MW. (2010). Ubiquity of *Poly-nucleobacter necessarius* ssp. *asymbioticus* in lentic freshwater habitats of a heterogeneous 2000 km² area. *Environ Microbiol* **12**: 658–669.
- Jiao N, Zhang Y, Zeng Y, Hong N, Liu R, Chen F *et al.* (2007). Distinct distribution pattern of abundance and diversity of aerobic anoxygenic phototrophic bacteria in the global ocean. *Environ Microbiol* **9**: 3091–3099.
- Jiao N, Zhang F, Hong N. (2010). Significant roles of bacteriochlorophyll *a* supplemental to chlorophyll *a* in the ocean. *ISME J* **4**: 595–597.
- Jochem FJ, Lavrentyev PJ, First MR. (2004). Growth and grazing rates of bacteria groups with different apparent DNA content in the Gulf of Mexico. *Mar Biol* **145**: 1213–1225.
- Kvist T, Ahring B, Lasken R, Westermann P. (2007). Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol* **74**: 926–935.
- Li W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinform* **10**: 359.
- Longnecker K, Homen DS, Sherr EB, Sherr BF. (2006). Similar community structure of biosynthetically active prokaryotes across a range of ecosystem trophic states. *Aquat Microb Ecol* **42**: 265–276.
- Longnecker K, Sherr BF, Sherr EB. (2005). Activity and phylogenetic diversity of bacterial cells with high and low nucleic acid content and electron transport system activity in an upwelling ecosystem. *Appl Environ Microbiol* **71**: 7737–7749.
- Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucl Acids Res* **32**: 1363–1371.
- Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG *et al.* (2007). Dissecting biological ‘dark matter’ with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* **104**: 11889–11894.
- Mašín M, Nedoma J, Pechar L, Koblížek M. (2008). Distribution of aerobic anoxygenic phototrophs in temperate freshwater systems. *Environ Microbiol* **10**: 1988–1996.
- McCarren J, DeLong EF. (2007). Proteorhodopsin photo-system gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ Microbiol* **9**: 846–858.
- McKinlay JB, Harwood CS. (2010). Carbon dioxide fixation as a central redox cofactor recycling mechanism in bacteria. *Proc Natl Acad Sci USA* **107**: 11669–11675.
- Meyer F, Paarmann D, D’Souza M, Olson R, Glass E, Kubal M *et al.* (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform* **9**: 386.
- Minin VN, Dorman KS, Fang F, Suchard MA. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* **21**: 3034–3042.
- Morgan JL, Darling AE, Eisen JA. (2010). Metagenomic sequencing of an *in vitro*-simulated microbial community. *PLoS One* **5**: e10209.
- Page KA, Cannon SA, Giovannoni SJ. (2004). Representative freshwater bacterioplankton isolated from Crater Lake, Oregon. *Appl Environ Microbiol* **70**: 6542–6550.
- Potvin M, Lovejoy C. (2009). PCR-based diversity estimates of artificial and environmental 18S rRNA gene libraries. *J Eukaryot Microbiol* **56**: 174–181.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* **35**: 7188–7196.
- Raghunathan A, Ferguson Jr HR, Bornarth CJ, Song W, Driscoll M, Lasken RS. (2005). Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71**: 3342–3347.
- Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The *Sorcerer II* global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.
- Sabehi G, Loy A, Jung K-H, Partha R, Spudich JL, Isaacson T *et al.* (2005). New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol* **3**: e273.
- Sharma AK, Sommerfeld K, Bullerjahn GS, Matteson AR, Wilhelm SW, Jezbera J *et al.* (2009). Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater *Actinobacteria*. *ISME J* **3**: 726–737.
- Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF. (2008). Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environ Microbiol* **10**: 1039–1056.
- Stamatakis A. (2006). RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.
- Stepanauskas R, Sieracki ME. (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* **104**: 9052–9057.
- Suyama T, Shigematsu T, Takaichi S, Nodasaka Y, Fujikawa S, Hosoya H *et al.* (1999). *Roseateles depolymerans* gen. nov., sp. nov., a new bacteriochlorophyll *a*-containing obligate aerobe belonging to the beta-subclass of the Proteobacteria. *Int J Syst Bacteriol* **49**: 449–457.
- Tabita FR, Hanson TE, Satagopan S, Witte BH, Kreel NE. (2008). Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the

- functional lessons provided by diverse molecular forms. *Philos Trans R Soc Ser B* **363**: 2629–2640.
- Tillett D, Neilan BA. (2000). Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. *J Phycol* **36**: 251–258.
- Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucl Acids Res* **35**: W71–W74.
- Wagner-Döbler I, Biebl H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annu Rev Microbiol* **60**: 255–280.
- Waidner LA, Kirchman DL. (2005). Aerobic anoxygenic photosynthesis genes and operons in uncultured bacteria in the Delaware River. *Environ Microbiol* **7**: 1896–1908.
- Wang Y, Hammes F, Boon N, Chami M, Egli T. (2009). Isolation and characterization of low nucleic acid (LNA)-content bacteria. *ISME J* **3**: 889–902.
- Warnecke F, Sommaruga R, Sekar R, Hofer JS, Pernthaler J. (2005). Abundances, identity, and growth state of Actinobacteria in mountain lakes of different UV transparency. *Appl Environ Microbiol* **71**: 5551–5559.
- Wernersson R, Pedersen AG. (2003). RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucl Acids Res* **31**: 3537–3539.
- Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.
- Yurkov VV, Beatty JT. (1998). Aerobic anoxygenic phototrophic bacteria. *Microbiol Mol Biol Rev* **62**: 695–724.
- Yutin N, Suzuki MT, Beja O. (2005). Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. *Appl Environ Microbiol* **71**: 8958–8962.
- Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**: 680–686.
- Zubkov MV, Allen JI, Fuchs BM. (2004). Coexistence of dominant groups in marine bacterioplankton community; a combination of experimental and modelling approaches. *J Mar Biol Assoc UK* **84**: 519–529.
- Zubkov MV. (2009). Photoheterotrophy in marine prokaryotes. *J Plankton Res* **31**: 933–938.
- Zwart G, Crump BC, Agterveld MPK-v, Hagen F, Han S-K. (2002). Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat Microb Ecol* **28**: 141–155.

Supplementary Information accompanies the paper on The ISME Journal website (<http://www.nature.com/ismej>)