SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
ENVIRONMENTAL MICROBIOLOGY
MICROBIAL ECOLOGY
MARINE MICROBIOLOGY
METAGENOMICS

Received
1 May 2013

Accepted
2 August 2013

Published
20 August 2013

Correspondence and requests for materials should be addressed to F.R.-V. (frvalera@umh.es)

Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria

Rohit Ghai[1], Carolina Megumi Mizuno[1], Antonio Picazo[2], Antonio Camacho[2] & Francisco Rodriguez-Valera[1]

[1]Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, San Juan de Alicante, 03550, Alicante, Spain, [2]Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia E-46100 Burjassot, Spain.

We describe a deep-branching lineage of marine Actinobacteria with very low GC content (33%) and the smallest free living cells described yet (cell volume ca. 0.013 μm^3), even smaller than the cosmopolitan marine photoheterotroph, '*Candidatus* Pelagibacter ubique'. These microbes are highly related to 16S rRNA sequences retrieved by PCR from the Pacific and Atlantic oceans 20 years ago. Metagenomic fosmids allowed a virtual genome reconstruction that also indicated very small genomes below 1 Mb. A new kind of rhodopsin was detected indicating a photoheterotrophic lifestyle. They are estimated to be ~4% of the total numbers of cells found at the site studied (the Mediterranean deep chlorophyll maximum) and similar numbers were estimated in all tropical and temperate photic zone metagenomes available. Their geographic distribution mirrors that of picocyanobacteria and there appears to be an association between these microbial groups. A new sub-class, '*Candidatus* Actinomarinidae' is proposed to designate these microbes.

Actinobacteria were considered to be typical soil dwellers. However, with the advent of the molecular approach, 16S rRNA genes indicative of actinobacterial descent were found in the ocean[1]. Later, more sequences retrieved from marine habitats could be more specifically connected to the cultivated actinobacterium *Candidatus* Microthrix parvicella and were designated as the OM1 clade[2–4]. Moreover, rRNA genes that were identified as Actinobacteria were also found in significant numbers in lakes and other freshwater habitats[5,6]. The diversity of freshwater Actinobacteria turned out to be very broad with several groups described based only on 16S rRNA analyses, distributed over two orders (Actinomycetales and Acidimicrobiales)[6–10]. Using fluorescence in situ hybridization (FISH) and examination of enrichment cultures it was concluded that these aquatic Actinobacteria were very small in size (biovolume <0.1 μm^3) and very abundant in oligotrophic freshwaters[7,8]. Recently, by using metagenomic approaches, aquatic Actinobacteria were shown to be low GC (mol% GC of genomic DNA 40–50%) compared to their high GC soil relatives[11,12]. The only other previously known low GC Actinobacteria were pathogens of the genus *Gardnerella*[11]. The higher surface:volume ratio of freshwater Actinobacteria likely improves their survival chances at the very low-nutrient concentrations found in oligotrophic freshwater bodies[13,14]. Two genomes of low-GC Actinobacteria are now available, one from a lake in Wisconsin, USA, determined using single-cell genomics[15] (GC content 42%) and another using a culture based approach[16] (GC 51.7%). Both of these organisms are photoheterotrophs, possessing rhodopsins (actinorhodopsins) to harvest light energy.

In addition, metagenomic studies, including the Global Ocean Sampling (GOS), KM3 station at the bathypelagic zone and the deep chlorophyll maximum (DCM) in the Mediterranean sea found sequences that could be classified as actinobacterial[17–19]. However, the absence of long scaffolds in which phylogenetically informative genes appear linked to significant fragments of their genomes has prevented a reliable assessment of their diversity, phylogenetic placement and genomic features.

Using a combination of metagenomics, flow cytometry and FISH, we describe here a widely distributed novel clade of marine Actinobacteria that have the lowest GC content reported so far as well as the smallest cells found among free-living prokaryotes. We propose the creation of a new sub-class '*Candidatus* Actinomarinidae' to denominate this group of microbes.

Results

Ribosomal rRNA phylogeny. The deep chlorophyll maximum (DCM) is a section of the photic zone water column, in stratified temperate or tropical oligotrophic ocean waters, where most of the photosynthetic activity takes place[17,20]. We have sequenced a large number of metagenomic fosmids from the Mediterranean DCM (MedDCM; see Methods). Fosmids provide discrete, natural contigs that can be efficiently assembled to obtain genomic fragments from all members in the community, even from those that are less prevalent and less accessible to direct sequencing. During a search for rRNA genes in the assembled contigs, we identified two nearly complete rRNA operons classified as actinobacterial by the 16S rRNA Ribosomal Database Project (RDP, http://rdp.cme.msu.edu) classifier[21] and the 23S rRNA SILVA large subunit (LSU, http://www.arb-silva.de) database[22] (MedDCM-OCT-S38-C68 and MedDCM-OCT-S40-C95). Surprisingly, the GC content of both of these rRNA containing contigs (33% and 32%), was far lower even than the recently described low GC freshwater Actinobacteria (GC% 42)[11,12,15]. Both contigs were syntenic to each other and showed high sequence similarity. Additionally, we identified another contig (MedDCM-OCT-S43-C55) (GC% 29.6) that overlapped with both rRNA-containing contigs, extending the reconstructed genomic fragment (Fig. 1a). A careful inspection indicated that the majority of genes in these contigs were similar to genes in actinobacterial genomes, providing additional evidence of their affiliation to this group.

We examined whether similar sequences had been assembled before by searching the 16S rRNA gene in the entire collection of assembled scaffolds from the GOS dataset[19]. This way, 13 GOS scaffolds were retrieved using a stringent cut-off of 98% nucleotide identity over 97% of the 16S rRNA gene sequence (species threshold

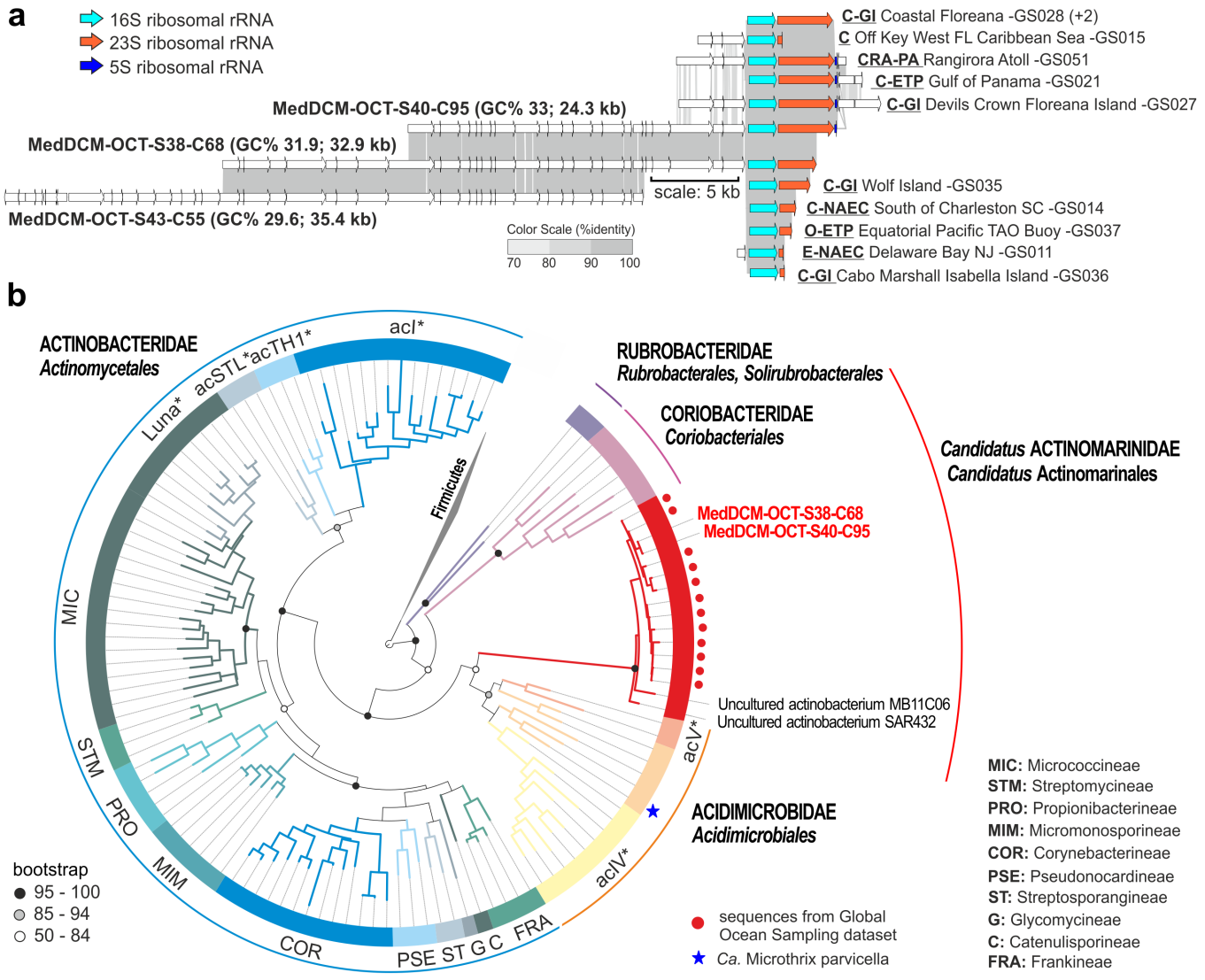

Figure 1 | (a), Comparison of marine low GC Actinobacterial contigs containing rRNA genes to scaffolds from the Global Ocean Sampling (GOS) dataset (using BLASTN). The oceanic habitat (C-Coastal, CRA-Coral Reef Atoll, O-Open Ocean, E-Estuary), sampling locations (NAEC: North American East Coast, GI-Galapagos Islands, ETP-Eastern Tropical Pacific, PA-Polynesia Archipelagos) and the GOS dataset identifier are shown next to each GOS scaffold. Numbers in brackets indicate additional identical sequences found at the same location. All ribosomal RNA genes are highlighted in color and sequence identity amongst the contigs is shown in shades of grey (see color scale). (b), 16S rRNA phylogeny. 16S rRNA gene sequences from the assembled contigs and GOS scaffolds in the context of the entire Actinobacteria phylum, with Firmicutes as the outgroup. Actinobacterial *Sub-Classes* are in bold uppercase and *Orders* in bold italics. Sub-orders are shown in different colors in the tree and labeled (key is shown on bottom right). Freshwater actinobacterial clades are additionally marked with an asterisk. '*Ca.* Microthrix parvicella', to which the Actinobacteria OM1 clade is related, is marked with a blue star. The novel branch with sequences attributed to sub-class '*Candidatus* Actinomarinidae' is shown in red. Bootstrap values (shown as percentages) for all major branches are shown in colored circles (see key bottom left).

level), and an additional 25 at >95% identity at 95% coverage. Even the comparison between the 16S–23S rRNA intergenic spacer region (ITS) of our contigs and those of the GOS indicated a high degree of conservation of these rRNA operons (Fig. S1). Most GOS scaffolds were short and contained only the rRNA operon, but some also presented a few more genes, which were remarkably syntenic to our contigs, although at a lower sequence identity (Fig. 1a). It is also interesting to note that the GOS scaffolds were all from temperate or tropical regions but geographically very distant from each other (e.g. Gulf of Panama, Equatorial Pacific). Moreover, we also found 250 sequences in 16S rRNA clone libraries[21] (%identity >98% and coverage >98% of complete gene). These results independently confirm the genuine nature of our assembled contigs and show that they originate from a widely distributed group of ultra-low GC Actinobacteria only known through their 16S rRNA sequences.

We generated maximum-likelihood trees for the alignments of 16S, 23S rRNAs and wherever possible to improve phylogenetic resolution, a concatenated alignment of both 16S and the 23S, in context of all known Actinobacteria (Fig. 1b, Fig. S2 and Fig. S3). All three analyses produced consistent results and unambiguously placed the rRNA sequences from the ultra-low GC Actinobacteria as a deep branching lineage, divergent enough to be a new subclass within the phylum. In one of the earliest studies using PCR amplification of the 16S rRNA gene performed in the Pacific and the Atlantic Oceans[1] a few deeply branching sequences belonging to Gram-positive bacteria were discovered, some of which were nearly identical to each other, even though they came from sampling sites that were quite far apart. This lineage was again recovered from the Sargasso sea, and described as the marine Actinobacterial clade[23]. Subsequent studies also confirmed the presence of another actinobacterial group (also referred to as Actinobacterial clade OM1) and estimated their abundance in the range of 1–5% of the total community[2,3]. Our analysis of all these short 16S rRNA sequences in the previous surveys indicates that these previously obtained sequences belong to two different groups. The Actinobacterial OM1 group has been previously recognized to be related to '*Candidatus* Microthrix parvicella'[2,3], and all the sequences in this group belong to the order Acidimicrobiales. However, sequences from the first two surveys[1,23] are related to the sequences retrieved by our metagenomic fosmids, and belong in an independent well defined clade. Therefore, with additional evidence of the complete 16S and the 23S genes at hand, we propose the creation of the new sub-class, '*Candidatus* Actinomarinidae', (order '*Ca.* Actinomarinales', sub-order '*Ca.* Actinomarineae', family '*Ca.* Actinomarinaceae') for the taxonomic placement of this group of microbes.

FISH hybridization and flow cytometry. As another completely independent way to verify the presence and abundance of these new Actinobacteria in the MedDCM, we used the 16S rRNA gene sequence to design a lineage-specific probe (LGC722; Table S1) and visualize them directly by FISH[24] (see Methods). The cells labeled with this probe were extremely small, even compared to *Prochlorococcus* cells that are less than ~1 μm in diameter (Fig. 2a–d). Image analysis indicates that the cells are probably spherical, and are among the smallest free-living marine microbes identified to date. Analysis of the size spectrum of bacterioplankton from MedDCM samples by combined flow cytometry-FISH techniques (Fig. 2e) gave biovolume estimations for the cells matching the lineage-specific probe ranging between 0.006–0.024 μm^3 (±SD 0.006 μm^3) and an average diameter of 0.292 μm (±SD 0.044 μm). Assuming a spherical shape, the average cell volume calculated was only ~0.013 μm^3. This extremely low biovolume is by far the lowest described for any planktonic prokaryote thus far (Table S2)[25–38]. In comparison, '*Candidatus* Pelagibacter ubique', considered the smallest autonomously replicating free-living cell, has a volume ranging from 0.019 to 0.039 μm^3 [37]. Microscopy abundance estimates from the fluorescently labeled cells indicated that they comprised nearly 4% of total bacterioplankton (~5×10^3 cell ml^{-1}) and represented ~80% of the cells hybridizing with a general actinobacterial probe (HGC236; Table S1). Given their extremely small size, we propose the taxonomic name '*Candidatus* Actinomarina minuta' for these microbes.

Genome reconstruction. For a better understanding of the lifestyle of the ultra-small Actinobacteria, we identified more assembled contigs from our MedDCM metagenomic fosmids that could belong to this group. In addition to the strict criteria employed for selection (see Methods), all contigs were manually examined. Moreover, a tight clustering of these contigs was revealed by principal component analysis (PCA) of tetranucleotide frequencies indicating that they likely belong to highly related microbes (probably at the level of the same genus; Fig. S4). This method of studying genomes retrieved from metagenomic datasets has been shown to work very well previously[13,39]. We were able to retrieve

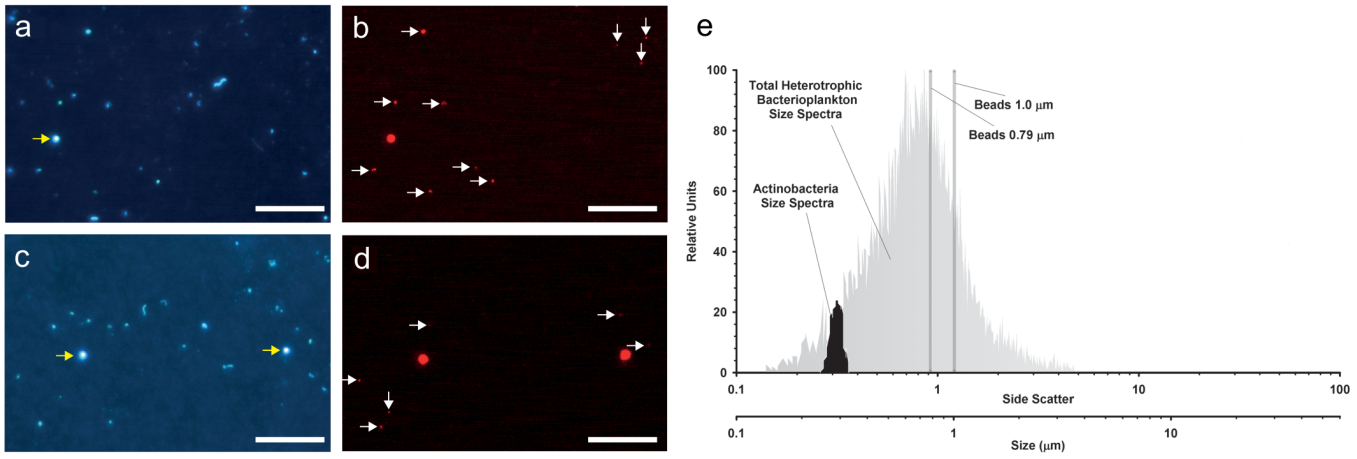

Figure 2 | (a–d), Microscopic fluorescence in situ hybridization (FISH) image of samples from the Mediterranean deep chlorophyll maximum (MedDCM). The micrographs show two pairs of identical microscopic fields, with samples stained with DAPI (left) and with the new lineage specific low GC Actinobacteria probe (LGC722) labeled with Cy3 (right). Yellow arrows (left) indicate autofluorescent *Prochlorococcus*, and white arrows (right) mark LGC722 signal also detected by DAPI. Bar: 10 μm (all four panels). (e), Abundance and bacterial structure size by flow cytometry. The size structure of the heterotrophic bacterioplankton population is shown. Size distribution of targeted Actinobacteria according to FISH measurements is shown in black. Note that the left tail of the size distribution is mostly due to instrumental noise and not due to bacterioplankton size.

43 contigs (longest 45.6 kb, shortest 7.3 kb, median GC 33.4%), which can be treated as a virtual (if incomplete) genome (Fig. 3a). We identified several overlapping contigs, but it is important to emphasize that a wide variation in the degree of relatedness was found among the overlaps (Fig. 3b). While some contigs were nearly identical at nucleotide level, others showed the average nucleotide identity expected for members of different species within a genus. Synteny was largely preserved in all cases of overlapping contigs, suggesting that multiple lineages of these microbes are present concurrently at the same location. The combined length of these 43 contigs is 1317 kb and once coalesced they span only ~700 kb (~800 genes). We analyzed the contigs for the presence of 35 orthologous markers defined previously[40] to estimate the completeness of the recovered virtual genome. Identification of 30 of these markers indicated 85% genome recovery. Another estimate using the core genes of all complete actinobacterial genomes suggested that 68% of the genome was recovered. Taken together, they result in an expected, but still remarkably small, genome size in the range of 823–1029 kb (Fig. 3a). Moreover, the median length of intergenic spacers was 3 bp comparable only to '*Candidatus* Pelagibacter ubique'[41], confirming a highly streamlined genome (Fig. S5).

Comparison of the reconstructed genome with the only sequenced freshwater low-GC actinobacterial (acI cluster) genome[15] did not show any conserved synteny. However, they shared 418 orthologous genes (albeit with low average similarities, ~57%), a remarkably high proportion considering how phylogenetically distant the two microbes are (Fig. 1b). There were also a number of surprising parallels between the two genomes. Both microbes are putative photoheterotrophs containing rhodopsins. Rhodopsins are known to be important for light-harvesting in the photic zone of all aquatic environments[42,43]. We identified two rhodopsin-containing contigs (Fig. 3c) and retrieved 27 additional sequences (%similarity >95%, gene coverage 90%) from the GOS dataset (14 in scaffolds and 13 in metagenomic reads) (Fig. S6). These rhodopsins are distantly related to all other rhodopsins known so far, forming a novel branch in the phylogenetic tree (Fig. 3d). We suggest the name MACrhodopsins (**M**arine **a**ctinobacterial **c**lade rhodopsins) for this new clade. It is quite likely that these rhodopsins are used as a supplementary energy source to their main chemoheterotrophic metabolism as shown for other marine microbes[37,44]. The rhodopsin flanking genes in these metagenomic contigs were also conserved, a photolyase, common in organisms exposed to light, and a thiol-disulfide reductase, also linked to the rhodopsin gene in '*Candidatus* Pelagibacter'[41] (Fig. 3c

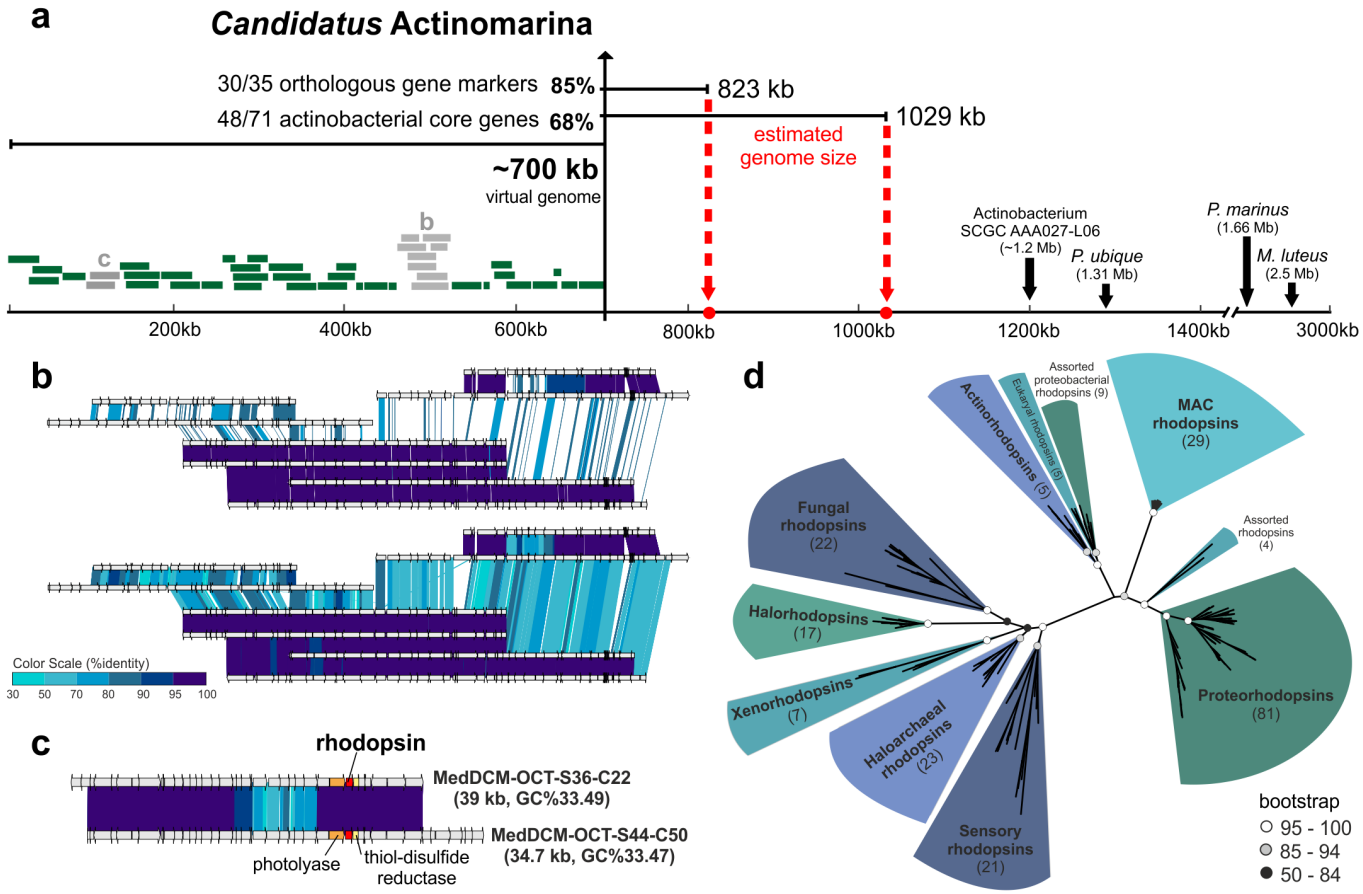

Figure 3 | (a), Linear representation of '*Candidatus* Actinomarina' contigs showing their overlaps. Estimates of the genome size based on different indicators are shown to the right with some reference small genome sizes. Two groups of contigs are highlighted in grey and are shown in greater detail in the panels below. (b), Multiple, highly related lineages. A group of contigs with overlaps indicating nucleotide identity (BLASTN, top) and translated protein identity (TBLASTX, below). A color scale is shown below. (c), Synteny amongst two rhodopsin containing contigs. The rhodopsin gene is shown in red. Overlaps are colored according to the color scale as shown (comparison performed with TBLASTX). (d), Marine Actinobacterial Clade Rhodopsins (MACrhodopsins). A maximum likelihood tree of all known types of rhodopsins is shown. The number of sequences in each clade of rhodopsins is indicated in brackets. 29 sequences from several Global Ocean Sampling (GOS) datasets were also identified using the novel sequences from the Mediterranean deep chlorophyll maximum and are part of the MACrhodopsin clade. Bootstrap values (shown as percentages) are indicated by circles (see key on bottom right).

and Fig. S6). Analysis of the critical amino acids determining wavelength selection for light absorption[43] indicated that they absorb light in the green region of the visible spectrum. Green tuned rhodopsins are correlated with highly-productive marine environments[45], such as coastal waters and the DCM. Genes involved in beta-carotene biosynthesis, e.g. geranylgeranyl diphosphate (GGPP) synthase and geranylgeranyl diphosphate reductase, were also found. Another interesting parallel with the acI genome available was the presence of a cyanophycinase. Cyanophycin is an amino acid polymer used as carbon and nitrogen storage material by several *Cyanobacteria* e.g. *Synechococcus*[46].

Other general metabolic pathways associated with aerobic life were shared by the two microbes such as several components of the TCA cycle, glycolysis, pentose phosphate pathway, superoxide dismutase and cytochrome *c*. No flagellar genes were present in either genome. Some other actinobacterial specific genes, e.g. for mycothiol biosynthesis and coenzyme F420-dependent enzymes, were also present in both genomes. On the other hand, some specific marine adaptations were found in '*Ca.* Actinomarina', including a phosphotransferase sugar transport system (PTS). PTS systems can transport several sugars, as well as *N*-acetyl glucosamine[47], which is widely available in the sea. Also consistent with the marine habitat was the presence of several Na^+ symporters (Na^+/H^+, Na^+/bile acid, Na^+/phosphate) and operons for the uptake of phosphate and phosphonate (the Mediterranean sea being a phosphate-limited habitat).

Biogeography and ecology. We examined the worldwide distribution of '*Ca.* Actinomarina' using the 16S rRNA as a probe in several metagenomic datasets and also in the entire Ribosomal Database Project (RDP)[21] (see Methods) using extremely stringent cut-offs (Fig. 4a, Fig. S7). It appears that the representatives of this group are widely distributed in the photic zone of the ocean, both in the tropical and temperate belt, not unlike the distribution of picocyanobacteria, particularly *Synechococcus*[48]. This distribution is also well supported by the high number of reads recruited at very high similarity at both central North Pacific and North Atlantic gyres (Hawaii Ocean Time Series-HOTS and Bermuda Atlantic Time Series-BATS metagenomes[49,50]) (Fig. 4b). However, like the picocyanobacteria, they are prominently absent from polar regions and from meso or bathypelagic depths (Fig. 4a and Fig. S7). Further evidence of their preferential abundance in the photic zone is seen in HOTS and BATS metagenomic depth-profiles reinforcing their absence in deeper waters (Fig. 4c). The abundance of '*Ca.* Actinomarina' along the depth profile remarkably mirrors that of *Synechococcus*. Along these lines, *Synechococcus* is known to produce cyanophycin[46] while in

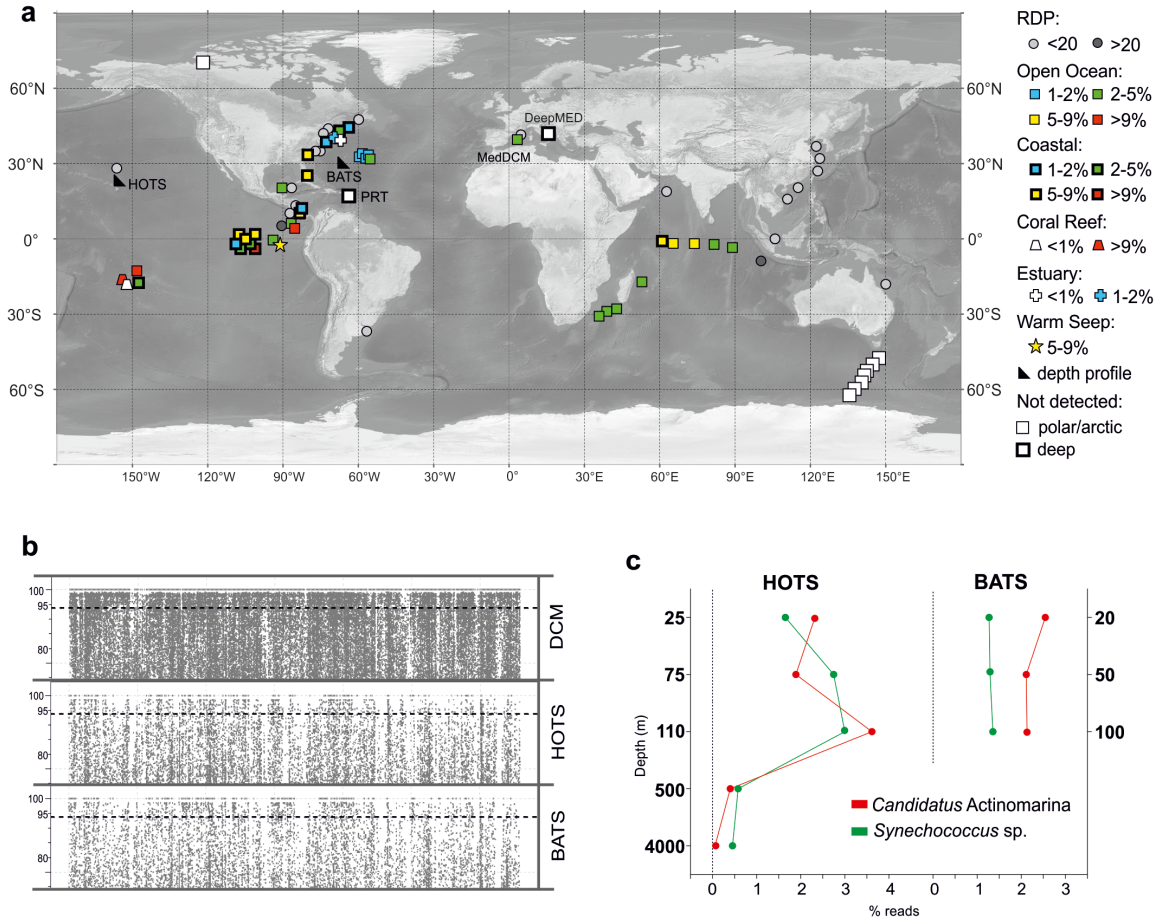

Figure 4 | (a), Worldwide distribution of 16S ribosomal rRNA of '*Candidatus* Actinomarina'. Several metagenomes and the Ribosomal Database Project (RDP) database were examined. Locations where the 16S rRNA gene of '*Candidatus* Actinomarina' was detected in the RDP database (%identity >98% and coverage >98% of complete gene) are shown in circles shaded according to the number of sequences (see key on the right). The number of reads detected in several metagenomes (GOS Open Ocean, Coastal, Coral Reef, Estuary, Warm Seep) are shown in percentages of total rRNA reads (%identity >98% and coverage 98% of metagenomic read) (see key on the right). Also shown (in white squares) are locations where no reads were detected. The world map shown here is a modified version of a freely available map made with Natural Earth at www.naturalearthdata.com. (b), Fragment recruitment. Metagenomic reads recruited (TBLASTX) by the '*Candidatus* Actinomarina' contigs in three metagenomes, the Mediterranean deep chlorophyll maximum (DCM), BATS and HOTS. (c), Depth profile. Percentage of metagenomic reads assigned to '*Candidatus* Actinomarina' genome in a depth profile of the HOTS and the BATS stations in comparison to *Synechococcus*.

Prochlorococcus this storage material seems to be absent as our search for cyanophycin-synthetase in all available *Prochlococcus* genomes did not reveal any such gene . The presence of the cyanophycinase gene also supports the *Synechococcus*-Actinomarina connection.

Discussion

The existence of new groups of aquatic Actinobacteria has been known for some time, but the difficulty in isolating these microbes in pure culture has hampered the advancement of knowledge about them. Single cell genomics has been used to describe the genome of one acI representative[15]. Here we have used metagenomic fosmids to partially reconstruct the genomes of uncultured marine Actinobacteria. The reconstruction of genomes from metagenomes is extremely unreliable mostly due to the high intraspecies diversity that is characteristic of most prokaryotes. Similar observations have been made for the recently described Group II Euryarchaeota virtual genome assembled from metagenomic data[51]. However, the large contigs provided by fosmids allow the inference of many properties of the microbes represented by them. The access to complete rRNA operons has allowed a refined phylogenetic placement of the microbes and the proposal of a new taxon at the subclass level. Besides, complete sequences allowed the development of FISH probes that provided independent confirmation of the presence and abundance of these microbes at a typical off-shore marine habitat. The DCM is one of the most characteristic ecological features of the stratified marine water column representing the most productive segment of the photic zone.

The actinobacterial cells characterized here are among the smallest free living cells described to date and fit very well with the characteristics of the typical photoheterotrophic cells that inhabit the pelagic niche of the oligotrophic ocean. The highly streamlined genome and the presence of rhodopsins that allow the cells a photoheterotrophic metabolism are common characteristics of the typical inhabitants of this niche.

Thus far, all the abundant aquatic Actinobacteria found appear to belong to two orders, the Acidimicrobiales, found mostly in freshwater but also in marine habitats (this is the most probable affiliation of the OM1 clade) or the '*Ca.* Actinomarinales'. Further work of genome reconstruction coupled to single cell genomics or (ideally) to the retrieval in pure culture of one or more representatives will allow a better understanding of this remarkable group of marine prokaryotes, which considering their widespread presence might have an important role in the global carbon cycle.

Methods

Sequencing, assembly and annotation. DNA from ~6000 fosmids (each fosmid ~40 kb) was extracted and pooled in 24 batches, with ~250 fosmids in each batch. These were sequenced using Illumina PE 300 bp reads (HiSeq 2000, Macrogen, South Korea) in a single lane (total output 42 Gb) which was expected to provide nearly ~175× coverage for each fosmid. Sequences were quality trimmed and vector sequences were clipped. Assembly was performed separately for each batch using Velvet[52] and gene predictions on the assembled fosmids were done using Prodigal in metagenomic mode[53] and tRNAs were predicted using tRNAscan-SE[54]. Ribosomal genes were identified using ssu-align[55] and meta_rna[56]. Functional annotation was performed by comparison of predicted protein sequences against the NCBI NR database (available from ftp://ftp.ncbi.nih.gov/blast/db/) and domain predictions for the fosmids described in this work were performed manually using NCBI-CD search[57] and the HHpred server[58]. Local BLAST searches against the latest NCBI-NR database were performed whenever necessary. Tetranucleotide frequencies were computed using the wordfreq program in the EMBOSS package[59]. Principal components analysis was performed using the FactoMineR package in R[60].

Phylogenetic analysis. Reference 16S rRNA sequences for all major actinobacterial lineages defined using 178 type strains, all known lineages of uncultured freshwater Actinobacteria (72 sequences), the closest BLAST hits to the Mediterranean actinobacterial sequences to the RDP database (available from http://rdp.cme.msu.edu/) (27 sequences) and the GOS dataset (available from http://camera.calit2.net/) (255 sequences) were collected to examine the phylogenetic relatedness of the low GC actinobacterial sequences. All sequences were screened and trimmed using ssu-align[55]. Only sequences more than 800 bp in length were retained. Sequences were aligned using MUSCLE[61] and a maximum likelihood tree was constructed using FastTree2[62] using GTR + CAT model and a gamma approximation. Bootstrapping (1000 bootstraps) was done using the seqboot program in the PHYLIP package[63]. Assembled site-specific GOS scaffolds were screened for the presence of 16S genes and a stringent cut-off of >98% identity and >800 bp length was used to select scaffolds that belonged to the same lineage as the Mediterranean actinobacterial 16S sequences assembled from the fosmids. In addition, alignments were constructed using 16S rRNA secondary structure aware ssu-align[55] and phylogenetic trees were reconstructed. Similar results were obtained as above. For the rhodopsin tree, sequences were selected based on existing literature, PFAM domain searches, and BLAST searches against NCBI-NR and the GOS dataset metagenomic reads. Sequences were aligned using MUSCLE[61] and a maximum likelihood tree was constructed with RAxML[64], using a JTT model a gamma approximation with 100 rapid bootstrap inferences.

Proteome comparison to freshwater Actinobacteria. Owing to the occurrence of several overlaps in the 43 actinobacterial contigs, some genes were represented more than once. Prior to comparison with the acI genome, the 1452 proteins from the 43 actinobacterial contigs were clustered using USEARCH[65] at 90% identity. The clustering resulted in a smaller dataset of 1177 proteins, representing a non-redundant proteome of the marine Actinobacteria. This set was compared to the 1244 proteins of the acI genome using a reciprocal best blast hit analysis to identify orthologs. Of these 1177 marine actinobacterial genes, 418 genes were found to be orthologous to the freshwater actinobacterial genes.

Genome size estimation. Genome size was estimated by two methods. First, a set of previously described 35 orthologous gene markers[40] was used. We were able to identify 30 of these genes in the 43 contigs. This suggests that the genome was 85% complete. In the second method, 4203 TIGRFAMs (available from ftp://ftp.jcvi.org/pub/data/TIGRFAMs/) were searched in all known complete actinobacterial genomes (n = 232). A set of 71 TIGRFAMs was identified in all known Actinobacteria, forming a core set of genes. This core set of genes was tested against the nearly complete genome of the freshwater actinobacterium SCGC AAA027-L06, which was estimated to be 97.5% complete by using 138 complete actinobacterial genomes. We found 69 core TIGRFAMs in this genome, providing an estimate of 97.1%, consistent with the previous estimate. The 43 contigs of '*Ca.* Actinomarina' contained 48 core TIGRFAMs, indicating that 67.6% of the genome was recovered.

Metagenomic recruitment. Recruitments were performed using TBLASTX[66], and a hit was considered only when it was at least 50 amino acids (aa) long with an e-value <= 1e − 5. For estimating the abundance of '*Candidatus* Actinomarina', *Synechococcus*, *Prochlorococcus* and '*Candidatus* Pelagibacter' in the HOTS (25 m, 75 m, 110 m, 500 m, 4000 m) and BATS (20 m,50 m,100 m) datasets of depth profiles, the entire metagenomic datasets (for each depth) were compared to a customised NR protein database to which the '*Ca.* Actinomarina' proteins were added (BLASTX). Only the best hits with an evalue <= 1e − 5 and at least 50 aa length were considered towards the calculations of abundance for each taxon.

16S ribosomal rRNA search across metagenomic datasets. The complete 16S rRNA gene sequence of '*Ca.* Actinomarina' was used as a probe to identify related sequences across several marine metagenomic datasets e.g. the GOS dataset[19], the Mediterranean DCM dataset[17], Arctic Metagenome (NCBI SRA accession ERR071289), Puerto Rico Trench Metagenome[67], Antarctica transect metagenome[68], HOTS datasets[50], and BATS datasets[49]. In addition, the entire RDP[21] was also searched to identify previously sequenced relatives. 16S rRNA gene sequences of all sequenced *Prochlorococcus*, *Synechococcus* and '*Ca.*Pelagibacter' genomes were used as controls.

16S ribosomal rRNA comparison with known marine actinobacterial sequences. All short 16S rRNA gene sequences described previously in surveys of actinobacterial diversity[1–3,23] were obtained from GenBank and were aligned to the reference actinobacterial 16S alignment using a phylogeny aware read-alignment[69] and placement on the reference actinobacterial tree using an evolutionary placement algorithm[70]. Moreover, sequence identities to the reference sequences indicated that the Actinobacterial OM1 clade always had >95% identity along their entire length to sequences belonging to the order Acidimicrobiales.

FISH and bacterial size structure. For microscopic counts of autotrophic picoplankton and heterotrophic bacterioplankton, water samples were fixed with a paraformaldehyde: glutaraldehyde solution to a final concentration (w/v) in the sample of 1% : 0.05% (w/v)[71]. Once in the laboratory, subsamples of 5–10 ml were filtered through 0.2 μm pore size black filters (Nuclepore™)(Whatman) at low pressure (<100 mbar). For the autotrophic picoplankton (0.2–2.0 μm), a quarter of a filter was directly inspected under an inverted Zeiss III RS epifluorescence microscope (1250×, resolution 0.02857 μm/pixel) (Zeiss), and cells classified as prokaryotes or photosynthetic eukaryotes depending on their autofluorescence characteristics, shape, cell size and the presence of chloroplasts. For heterotrophic bacterioplankton quantification was made on another quarter of the filter that was stained with 4′, 6-diamidino-2-phenylindole (DAPI)[72] (Sigma) and counted with the same microscope (1250×). Autofluorescence and DAPI-generated fluorescence were determined by using a standard filter set for green and blue light excitation[73].

For FISH detection of Actinobacteria, water samples were fixed with a paraformaldehyde 4% 1 : 1 to 2% final concentration and filtered within the next two hours. We used a general probe HGC236[7] (we discarded HGC664 and HGC840 for

the high mismatch with our Actinobacteria) and a new probe specifically designed for the targeted low GC Actinobacteria (Supplementary Table S1). For the design of the specific probes the Primer3 tool was used[74]. Four different oligonucleotide probes were constructed and tested; only LGC722 was used after checking for its specificity with the RDP[21]. All probes used were labeled with the indocarbocyanine dye Cy3 (Thermo Scientific, Waltham, MA, USA). FISH was performed on white polycarbonate filter (0.2 μm) sections with the different oligonucleotide probes, also stained with DAPI, and mounted for microscopic evaluation. The protocol was performed as described in Sekar *et al.*[75]. Hybridization conditions for the probe LGC722 were adjusted by formamide (VWR BDH Prolabo) series applied to different subsamples. A minimum of 500 DAPI and probe-stained cells were measured per sample in an inverted Zeiss III RS epifluorescence microscope with the adequate set of filters. Absolute densities of hybridized bacteria were calculated as the product of their relative abundances on filter sections (percentage of DAPI-stained objects) and the DAPI-stained direct cell counts. Images from FISH were analyzed using NIH ImageJ Software to determine cell dimensions for a minimum of 500 cells (http://rsb.info.nih.gov/ij/index.html). The biovolume of coccoid Actinobacteria was calculated as a sphere.

For cytometric identification, quantification and size structure approximation[76] of the bacterioplankton and autotrophic picoplankton (APP) cells, a Coulter Cytomics FC500 flow cytometer (Brea, California, USA) equipped with an argon laser (488 excitation), a red emitting diode (635 excitation), and five filters for fluorescent emission (FL1–FL5), was used. Bacterioplankton abundance and size structure was determined with argon laser by green fluorescence (Sybr Green I, Sigma-Aldrich, Missouri, USA) using a FL1 detector (525 nm). APP abundance was determined by combining the argon laser and red diode with red fluorescence (Chlorophyll a and phycobiliproteins autofluorescence) using a FL4 detector (675 nm). For size calibration, beads (polystyrene fluorospheres) of different sizes were measured (0.79 μm, 1 μm, 4.9 μm and 10 μm). In addition, *Prochlorococcus* cells were also used as controls. The lower and upper size limits of measurement are 0.25 μm to 40 μm respectively. The measured diameter of '*Ca.* Actinomarina' cells is 0.29 μm, which is at the lower end of the scale.

1. Fuhrman, J., McCallum, K. & Davis, A. Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Applied and Environmental Microbiology* **59**, 1294–1302 (1993).
2. Morris, R. M., Frazar, C. D. & Carlson, C. A. Basin-scale patterns in the abundance of SAR11 subclades, marine Actinobacteria (OM1), members of the Roseobacter clade and OCS116 in the South Atlantic. *Environmental Microbiology* **14**, 1133–1144 (2012).
3. Morris, R. M. *et al.* Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnol Oceanogr* **50**, 1687–1696 (2005).
4. Treusch, A. H. *et al.* Seasonality and vertical structure of microbial communities in an ocean gyre. *The ISME Journal* **3**, 1148–1163 (2009).
5. Warnecke, F., Amann, R. & Pernthaler, J. Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. *Environ Microbiol* **6**, 242–253 (2004).
6. Hahn, M. W. Description of seven candidate species affiliated with the phylum Actinobacteria, representing planktonic freshwater bacteria. *Int J Syst Evol Microbiol* **59**, 112–117 (2009).
7. Glockner, F. O. *et al.* Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl Environ Microbiol* **66**, 5053–5065 (2000).
8. Hahn, M. W. *et al.* Isolation of novel ultramicrobacteria classified as actinobacteria from five freshwater habitats in Europe and Asia. *Appl Environ Microbiol* **69**, 1442–1451 (2003).
9. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews* **75**, 14–49 (2011).
10. Newton, R. J., Jones, S. E., Helmus, M. R. & McMahon, K. D. Phylogenetic ecology of the freshwater Actinobacteria acI lineage. *Appl Environ Microbiol* **73**, 7169–7176 (2007).
11. Ghai, R., McMahon, K. D. & Rodriguez-Valera, F. Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environmental Microbiology Reports* **4**, 29–35 (2012).
12. Ghai, R. *et al.* Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6**, e23785 (2011).
13. Ghai, R. *et al.* Metagenomes of Mediterranean coastal lagoons. *Sci. Rep.* **2**, 490 (2012).
14. Poindexter, J. Oligotrophy Fast and Famine Existence. (1981).
15. Garcia, S. L. *et al.* Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. *The ISME Journal* (2012).
16. Kang, I. *et al.* Genome Sequence of "Candidatus Aquiluna" sp. Strain IMCC13023, a Marine Member of the Actinobacteria Isolated from an Arctic Fjord. *Journal of Bacteriology* **194**, 3550–3551 (2012).
17. Ghai, R. *et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *The ISME Journal* **4**, 1154–1166 (2010).
18. Martin-Cuadrado, A. B. *et al.* Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One* **2**, e914 (2007).
19. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**, e77 (2007).
20. Estrada, M., Henriksen, P., Gasol, J. M., Casamayor, E. O. & Pedros-Alio, C. Diversity of planktonic photoautotrophic microorganisms along a salinity gradient as depicted by microscopy, flow cytometry, pigment analysis and DNA-based methods. *FEMS Microbiol Ecol* **49**, 281–293 (2004).
21. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141–145 (2009).
22. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**, 7188–7196 (2007).
23. Rappé, M. S., Gordon, D. A., Vergin, K. L. & Giovannoni, S. J. Phylogeny of actinobacteria small subunit (SSU) rRNA gene clones recovered from marine bacterioplankton. *Systematic and Applied Microbiology* **22**, 106–112 (1999).
24. Glöckner, F. O. *et al.* Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology* **66**, 5053–5065 (2000).
25. Fagerbakke, K. M., Heldal, M. & Norland, S. Content of carbon, nitrogen, oxygen, sulfur and phosphorus in native aquatic and cultured bacteria. *Aquat Microb Ecol* **10**, 15–27 (1996).
26. Felip, M., Andreatta, S., Sommaruga, R., Straskrábová, V. & Catalan, J. Suitability of flow cytometry for estimating bacterial biovolume in natural plankton samples: comparison with microscopy data. *Applied and Environmental Microbiology* **73**, 4508–4514 (2007).
27. La Ferla, R. & Leonardi, M. Ecological implications of biomass and morphotype variations of bacterioplankton: an example in a coastal zone of the Northern Adriatic Sea (Mediterranean). *Marine Ecology* **26**, 82–88 (2005).
28. Lee, S. & Fuhrman, J. A. Relationships between biovolume and biomass of naturally derived marine bacterioplankton. *Applied and Environmental Microbiology* **53**, 1298–1303 (1987).
29. Loferer-Kröβbacher, M., Klima, J. & Psenner, R. Determination of bacterial cell dry mass by transmission electron microscopy and densitometric image analysis. *Applied and Environmental Microbiology* **64**, 688–694 (1998).
30. Malmstrom, R. R., Cottrell, M. T., Elifantz, H. & Kirchman, D. L. Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the Northwest Atlantic Ocean. *Applied and Environmental Microbiology* **71**, 2979–2986 (2005).
31. Nicastro, D. *et al.* Three-dimensional structure of the tiny bacterium Pelagibacter ubique studied by cryo-electron tomography. *Microscopy and Microanalysis* **12**, 180–181 (2006).
32. Norland, S. in *Aquat Microb Ecol* (eds Kemp, P. F., Sherr, B. F., Cole, J. J. & Sherr, E. B.) 303–307 (Lewis Publishers, 1993).
33. Norland, S., Heldal, M. & Tumyr, O. On the relation between dry matter and volume of bacteria. *Microbial Ecology* **13**, 95–101 (1987).
34. Posch, T. *et al.* Precision of bacterioplankton biomass determination: a comparison of two fluorescent dyes, and of allometric and linear volume-to-carbon conversion factors. *Aquat Microb Ecol* **25**, 55–63 (2001).
35. Salcher, M. M., Pernthaler, J. & Posch, T. Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limnology and Oceanography* **55**, 846 (2010).
36. Salcher, M. M., Pernthaler, J. & Posch, T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves'(LD12). *The ISME Journal* **5**, 1242–1252 (2011).
37. Steindler, L., Schwalbach, M. S., Smith, D. P., Chan, F. & Giovannoni, S. J. Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. *PLoS One* **6**, e19725 (2011).
38. Theil-Nielsen, J. & Søndergaard, M. Bacterial carbon biomass calculated from biovolumes. *Arch Hydrobiol* **141**, 195–207 (1998).
39. Ghai, R. *et al.* New Abundant Microbial Groups in Aquatic Hypersaline Environments. *Sci. Rep.* **1**, 135 (2011).
40. Raes, J., Korbel, J. O., Lercher, M. J., Von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**, R10 (2007).
41. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245 (2005).
42. Béja, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
43. Fuhrman, J. A., Schwalbach, M. S. & Stingl, U. Proteorhodopsins: an array of physiological roles? *Nature Reviews Microbiology* **6**, 488–494 (2008).
44. Riedel, T. *et al.* Genomics and Physiology of a Marine Flavobacterium Encoding a Proteorhodopsin and a Xanthorhodopsin-Like Protein. *PLoS One* **8**, e57487 (2013).
45. Sharma, A. K., Zhaxybayeva, O., Papke, R. T. & Doolittle, W. F. Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environ Microbiol* **10**, 1039–1056 (2008).
46. Wingard, L. L. *et al.* Cyanophycin production in a phycoerythrin-containing marine Synechococcus strain of unusual phylogenetic affinity. *Applied and Environmental Microbiology* **68**, 1772–1777 (2002).
47. Riemann, L. & Azam, F. Widespread N-acetyl-D-glucosamine uptake among pelagic marine bacteria and its ecological implications. *Applied and Environmental Microbiology* **68**, 5554–5562 (2002).
48. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiology and Molecular Biology Reviews* **73**, 249–299 (2009).

49. Coleman, M. L. & Chisholm, S. W. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences* **107**, 18634–18639 (2010).
50. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
51. Iverson, V. *et al.* Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**, 587–590 (2012).
52. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821–829 (2008).
53. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
54. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 0955–0964 (1997).
55. Nawrocki, E. P. *Structural RNA Homology Search and Alignment using Covariance Models* Ph.D. thesis, Washington University (2009).
56. Huang, Y., Gilna, P. & Li, W. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**, 1338–1340 (2009).
57. Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* **39**, D225–D229 (2011).
58. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244-W248 (2005).
59. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277 (2000).
60. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25**, 1–18 (2008).
61. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
62. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
63. Felsenstein, J. PHYLIP: phylogenetic inference package, version 3.5 c. (1993).
64. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463 (2005).
65. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
66. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
67. Eloe, E. A. *et al.* Going deeper: metagenome of a hadopelagic microbial community. *PLoS One* **6**, e20388 (2011).
68. Wilkins, D. *et al.* Key microbial drivers in Antarctic aquatic environments. *Fems Microbiol. Rev.* (2012).
69. Berger, S. A. & Stamatakis, A. Aligning short reads to reference alignments and trees. *Bioinformatics* **27**, 2068–2075 (2011).
70. Stamatakis, A., Komornik, Z. & Berger, S. A. in *Computer Systems and Applications (AICCSA), 2010 IEEE/ACS International Conference on.* 1–8 (IEEE).
71. Marie, D., Partensky, F., Jacquet, S. & Vaulot, D. Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. *Applied and Environmental Microbiology* **63**, 186–193 (1997).
72. Porter, K. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnology and Oceanography* **25** (1980).
73. MacIsaac, E. & Stockner, J. G. Enumeration of phototrophic picoplankton by autofluorescence microscopy. *Handbook of methods in aquatic microbial ecology. Lewis Publishers, Boca Raton, Fla* 187–197 (1993).
74. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365–386 (2000).
75. Sekar, R. *et al.* An improved protocol for quantification of freshwater Actinobacteria by fluorescence in situ hybridization. *Applied and Environmental Microbiology* **69**, 2928–2935 (2003).
76. Bouvier, T., Troussellier, M., Anzil, A., Courties, C. & Servais, P. Using light scatter signal to estimate bacterial biovolume by flow cytometry. *Cytometry* **44**, 188–194 (2001).

Acknowledgements

This work was supported by projects MAGYK (BIO2008-02444), MICROGEN (Programa CONSOLIDER-INGENIO 2010 CDS2009-00006), CGL2009-12651-C02-01 from the Spanish Ministerio de Ciencia e Innovación, DIMEGEN (PROMETEO/2010/089) and ACOMP/2009/155 from the Generalitat Valenciana and MaCuMBA Project 311975 of the European Commission FP7. FEDER funds supported this project. Work by AC and AP was also supported by project CGL2012-38909 (ECOLAKE) from the Spanish Ministerio de Economia y Competitividad. RG was supported by a Juan de la Cierva scholarship from the Spanish Ministerio de Ciencia e Innovación. The authors would like to thank Ana-Belen Martin-Cuadrado and Rebeca Ubeda-Lopez for assistance with sequencing.

Author contributions

R.G. and C.M.M. performed all the metagenomic analyses. A.P. and A.C. performed the FISH and flow-cytometric work. F.R.V. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Additional information

Accession Numbers: All the 43 assembled contigs described here have been deposited in GenBank and can be accessed using the accession numbers KC811108-KC811150.

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ghai, R., Mizuno, C.M., Picazo, A., Camacho, A. & Rodriguez-Valera, F. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. *Sci. Rep.* **3**, 2471; DOI:10.1038/srep02471 (2013).