Open

The ISME Journal (2013) 7, 137–147
© 2013 International Society for Microbial Ecology All rights reserved 1751-7362/13
www.nature.com/ismej

ORIGINAL ARTICLE

Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton

Sarahi L Garcia[1], Katherine D McMahon[2,3], Manuel Martinez-Garcia[4,9], Abhishek Srivastava[5], Alexander Sczyrba[6,7], Ramunas Stepanauskas[4], Hans-Peter Grossart[5,8], Tanja Woyke[7] and Falk Warnecke[1]

[1]*Jena School for Microbial Communication (JSMC) and Microbial Ecology Group at Friedrich Schiller University Jena, Jena, Germany;* [2]*Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA;* [3]*Department of Bacteriology, University of Wisconsin-Madison Madison, WI, USA;* [4]*Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA;* [5]*Department of Limnology of Stratified Lakes, Leibniz-Institute for Freshwater Ecology and Inland Fisheries, Stechlin, Germany;* [6]*Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany;* [7]*DOE Joint Genome Institute, Walnut Creek, CA, USA and* [8]*Institute for Biochemistry and Biology, Postdam University, Postdam, Germany*

Actinobacteria within the acI lineage are often numerically dominating in freshwater ecosystems, where they can account for >50% of total bacteria in the surface water. However, they remain uncultured to date. We thus set out to use single-cell genomics to gain insights into their genetic make-up, with the aim of learning about their physiology and ecological niche. A representative from the highly abundant acI-B1 group was selected for shotgun genomic sequencing. We obtained a draft genomic sequence in 75 larger contigs (sum = 1.16 Mb), with an unusually low genomic G + C mol% (~42%). Actinobacteria core gene analysis suggests an almost complete genome recovery. We found that the acI-B1 cell had a small genome, with a rather low percentage of genes having no predicted functions (~15%) as compared with other cultured and genome-sequenced microbial species. Our metabolic reconstruction hints at a facultative aerobe microorganism with many transporters and enzymes for pentoses utilization (for example, xylose). We also found an *actinorhodopsin* gene that may contribute to energy conservation under unfavorable conditions. This project reveals the metabolic potential of a member of the global abundant freshwater Actinobacteria.

The ISME Journal (2013) **7,** 137–147; doi:10.1038/ismej.2012.86; published online 19 July 2012
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: freshwater Actinobacteria; metabolic potential; single-cell genomics

Introduction

Freshwater bacteria have a critical role in the recycling of many biologically active elements (Lindeman, 1942; Newton *et al.,* 2011). In lakes, A*ctinobacteria* are often the numerically dominant phylum, where they can constitute >50% of epilimnion bacterioplankton (Glöckner *et al.*, 2000; Warnecke *et al.*, 2005). Of the four major lineages of freshwater, planktonic Actinobacteria, acI is the most abundant (Warnecke *et al.*, 2004; Allgaier *et al.*, 2007). A number of previous studies have focused on the ecophysiology and distribution of acI members through time and space. Their abundance was positively correlated to UV transparency in high mountain lakes (Warnecke *et al.*, 2005). Clades within the lineage appear to niche partition across lakes based on pH with some seeming to favor slightly alkaline lakes (including the cosmopolitan acI-B1 group), whereas others clearly favor more acidic systems (Newton *et al.*, 2007). Similarly, the clades show contrasting dynamics through time, likely due to a combination of changes in the availability of specific carbon substrates (Buck *et al.*, 2009), variability in predation pressure (Pernthaler *et al.*, 2001; Eckert *et al.*, 2011), changes in nutrient availability (Haukka *et al.*, 2006; Newton and McMahon, 2011) and seasonal warming and cooling (Allgaier and Grossart, 2006; Eiler *et al.,* 2012).

Correspondence: F Warnecke, Jena School for Microbial Communication (JSMC), Microbial Ecology Group, Friedrich Schiller University of Jena, Philosophenweg 12, Jena 07743, Germany.
E-mail: falk.warnecke@uni-jena.de
[9]Current address: University of Alicante, Alicante, Spain.
Received 3 April 2012; revised 11 June 2012; accepted 15 June 2012; published online 19 July 2012

Finally, acI members seem to have an important role in the mineralization of *N*-acetylglucosamine, a breakdown product of chitin and bacterial cell walls (Beier and Bertilsson, 2011).

Despite their global abundance in the environment, no isolate representatives of acI have been obtained to date. Recently a stable co-culture was established including <5.6% of a member of the acI-A clade within the acI lineage. Unfortunately, its low abundance did not enable much physiological insight into the lifestyle of these abundant players (Jezbera *et al.*, 2009). Other attempts to elucidate acI physiological and ecological roles based on gene content include analyses of fosmid-cloned genomic fragments from Lake Kinneret (Philosof *et al.*, 2009) and metagenomic analyses from lakes, estuaries (Ghai *et al.*, 2011a) and rivers (Ghai *et al.*, 2011b). From these studies it became apparent that to better study specific functional characteristics, the use of a different culture-independent method was needed.

Robust protocols were recently developed for genomic sequencing from individual microbial cells (Marcy *et al.*, 2007; Woyke *et al.*, 2009; Swan *et al.*, 2011; Martinez-Garcia *et al.*, 2011b). With these tools, it is possible to reconstruct a genome from an uncultured bacterium to analyze its metabolic potential. In this study, we employed single-cell genomics techniques to analyze a nearly complete genome of an individual cell representing the acI lineage, actinobacterium SCGC AAA027-L06. We selected this particular single amplified genome (SAG) for sequencing because it is a member of the acI-B1 tribe, the acI genotype that is most commonly detected in freshwater lake epilimina using 16S rRNA gene-based methods (Newton *et al.*, 2011).

Materials and methods

Recovery of single-cell genomic DNA

Genomic DNA was recovered from an individual cell representing the acI lineage at the Bigelow Laboratory Single Cell Genomics Center (http://www.bigelow.org/scgc), as previously described (Martinez-Garcia *et al.*, 2011b). The water sample was collected from 1 m depth of the Lake Mendota, WI, USA (43° 6′19.58″N, 89°24′28.71″W) on May 19 2009, and cryopreserved with 6% glycine betaine (Sigma) at −80 °C until used. Before cell sorting, the sample was diluted 10× with sterile-filtered Lake Mendota water and pre-screened through a 70 μm mesh-size cell strainer (Becton Dickinson, San Jose, CA, USA). For prokaryote detection, diluted subsamples (1–3 ml) were incubated for 10–120 min with SYTO-9 DNA stain (5 μM final concentration; Life Technologies, Carlsbad, CA, USA). Cell sorting was performed with a MoFlo (Beckman Coulter, Brea, CA, USA) flow cytometer using a 488-nm argon laser for excitation, a 70-μm 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. High nucleic acid content prokaryote cells were deposited into 384-well plates containing 0.6 μl 1× TE buffer per well 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 accuracy of 10 μm 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 <0.5% wells were found to contain more than one bead. 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.

The cells were lysed and their DNA was denatured using cold KOH and then amplified using multiple displacement amplification (MDA) (Dean *et al.*, 2002; Raghunathan *et al.*, 2005). The 10-μl MDA reactions contained 2 U μl^{-1} Repliphi polymerase (Epicentre), 1× reaction buffer (Epicentre), 0.4 mM each dNTP (Epicentre), 2 mM DTT (Epicentre), 50 mM phosphorylated random hexamers (IDT) and 1 μM SYTO-9 (Life Technologies) (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 previously described (Stepanauskas and Sieracki, 2007; Woyke *et al.*, 2011). Cell sorting and MDA setup were performed in a HEPA (high-efficiency particulate air)-filtered environment. As a quality control, the kinetics of all MDA reactions was monitored by measuring the SYTO-9 fluorescence using FLUOstar Omega (BMG Labtech, 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). The Cp values were significantly lower in 1-cell wells compared with 0-cell wells ($P<0.05$; Wilcoxon two sample test). Our previous studies demonstrate the reliability of our methodology with insignificant levels of DNA contamination (Stepanauskas and Sieracki, 2007; Woyke *et al.*, 2009; Martinez-Garcia *et al.*, 2011a; Fleming *et al.*, 2011; Heywood *et al.*, 2011; Swan *et al.*, 2011; Woyke *et al.*, 2011).

The MDA products were diluted 50-fold in sterile TE buffer. Then, 0.5 μl aliquots of the dilute MDA products served as templates in 5 μl real-time PCR. The small subunit rRNA and *rhodopsin* genes were targeted in these PCR using primers and thermal

cycling conditions specified in Martinez-Garcia *et al.* (2011b) and sequenced from both ends using Sanger technology at Beckman Coulter Genomics.

To obtain sufficient quantity of genomic DNA for shotgun sequencing, the original MDA products were re-amplified using similar MDA conditions as above: eight replicate 125 μl reactions were performed and then pooled together, resulting in ~100 μg of genomic dsDNA.

Selection of the SAG for sequencing

The SCGC AAA027-L06 SAG was selected from among 188 SAGs yielding amplifiable 16S rRNA genes during the library screening step, described elsewhere (Martinez-Garcia *et al.*, 2011b), because its 16S rRNA gene was clearly affiliated with the acI-B1 tribe (Figure 1). The 16S rRNA gene shared 99.9% identity with near-full-length 16S rRNA gene fragments recovered from Lake Mendota previously (Newton *et al.*, 2007), as well as many other acI-B1 sequences (Newton *et al.*, 2011). This SAG was also selected because it harbored an *actinorhodopsin* gene (Martinez-Garcia *et al.*, 2011b) and had a relatively low MDA Cp, indicative of a significant amount of DNA readily available to the phi29 polymerase.

Genome sequencing and assembly

A combination of Illumina (San Diego, CA, USA) and 454 shotgun sequencing was performed on the single-cell re-MDA product for actinobacterium SCGC AAA027-L06. For Illumina sequencing, a normalized 0.3-kb shotgun library was constructed. Briefly, 3 μg MDA product was sheared in 100 μl using the Covaris E210 (Life technologies, Carlsbad, CA, USA) with the setting of 10% duty cycle, intensity 5, and 200 cycle per burst for 3 min per sample and the fragmented DNA was purified using QIAquick columns (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. The sheared DNA was end repaired and A-tailed according to the Illumina standard PE protocol and purified using the MinElute PCR Purification Kit (Qiagen) with a final elution in 12 μl of Buffer EB. After quantification using a Bioanalyzer DNA 1000 chip (Agilent), the fragments were ligated to the Illumina adapters according to the Illumina standard PE protocol, followed by a purification step of the ligation product using AMPure SPRI beads. The library was quantified using a Bioanalyzer DNA High Sensitivity chip (Agilent, Santa Clara, CA, USA) and 300 ng of DNA (in 6 μl) then underwent normalization using the Duplex-Specific Nuclease (DSN) Kit (Axxora, Farmingdale, NY, USA) (Bogdanova *et al.*, 2009). For normalization, the dsDNA was denatured for 3 min at 98 °C, followed by a hybridization step at 68 °C for 5 h and DSN treatment at 68 °C for 20 min. The normalized library was amplified by PCR for 12 cycles, gel-purified and QC assessed on a Bioanalyzer DNA High Sensitivity chip (Agilent), and then sequenced using an Illumina GAIIx sequencer generating 7.4 Gb (96.8 M reads, 2×76 bp). For 454 pyrosequencing, a 4-kb paired-end library was constructed and sequenced generating 79.5 Mb (260 428 reads). All general aspects of and detailed protocols for library construction and sequencing can be found on the JGI website (http://www.jgi.doe.gov/).

All raw Illumina sequence data was passed through a filtering program developed at JGI, which filters out known Illumina sequencing and library preparation artifacts. Specifically, all reads containing sequencing adapters, low complexity reads and reads containing short tandem repeats were removed. Duplicated read pairs derived from PCR amplification during library preparation were identified and consolidated into a single consensus read pair. The MDA introduces a tremendous bias in the sequencing coverage of the single-cell genome. The artifact-filtered sequence data was screened and trimmed according to the k-mers present in the data set. High-depth k-mers, presumably derived from MDA amplification bias, cause problems in the assembly, especially if the k-mer depth varies in several orders of magnitude for different regions of the genome. We therefore removed reads representing high-abundance k-mers ($>64\times$ k-mer coverage, $k=31$) and trimmed reads that contain unique k-mers. These filtering steps reduced the data set from 96.8 M to 2.1 M reads with an average length of 64 bp ± 16 bp.

Assembly was performed in several steps: (1) filtered Illumina reads were assembled using Velvet version 1.1.02 (Zerbino and Birney, 2008). The VelvetOptimiser script (version 2.1.7) was used with default optimization functions (n50 for k-mer choice, total number of base pairs in large contigs for cov_cutoff optimization). (2) The Velvet contigs were used to simulate reads from long-insert libraries, which were used together with the filtered reads as input for Allpaths-LG (Gnerre *et al.*, 2011) assembly. (3) Next, Allpaths contigs larger than 1 kb were shredded into 1-kb pieces with 200 bp overlaps. (4) Lastly, the Allpaths shreds and raw 454 pyrosequence reads were assembled using the 454 Newbler assembler version 2.4 (Roche/454 Life Sciences, Branford, CT, USA). This resulted in a total assembly size of 1 163 583 bp (75 contigs, N50: 34 337 bp). The 454 data was deposited into SRA under access SRA050785.

Genome size estimate

An estimate of complete genome size was obtained for SCGC AAA027-L06 using core gene analysis. To identify the core genes, 138 Actinobacteria class finished genomes, currently available at the DOE Joint Genome Institute Integrated Microbial Genomes site (IMG, http://img.jgi.doe.gov/, (Markowitz *et al.*, 2009)) were included in the analysis (Supplementary Table S1). The analysis was carried

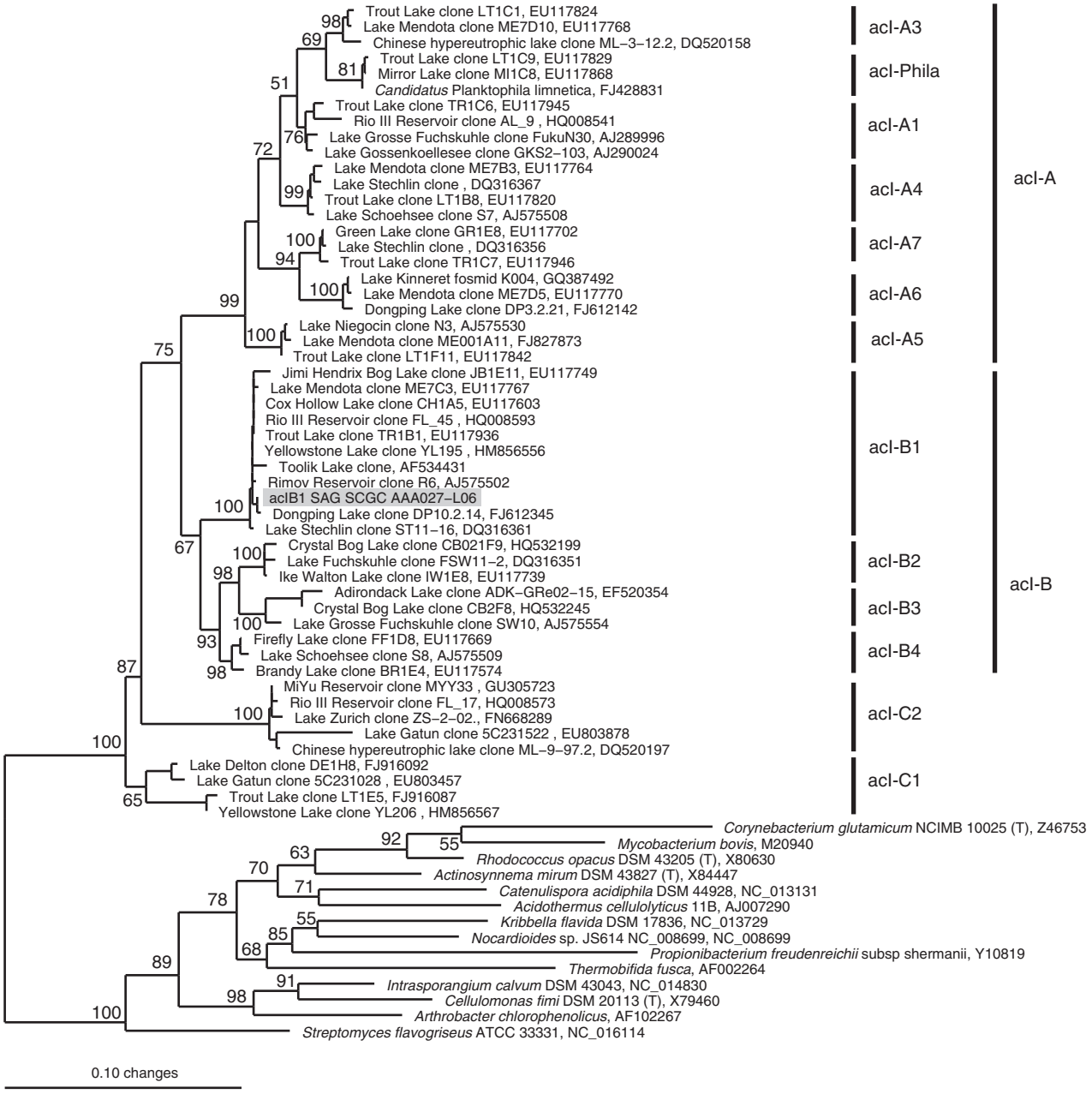

Figure 1 Phylogenetic placement of the acI-B1 AAA027-L06 SAG within the acI lineage and relative to other sequenced actinobacterial genomes. Phylogenetic reconstruction was conducted by maximum likelihood (RAxML) (Stamatakis *et al.*, 2008) with 1000 bootstrap runs on the CIPRES web portal (http://www.phylo.org) using near-full-length reference 16S rRNA gene sequences from a manually curated alignment (Newton *et al.*, 2011) and a 50% base frequency filter (total 1402 positions). Bootstrap values are indicated above nodes with greater than 50% support and the scale bar represents 10 base substitutions per 100 nucleotide positions.

out using Phylogenetic profilers. First, one of the 138 genomes was randomly selected as 'seed' to quantify single-copy genes. Then additional, randomly selected genomes were sequentially added, until all 138 genomes were included, and the number of core genes shared among the analyzed group of genomes, was quantified for each genome combination. The same process using the same 'seed' genome was done but this time including the SAG SCGC AAA027-L06 genome. A ratio between the shared genes including and not including the SAG gave the percentage of core genes present in SAG. The entire process was reiterated five times, using new, randomly selected 'seed' genomes. The ratio of the five replicates was averaged. The estimated core genome was 129 ± 49 genes.

Genome annotation and comparative analysis

Genes were identified using Prodigal (Hyatt *et al.*, 2010). The predicted CDSs were translated and used

to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG and InterPro databases. The tRNAScanSE tool (Lowe and Eddy, 1997) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (Pruesse *et al.*, 2007). Other non-coding RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (Nawrocki *et al.*, 2009) (Inference of RNA alignments. http://infernal.janelia.org). Additional gene-prediction analysis and manual functional annotation was performed within the IMG platform.

Genes encoding carbohydrate-active enzymes were automatically annotated using the CAZymes Analysis Toolkit applying the association rule-learning algorithm described in (Park *et al.*, 2010) and the resulting annotation was carefully revised by using conserved domain BLAST (Marchler-Bauer *et al.*, 2011), BLASTP (Camacho *et al.*, 2009) against NCBI's non-redundant protein database and the resources of SWISS-MODEL (Kiefer *et al.*, 2009), PROSITE (Sigrist *et al.*, 2010) and CAZy database (Cantarel *et al.*, 2009). Bioinformatic resources of the IMG system were used to estimate the frequency of *glycoside hydrolase* genes (E.C. 3.2.1.x; see CAZy database) in the publicly available prokaryote genomes. Frequency was calculated for each one of the prokaryote groups by dividing the total number of genes annotated as glycoside hydrolases by the total number of genes annotated for that specific group. Prediction of signal peptide was performed with SignalP 3.0 Server. Cleavage site is indicated at the N-terminus of the protein sequence, which is used to direct the protein through the cellular membrane (Petersen *et al.*, 2011). All bacterial genes annotated as chitinases in IMG were downloaded and used to build a custom database by using the option makeblastdb implemented in the version BLAST 2.2.22. Next, a BLASTP search was performed by using the putative chitinase (Locus Tag: A27L6_0005.00000540) detected in the sequenced acI SAG against the in-house chitinase genes database.

Scaffolds derived from the acI-B1 SAG were compared with a fosmid recovered from Lake Kinneret, Israel (K004, Accession GQ387492) (Philosof *et al.*, 2009) and to a scaffold assembled from metagenomics reads recovered from Lake Gatun, Panama during the Global Ocean Survey (JCVI_SCAF_1097207254344, downloaded from the CAMERA database) (Rusch *et al.*, 2007; Sun *et al.*, 2011). Both sequences were re-annotated within IMG/M-ER for comparison to the SAG draft genome. Preliminary analyses showed that both K004 and the Gatun scaffold mapped to scaffold A27L6_scaffold00006 (63 580 bp) in the acI-B1 SAG. BLASTP was used to compare CDSs extracted from A27L6_scaffold00006 and K004, to calculate average protein sequence identity and similarity. For the same purpose, TBLASTN was used to compare CDSs extracted from A27L6_scaffold00006 to translated JCVI_SCAF_1097207254344. In both cases, only hits spanning at least 70% of the query sequence were used to calculate percent similarity or identity. Synteny was examined by gapped TBLASTX using AB-BLAST (version 3.0PE, derived from WU-BLAST2, Advanced Biocomputing LLC, St Louis, MO, USA) for pairwise comparisons of A27L6_scaffold00006, K004, and JCVI_SCAF_1097207254344. The following settings were used: tblastx -matrix = BLOSUM45 -altscore = '* any −80' -altscore = 'any * −80' -hspmax = 5000 -Q = 20 -R = 7. The results were visualized using custom perl scripts. Only hits with greater than 40% identity and alignment length greater than 40 amino acids were considered further.

Results and discussion

Single-cell general genomic features

Rigorous quality controls were implemented to detect potential contaminants and amplification artifacts. Only 0.65% of all sequences reads were identified as contaminants, and were removed from further analysis. Sequencing and genome assembly resulted in 1.16 Mb sequence data distributed among 75 contigs of SAG SCGC AAA027-L06, with contig length ranging 0.5–85 kb (Table 1). All the genes from a survey of 35 widely conserved, single-copy genes present in most prokaryotic genomes were found (Raes *et al.*, 2007). Overall, 32 tRNA genes and 23 tRNA synthetase genes were identified with specificities for 18 out of the 20 canonical amino acids. All these data suggest that although the genome was not closed, most of the actinobacterial genome sequence was recovered. Based on the analysis of core genes for Actinobacteria, it was estimated that the 75 contigs account for about 97% of the genome (see Materials and methods for more details). Thus, the predicted genome size approximates 1.2 Mb, which is slightly smaller than that of the streamlined *Pelagibacter ubique* strain HTCC1062 genome (Giovannoni *et al.*, 2005).

Table 1 General features of the single-cell genome assembly for actinobacterium SCGC AAA027-L06

Assembly statistic	
Assembly size (Mb)	1.163
Estimated genome recovery (%)	97
List of 35 orthologous markers (Raes *et al.*, 2007)	35
tRNA synthetase	18
Number of contigs	75
Largest contig (kb)	85.45
GC content (%)	41.69
Number of total predicted genes	1282
Number of rRNA operons	3
Number of tRNA genes	32
Number of protein-coding genes	1244
Number of genes with no function prediction	196

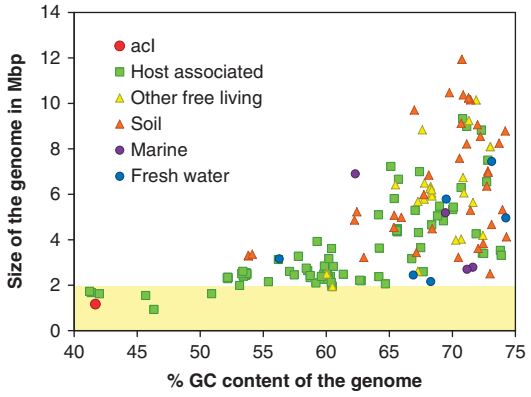

Figure 2 Distribution of 197 actinobacterial finished genomes according to GC content and genome size. The yellow area which is below 2 Mb contains only ~8% of the genomes, most of which are host associated.

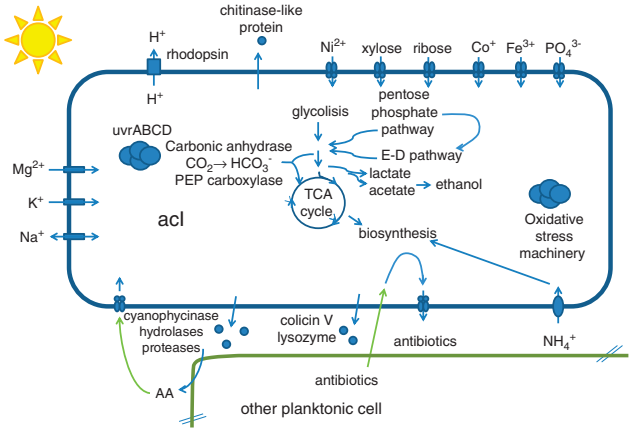

Figure 3 Physiology of acI as deduced from SAG sequencing.

When compared with other Actinobacteria, the acI-B1 draft genome is among the smallest: 92% of the 197 Actinobacteria finished genomes analyzed are larger than 2 Mb (Figure 2).

The average GC content of the actinobacterial SAG is 41.6% (Figure 2). This low GC content agrees with recently identified actinobacterial metagenomic sequences from lakes and estuaries (Ghai *et al.*, 2011a) and contradicts the traditional view of Actinobacteria as high G+C Gram-positive bacteria (Stackebrandt *et al.*, 1997). In total, 1282 genes were annotated in the SAG out of which 1244 are protein-coding genes. Approximately 84% of the genes are protein-coding genes with function prediction. For comparison, the total genes present in finished *Pelagibacter* strains are 1394 in HTCC1062 and 1482 in IMCC9063.

Primary metabolism

The main characteristics found in the predicted metabolism of the acI-B1 SAG relate perfectly to its preferred habitat (Figure 3). The SAG encodes glycolysis, pentose phosphate pathway, citrate cycle and oxidative phosphorylation, suggestive of aerobic respiration. Genes that enable pyruvate fermentation to lactate, acetate and ethanol are also present. This genetic evidence shows a potential versatility of the cell to switch between aerobic and anaerobic metabolisms depending on the oxygen availability. Being a facultative aerobic organism would be an advantage over strict aerobes or anaerobes during times of the year when the lake's oxygen increases or diminishes rapidly, such as spring and fall (stratification and mixing, respectively). Notably, active members of the acI-B clade have been detected in anoxic hypolimnia, providing further evidence for their ability to occupy lake habitats devoid of oxygen (Buck *et al.*, 2009).

Even though the SAG presents the full metabolic potential for performing glycolysis, glucose transporters have not been detected in the assembled reads. However, many genes encoding for ABC transporters of xylose, arabinose and/or ribose were found (Supplementary Table S2). Together with the transporters, all the enzymes for incorporation of ribose and xylose into the glycolysis pathway were accounted for (Figure 3). Previous studies suggest that members of the acI lineage reach their maximal abundances in late fall in temperate lakes (Allgaier and Grossart, 2006). Xylose is the second most abundant monosaccharide in terrestrial plants and it is especially prevalent in angiosperms. In woody angiosperms, D-xylose averages about 17% of the total dry weight, but in herbaceous angiosperms it can range up to 31% (Jeffries and Shi, 1999).

Genome sequence analysis also revealed that this SAG encodes a suit of carbohydrate-active enzymes (22 genes) such as glycoside hydrolases (Supplementary Figure S1A) involved in polysaccharide hydrolysis, which is a major bottleneck in the remineralization of high molecular weight dissolved organic matter (Arnosti, 2011). We found that the sequenced SAG contains an average proportion of bacterial and actinobacterial *glycoside hydrolase* genes (0.2% of total genes dedicated to polysaccharide hydrolysis) when compared with the 2990 publicly available prokaryote genomes (Supplementary Figure S1B). Interestingly, a putative chitinase-like protein was detected (Supplementary Table S3), containing the active site and the putative catalytic residues described for that family in EXPASy (InterPro), CAZy and Pfam databases as well as the signal peptide at the N-terminus (Supplementary Figure S2), which is used to direct the protein through the cellular membrane to the external surface, where the enzyme cleaves the polysaccharides (Arnosti, 2011; Petersen *et al.*, 2011). The closest annotated bacterial chitinase-like protein in IMG were those encoded by Firmicutes such as *Halothermothrix orenii* and *Clostridium acetobutylicum*, whereas the closest hit in the non-redundant database was that belonging to *Acidothermus cellulolyticus,* which is able to degrade chitin (Barabote *et al.*, 2009). Therefore,

in addition to the *N*-acetylglucosamine remineralization role (Beier and Bertilsson, 2011), the data indicates that acI-B1 may be also involved in the hydrolysis of chitin, a major polysaccharide in planktonic freshwater systems.

Thus, it seems that depending on the available substrate, acI-B1 may act as primary polysaccharide degraders in planktonic systems (that is, in the case of chitin) or as commensalist, taking up monosaccharides (that is, xylose) that are released during polysaccharide degradation by other heterotrophic microorganisms.

Looking at nitrogen and sulfur, the acI-B1 SAG lacks recognizable genes involved in the assimilation of sulfate, sulfite, nitrate or nitrite. However, the genome has cysteine synthase to incorporate hydrogen sulfide to form cysteine and several enzymes that incorporate ammonia into amino-acid production. An additional feature in the primary metabolism of the acI-B1 SAG is the presence of the enzymes carbonate anhydrase and phosphoenolpyruvate carboxylase. These enzymes have been found in other marine bacteria, where it was predicted that they may be involved in inorganic carbon fixation and were also associated with anaplerotic metabolism of cultivated marine photoheterotrophic bacteria (González *et al.*, 2008; Tang *et al.*, 2009).

Photometabolism

Microbial rhodopsins are transmembrane proteins that use retinal chromophores to harvest solar energy (Gómez-Consarnau *et al.*, 2010). These proteins were first discovered in the surface water of the ocean and much later in freshwater habitats (Sharma *et al.*, 2009). Recently, a study demonstrated that the acI-B1 lineage is one of the predominant freshwater rhodopsin-containing bacteria (Martinez-Garcia *et al.*, 2011b). We confirmed the presence of an *actinorhodopsin* gene and it was 99.4% identical at the nucleotide level to the gene fragment that had been amplified from the SAG before genomic sequencing (Genbank accession HQ663748.1) (Martinez-Garcia *et al.*, 2011b). As described by Martinez-Garcia *et al.* (2011b), this gene is closely related (with 95–98% sequence identity) at the amino-acid level to putative actinorhodpsin in *Candidatus Aquiluna rubra*, *Candidatus Rhodoluna planktonica*, *Candidatus Rhodoluna lacicola*, *actinobacterium* MWH-Uga1, *actinobacterium* MWH-EgelM2-3.D6 and bacteriorhodopsin of cyanobacterium *Gloeobacter violaceus* PCC 7421.

Interestingly, the acI-B1 SAG together with the previously analyzed marine bacteria harboring *rhodopsin* genes (González *et al.*, 2008; Woyke *et al.*, 2009) have more than one feature in common. First, all the genomes including the acI-B1 SAG are relatively small, that is, <2.9 Mb. Second, the previously analyzed genomes show the presence of *blh,* which is a gene encoding a monooxygenase that synthetizes retinal from β-carotene. In the case of the acI-B1 SAG, the presence of *blh* could not be confirmed with BLASTP. However, most of the genes of the *ctr*-gene cluster for β-carotene production were found, as well as an aromatic ring-cleaving dioxygenase (Supplementary Table S4). Although the ecological roles of microbial rhodopsins remain enigmatic, actinorhodopsin presence could be related to energy production in the presence of light (González *et al.*, 2008). It is clear that aquatic bacteria containing *rhodopsin* genes are abundantly distributed in their habitats, and this might be owing to a competitive advantage. Consistent with the findings of Martinez-Garcia *et al.* (2011b), we did not detect bacteriochlorophyll or the Calvin–Benson cycle genes, indicating that the analyzed SAG is a typical photoheterotrophic bacteria.

Stress resistance

Confirming previous hypotheses and observations (Warnecke *et al.*, 2004; Warnecke *et al.*, 2005; Debroas *et al.*, 2009), the acI-B1 SAG contains many genetic mechanisms involved in DNA repair (Supplementary Table S5). In total, more than 70 genes were found coding for replication, repair and recombination, representing 6.5% of the genes in the sequenced genome. On average, 6.3% of the genes in actinobacterial genomes code for replication, repair and recombination genes.

Living in surface waters could also mean that acI cells are continuously exposed to reactive oxygen molecules. Cellular damage caused by $^1O_2$ (Davies, 2005) needs to be repaired and the cellular redox homeostasis needs to be maintained during oxidative stress. Notably, members of the acI-B clade in a humic lake were particularly sensitive to $^1O_2$ exposure (Glaeser *et al.*, 2010). The acI-B1 SAG analyzed in this study contains several genes that encode for proteins responsible for oxidative and osmotic stress regulation (Supplementary Table S6). Among the several different mechanisms to cope with this type of stress, members of the peroxiredoxin family were found. These enzymes can reduce toxic peroxides with the help of reactive cysteine thiols and act as a potent cellular antioxidant (Dubbs and Mongkolsuk, 2007).

Many cultivated Actinobacteria have the capability of forming spores during unfavorable environmental conditions. Recently, a study based on freshwater metagenomics analysis of actinobacterial genomic fragments failed to detect the presence of sporulation genes (Ghai *et al.*, 2011a). The formation of spores by acI members in freshwater still remains unobserved in nature (Newton *et al.*, 2011), but in the acI-B1 SAG, six genes that might be related to sporulation formation were found (Supplementary Table S7). However the detection of these genes does not provide sufficient evidence for spore formation. These genes, as discussed in a previous publication (Traag *et al.*, 2010), are present in many microorganisms that do not form endospores, including species

outside of Firmicutes and Actinobacteria. Cultivation of acI or discovery of a novel sporulation mechanism would be necessary to determine if acI members can sporulate.

Transporters

Membrane transport proteins are particularly important for cells, as they are responsible for providing the cell with nutrients and for discarding toxic molecules. We found 118 cell membrane transporters, that is 101.7 transporters per Mb (Supplementary Table S2). Of all the transporters, 10% are specialized for drugs and antibiotics. The remaining transporters are mainly for amino acids, pentoses and ions. The number of transporters in the acI-B1 SAG is relatively small as compared with other aquatic bacteria, such as the marine *Roseobacter* cluster, which can harbor up to 330 transporters per genome, which are around 81.4 transporters per Mb (González *et al.*, 2008). The small number of transporters in SAG acI-B1 could be a reflection of the small genome size. As previously discussed by González *et al* (2008), all these facts together also point to a constrained cell metabolism and the high specialization of the lineage.

Other metabolic features

The acI-B1 SAG possess several genes encoding cell wall-associated hydrolases, colicin V production protein, lysozyme M1, hemolysins, predicted collagenase, related proteases and unsaturated glucuronyl hydrolase, and cyanophycinase. These genes could be involved in the repair and regeneration of the cell, but in a competitive environment or opportunistic scenario, acI members could have the capability of releasing lysozymes that can aid in lysing the cell walls of other microbes. Cyanophycinase is a hydrolytic enzyme that can release aspartic acid and arginine from cyanophycin (multi-L-arginyl-poly-L-aspartic acid), which is a storage compound in cyanobacteria. In the latter case, in the presence of lysed cyanobacterial cells, amino acids released from cyanophycin could be internalized by acI-B1 via ABC superfamily transporters and then metabolized for basal cellular metabolism (Richter *et al.*, 1999). This has important implications for niche specialization of acI-B1 during summer

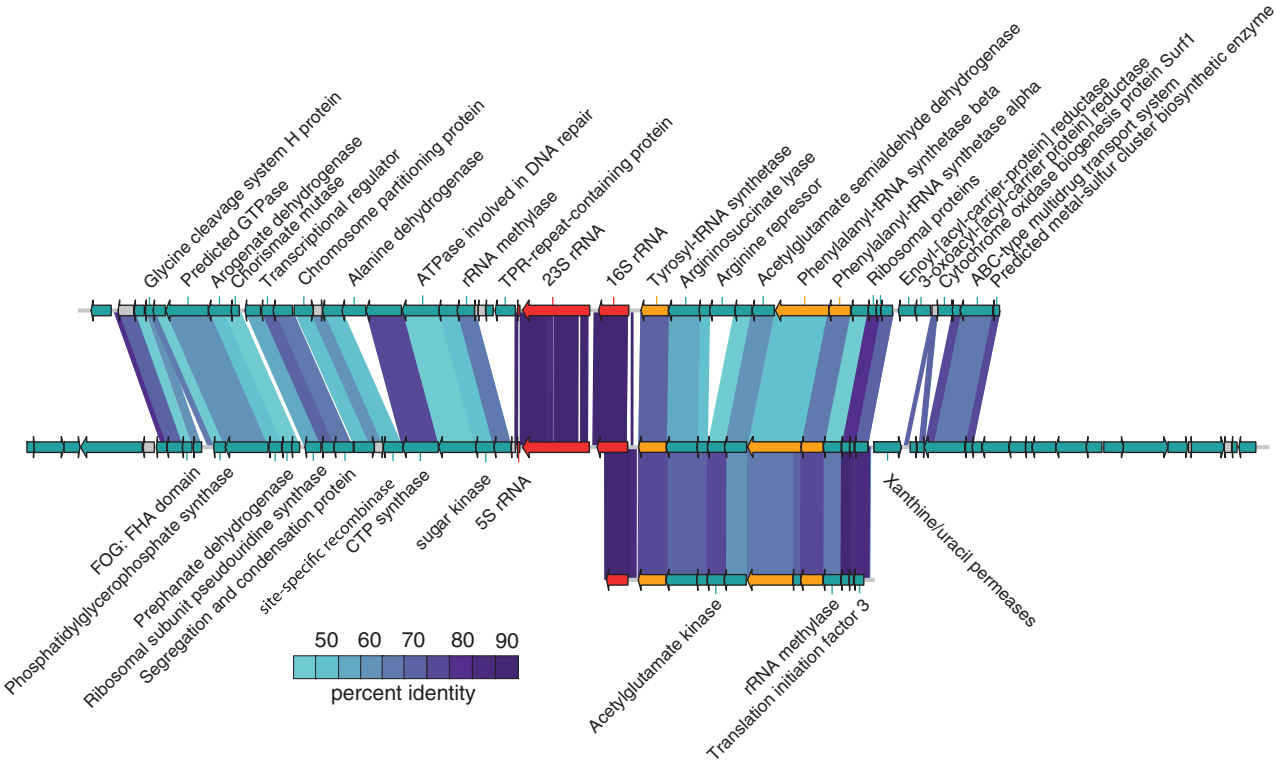

Figure 4 Synteny within acI genomes. The following scaffolds were compared using gapped TBLASTX to identify putative homologs: (top) acI-A6 fosmid K004 from Lake Kinneret (GQ387492); (middle) acI-B1 scaffold0006 from Lake Mendota SCGC AAA027-L06; (bottom) acI-B1 scaffold JCVI_SCAF_1097207254344 from Lake Gatun GOS20. The scaffolds from Lake Kinneret and Lake Gatun were re-annotated in the IMG pipeline to ensure consistency in gene calling and nomenclature. The blast hit color-coding represents percent protein sequence identity, except for the rRNA genes, which were analyzed separately using BLASTN. Open reading frames encoding tRNA synthetases are highlighted in orange. The alanine dehydrogenase that appears to be missing in the SAG was found elsewhere on another scaffold (see text), but none of the other regions unique to the Lake Kinneret fosmid were found in the SAG.

conditions in productive lakes, as decaying cyanobacterial blooms could serve as a significant carbon and energy source.

Conserved genome content and structure
Although this is the first near-complete genome to be sequenced within the acI lineage, two previous studies generated large genomic fragments that were clearly associated with acI. A fosmid library yielded a 40 818-bp scaffold (K004) from Lake Kinneret, Israel, that contained a 16S rRNA gene clearly affiliated with the acI-A6 tribe (Philosof *et al.*, 2009) and the Global Ocean Survey yielded one 12 359 bp scaffold that was assembled from metagenomics reads derived from Lake Gatun, Panama, containing a 16S rRNA gene clearly affiliated with the acI-B1 tribe (Rusch *et al.*, 2007; Ghai *et al.*, 2011a). We compared the draft acI-B1 SAG-predicted protein sequences with those from each of these two scaffolds and observed that the SAG shared 77% average protein sequence similarity (61% identity) with the acI-A6 scaffold from Lake Kinneret and 88% average protein sequence similarity (77% identity) with the acI-B1 scaffold from Lake Gatun. In both cases, gene order (synteny) was remarkably highly conserved, with no observed rearrangements and only a few deletions (Figure 4). Interestingly, a homolog to the alanine dehydrogenase that appears to be missing in the acI-B1 SAG scaffold was found elsewhere in the assembly, on A27L6_scaffold00012, with 68% protein sequence identity. The gene neighborhood around this open reading frame does not resemble the Lake Kinneret fosmid, suggesting a true gene order difference as opposed to an assembly error. None of the other regions present in the Lake Kinneret fosmid generated hits with more than 30% amino-acid identity within the acI-B1 SAG scaffolds. The relatively high degree of sequence similarity shared with a member of the acI-A clade and the observed synteny is particularly interesting and holds promise for comparative analysis of metagenomics data sets from samples containing diverse members of the acI lineage, using the acI-B1 SAG as a reference.

Conclusions

We present the first genomic analysis of a representative of the most abundant freshwater bacterioplankton lineage, the Actinobacteria cluster acI. A near-complete genome was recovered from an uncultured, individual cell that was collected from the eutrophic Lake Mendota in WI, USA. The genome suggests a heterotrophic, facultative aerobic metabolism. Additional speculated metabolic features include the potential for carbon fixation via anaplerotic pathways and actinorhodopsin-based photometabolism. The genome harbors genes for protection against stress from reactive oxygen species and for photorepair. The generality of the observed features remains to be tested, after sequencing additional genomes of the Actinobacteria cluster acI.

Acknowledgements

This work was supported by NSF grants DEB-841933 and OCE-821374 to RS. HPG and Abhishek Srivastava were supported by a grant given by the German Science foundation (DFG GR 1540/17-1). SLG and FW thank JSMC for funding and support. We also thank Drs Rohit Ghai and Francisco Rodriguez-Valera at the Universidad Miguel Hernandez, Alicante, Spain, for access to custom perl scripts. We thank Todd Miller for collecting the lake water sample used to recover the SAG sequence. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. KDM acknowledges funding from the United States National Science Foundation Microbial Observatories program (MCB-0702395), the Long Term Ecological Research program (NTL-LTER DEB-0822700), a CAREER award (CBET-0738309) and the Swedish Wenner-Gren Foundation.

References

Allgaier M, Brückner S, Jaspers E, Grossart H-P. (2007). Intra- and inter-lake variability of free-living and particle-associated Actinobacteria communities. *Environ Microbiol* **9**: 2728–2741.

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.

Arnosti C. (2011). Microbial extracellular enzymes and the marine carbon cycle. *Ann Rev Mar Sci* **3**: 401–425.

Barabote RD, Xie G, Leu DH, Normand P, Necsulea A, Daubin V *et al.* (2009). Complete genome of the cellulolytic thermophile Acidothermus cellulolyticus 11B provides insights into its ecophysiological and evolutionary adaptations. *Genome Res* **19**: 1033–1043.

Beier S, Bertilsson S. (2011). Uncoupling of chitinase activity and uptake of hydrolysis products in freshwater bacterioplankton. *Am Soc Limnol Oceanogr* **56**: 1179–1188.

Bogdanova E, Shagina I, Mudrik E, Ivanov I, Amon P, Vagner L *et al.* (2009). DSN depletion is a simple method to remove selected transcripts from cDNA populations. *Mol Biotechnol* **41**: 247–253.

Buck U, Grossart H-P, Amann R, Pernthaler J. (2009). Substrate incorporation patterns of bacterioplankton populations in stratified and mixed waters of a humic lake. *Environ Microbiol* **11**: 1854–1865.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinf* **10**: 421.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* **37**: D233–D238.

Davies MJ. (2005). The oxidative environment and protein damage. *Biochim Biophys Acta Prot* **1703**: 93–109.

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

Debroas D, Humbert J-F, Enault F, Bronner G, Faubladier M, Cornillot E. (2009). Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget – France). *Environ Microbiol* **11**: 2412–2424.

Dubbs JM, Mongkolsuk S. (2007). Peroxiredoxins in bacterial antioxidant defense. In: Flohé L, Harris JR (eds). *Peroxiredoxin Systems. Structures and Functions*. Springer: New York, USA, pp 143–193 (http://www.springerlink.com/content/h0l54kn511ku51q7/).

Eckert EM, Salcher MM, Posch T, Eugster B, Pernthaler J. (2011). Rapid successions affect microbial N-acetylglucosamine uptake patterns during a lacustrine spring phytoplankton bloom. *Environ Microbiol* **14**: 794–806.

Eiler A, Heinrich F, Bertilsson S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. *ISME J* **6**: 330–342.

Fleming EJ, Langdon AE, Martinez-Garcia M, Stepanauskas R, Poulton NJ, Masland EDP *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.

Ghai R, McMahon KD, Rodriguez-Valera F. (2011a). Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environ Microbiol Rep* **4**: 29–35.

Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, Cristina Souza de Oliveira T *et al.* (2011b). Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS ONE* **6**: e23785.

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

Glaeser SP, Grossart H-P, Glaeser J. (2010). Singlet oxygen, a neglected but important environmental factor: short-term and long-term effects on bacterioplankton composition in a humic lake. *Environ Microbiol* **12**: 3124–3136.

Glöckner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A *et al.* (2000). Comparative 16S rRNA analysis of Lake Bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of Actinobacteria. *Appl Environ Microbiol* **66**: 5053–5065.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108**: 1513–1518.

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

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

Haukka K, Kolmonen E, Hyder R, Hietala J, Vakkilainen K, Kairesalo T *et al.* (2006). Effect of nutrient loading on bacterioplankton community composition in Lake Mesocosms. *Microbiol Ecol* **51**: 137–146.

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.

Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, Hauser L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Jeffries T, Shi N-Q. (1999). Genetic engineering for improved xylose fermentation by yeasts recent progress in bioconversion of lignocellulosics. In: Tsao G, Brainard A, Bungay H, Cao N, Cen P, Chen Z *et al* (eds). *Recent Progress in Bioconversion of Lignocellulosics*. Springer: Berlin/Heidelberg, pp 117–161 (http://www.springerlink.com/content/tarvlh74 ea28lfw8/).

Jezbera J, Sharma AK, Brandt U, Doolittle WF, Hahn MW. (2009). 'Candidatus Planktophila limnetica', an actinobacterium representing one of the most numerically important taxa in freshwater bacterioplankton. *Int J Sys Evol Microbiol* **59**: 2864–2869.

Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. (2009). The SWISS-MODEL repository and associated resources. *Nucleic Acids Res* **37**: D387–D392.

Lindeman RL. (1942). The trophic-dynamic aspect of ecology. *Ecol* **23**: 399–417.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 0955–0964.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C *et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**: D225–D229.

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.

Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC. (2009). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**: 2271–2278.

Martinez-Garcia M, Brazel D, Poulton NJ, Swan BK, Gomez ML, Masland D *et al.* (2011a). Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *ISME J* **6**: 703–707.

Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME *et al.* (2011b). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**: 113–123.

Nawrocki EP, Kolbe DL, Eddy SR. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75**: 14–49.

Newton RJ, Jones SE, Helmus MR, McMahon KD. (2007). Phylogenetic ecology of the freshwater actinobacteria acI lineage. *Appl Environ Microbiol* **73**: 7169–7176.

Newton RJ, McMahon KD. (2011). Seasonal differences in bacterial community composition following nutrient

additions in a eutrophic lake. *Environ Microbiol* **13**: 887–899.
Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC. (2010). CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiol* **20**: 1574–1584.
Pernthaler J, Posch T, Šimek K, Vrba J, Pernthaler A, Glöckner FO *et al.* (2001). Predator-specific enrichment of actinobacteria from a cosmopolitan freshwater clade in mixed continuous culture. *Appl Environ Microbiol* **67**: 2145–2155.
Petersen TN, Brunak S, von Heijne G, Nielsen H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785–786.
Philosof A, Sabehi G, Béjà O. (2009). Comparative analyses of actinobacterial genomic fragments from Lake Kinneret. *Environ Microbiol* **11**: 3189–3200.
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. *Nucleic Acids Res* **35**: 7188–7196.
Raes J, Korbel J, Lercher M, von Mering C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.
Raghunathan A, Ferguson HR, Bornarth CJ, Song WM, Driscoll M, Lasken RS. (2005). Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71**: 3342–3347.
Richter R, Hejazi M, Kraft R, Ziegler K, Lockau W. (1999). Cyanophycinase, a peptidase degrading the cyanobacterial reserve material multi-L-arginyl-poly-L-aspartic acid (cyanophycin). *Eur J Biochem* **263**: 163–169.
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.
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.
Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A *et al.* (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* **38**: D161–D166.
Stackebrandt E, Rainey FA, Ward-Rainey NL. (1997). Proposal for a New Hierarchic Classification System, Actinobacteria classis nov. *Int J Sys Bacteriol* **47**: 479–491.
Stamatakis A, Hoover P, Rougemont J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Sys Biol* **57**: 758–771.
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.
Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–D551.
Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**: 1296–1300.
Tang K-H, Feng X, Tang YJ, Blankenship RE. (2009). Carbohydrate metabolism and carbon fixation in Roseobacter denitrificans OCh114. *PLoS One* **4**: e7233.
Traag BA, Driks A, Stragier P, Bitter W, Broussard G, Hatfull G *et al.* (2010). Do mycobacteria produce endospores? *Proc Natl Acad Sci USA* **107**: 878–881.
Warnecke F, Amann R, Pernthaler J. (2004). Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. *Environ Microbiol* **6**: 242–253.
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.
Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S *et al.* (2011). Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One* **6**: e26161.
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.
Zerbino DR, Birney E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.
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.

This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

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