The ISME Journal (2017) 11, 201–211

www.nature.com/ismej

ORIGINAL ARTICLE

Speciation and ecological success in dimly lit waters: horizontal gene transfer in a green sulfur bacteria bloom unveiled by metagenomic assembly

Tomàs Llorens–Marès[1], Zhenfeng Liu[2,6], Lisa Zeigler Allen[3], Douglas B Rusch[4,7], Matthew T Craig[3], Chris L Dupont[3], Donald A Bryant[2,5] and Emilio O Casamayor[1]

[1]*Integrative Freshwater Ecology Group, Centro de Estudios Avanzados de Blanes, CEAB-CSIC. Accés Cala Sant Francesc, Girona, Spain;* [2]*Department of Biochemistry and Molecular Biology, The Pennsylvania State University, State College, PA USA;* [3]*Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA, USA;* [4]*Informatics Group, J. Craig Venter Institute, Rockville, MD, USA and* [5]*Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT, USA*

A natural planktonic bloom of a brown-pigmented photosynthetic green sulfur bacteria (GSB) from the disphotic zone of karstic Lake Banyoles (NE Spain) was studied as a natural enrichment culture from which a nearly complete genome was obtained after metagenomic assembly. We showed *in situ* a case where horizontal gene transfer (HGT) explained the ecological success of a natural population unveiling ecosystem-specific adaptations. The uncultured brown-pigmented GSB was 99.7% identical in the 16S rRNA gene sequence to its green-pigmented cultured counterpart *Chlorobium luteolum* DSM 273^T. Several differences were detected for ferrous iron acquisition potential, ATP synthesis and gas vesicle formation, although the most striking trait was related to pigment biosynthesis strategy. *Chl. luteolum* DSM 273^T synthesizes bacteriochlorophyll (BChl) *c*, whereas *Chl. luteolum* CIII incorporated by HGT a 18-kbp cluster with the genes needed for BChl *e* and specific carotenoids biosynthesis that provided ecophysiological advantages to successfully colonize the dimly lit waters. We also genomically characterized what we believe to be the first described GSB phage, which based on the metagenomic coverage was likely in an active state of lytic infection. Overall, we observed spread HGT and we unveiled clear evidence for virus-mediated HGT in a natural population of photosynthetic GSB.
The ISME Journal (2017) **11,** 201–211; doi:10.1038/ismej.2016.93; published online 8 July 2016

Introduction

Green sulfur bacteria (GSB, *Chlorobiaceae*) form massive blooms, often of monoclonal nature, in the twilight zone of stratified aquatic environments (<1% of surface incident irradiance) with euxinic (anoxic and sulfidic) bottom waters (Gregersen *et al.*, 2009). GSB are anaerobic photoautotrophs that couple anoxic oxidation of sulfide and CO_2 fixation, and their specific contents in carotenoids and bacteriochlorophylls (BChl *c*, *d*, *e* and *a*) are ecological traits that dictate both their light-harvesting capacities and their potential ecological success (Montesinos *et al.*, 1983; Van Gemerden and Mas 1995; Bryant *et al.*, 2012). Usually brown-colored GSB (cells mostly containing BChl *e* and isorenieratene) bloom deeper than green-colored GSB (cells with BChl *c* and chlorobactene as dominant pigments) (Montesinos *et al.*, 1983). Within the GSB, there are examples of genomes with a high proportion of horizontal gene transfer (HGT, up to 24% of all genes in *Chlorobaculum tepidum* TLS (Nakamura *et al.*, 2004), formerly *Chlorobium tepidum* TLS). HGT is a major mechanism for bacterial innovation and adaptation in order to colonize new ecological niches and to improve *in situ* performance, thus acting as a trigger for prokaryotic speciation (Ochman *et al.*, 2000; Wiedenbeck and Cohan, 2011). HGT may be driven by transformation (naturally incorporated environmental DNA), conjugation (genetic material acquired through plasmid exchange between cells) and transduction (exchange through phage infection). The high proportion of HGT in *Chl. tepidum* is probably related to the fact that this bacterium is naturally transformable (Frigaard and Bryant, 2001). Likely examples of transduction may also be found in GSB; the *sox* cluster for thiosulfate utilization is a well-known example of

Correspondence: CL Dupont, Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA, USA.
E-mail: cdupont@jcvi.org
or E Casamayor, Integrative Freshwater Ecology Group, Centro de Estudios Avanzados de Blanes, CEAB-CSIC, Accés Cala Sant Francesc, 17300 Girona, Spain.
E-mail: casamayor@ceab.csic.es
[6]Current address: Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA.
[7]Current address: Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN, USA.
Received 6 January 2016; revised 6 May 2016; accepted 7 June 2016; published online 8 July 2016

lateral gene transfer in *Chlorobium phaeovibrioides* DSM 265 (Frigaard and Bryant, 2008). However, phages infecting GSB have not been described so far (Frigaard and Bryant, 2008).

The use of comparative genomic analysis on isolated strains has shown that HGT can have a key role in response to environmental selection pressures (Rocap *et al*., 2003; Martiny *et al*., 2009) and in the adaptation process *in vitro* (Wu *et al*., 2011). Genome comparisons of closely related strains provide clues to identify the role of HGT in ecotype formation and ecological diversification (Cohan and Koeppel, 2008; Rocap *et al*., 2003), and genetic explanations for the success of widespread marine plankton, for example, ecotypes of *Prochlorococcus* spp. and *Pelagibacter* spp. (Lindell *et al*., 2004; Zhao *et al*., 2013). Therefore, more comprehensive studies on the complex interplay between genetic recombination and ecology are needed (Polz *et al*., 2013). When pure cultures are difficult to obtain, metagenomic approaches can capture genomic differences in natural populations (Bhaya *et al*., 2007; Andersson and Banfield, 2008; Palenik *et al*., 2009; Klatt *et al*., 2011) and avoid the necessity to mimic the scale of natural ecosystems in laboratory experiments, which may discount the effect of potentially important variables for HGT (Aminov, 2011). The reconstruction of microbial genomes directly from environmental DNA through metagenomics is difficult (Luo *et al*., 2012). Initial studies focused on simple communities, such as a low-complexity acid mine drainage microbial biofilm with six estimated species (Tyson *et al*., 2004). Other studies used alternatives to help simplify the community, such as the use of artificially enriched communities (Martin *et al*., 2006), sequencing multiple metagenomes of the same community (Albertsen *et al*., 2013) or a dual approach of single-cell sequencing with co-assembly and binning of multiple metagenomes (Dupont *et al*., 2012).

In the present investigation, we show a case where HGT explains the ecological success of a population *in situ*. We reconstructed the consensus genome of a natural blooming GSB population without previous culturing using metagenomics. This dominant population serves as a natural enrichment culture from which a nearly complete genome was assembled and used to study ecosystem-specific adaptations. The presence of putative phage assemblies with homology to the consensus genome unveiled consistent evidence for virus-mediated HGT in a natural population of GSB.

Materials and methods

Environment and sample analysis

A brown-colored water sample was collected from deep (24 m) euxinic waters of the meromictic basin III (CIII) in the karstic Lake of Banyoles (NE Spain, 42 18'N, 21 45'E) on 9 May 2010. Data for environmental parameters and the water column vertical profile have been recently published (see Table 1 and Figure 1 in Llorens-Mares *et al*., 2015), and additional information for the bacterial 16S rRNA gene composition in Llorens-Marès *et al.* (2016). Brown-pigmented GSB massively and persistently bloom in CIII (Montesinos *et al*., 1983), and Chlorobi microscopic counts in the Banyoles area can usually reach $>10^6$ cells ml^{-1} (Casamayor *et al*., 2000). Sampling, environmental analysis, water filtering and DNA extraction were carried out as recently reported (Llorens-Mares *et al*., 2015). The size fraction 0.8–3 μm was targeted for assembly. In this sample, we had measured high concentrations of BChl *e*, the characteristic pigment of brown-colored species of *Chlorobium* with a high relative abundance (>50%) of 16S rRNA gene closely matching (99.7% identity) the green-colored species *Chlorobium luteolum* DSM 273^T (Llorens-Mares *et al*., 2015) indicating a bloom of

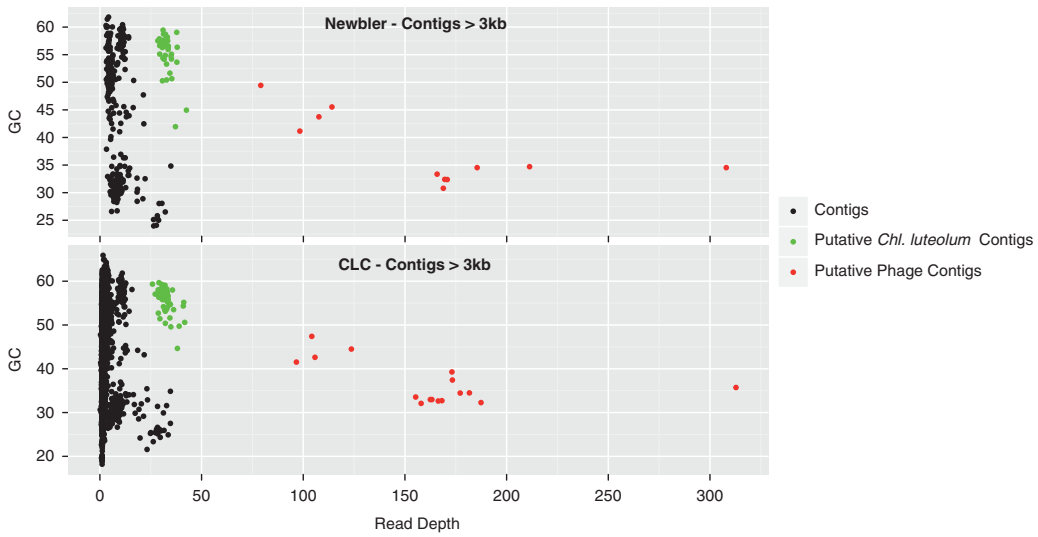

Figure 1 GC content versus read depth for each contig over 3 kb after Newbler and CLC assemblies. Green dots show the sequences related to the blooming chlorobi and selected for subsequent analysis. Red dots show the sequences related to the putative infecting phage.

brown-pigmented GSB. *Chl. luteolum* DSM 273 was formerly named *Pelodictyon luteolum* DSM 273, the cells have gas vesicles in contrast to the majority of GSB, and it was isolated from a coastal meromictic lake in Norway, requiring >1% salt concentration for growth (http://genome.jgi.doe.gov/pellu/pellu.home.html). Accordingly, the natural population from which we obtained the DNA for metagenomic sequencing and genome reconstruction was named *Chl. luteolum* CIII.

Metagenomics analyses

A total of 492 615 reads with 401nt average read length were generated by shotgun metagenomics (Life Technologies 454 titanium, Branford, CT, USA) as recently reported (Llorens-Mares *et al.*, 2015). Reads were assembled using two different software assemblers: Newbler Assembler (454 Life Sciences, Branford, CT, USA) and CLC Assembly Cell (CLC bio, Qiagen, Waltham, MA, USA), which overall produced 2971 contigs (6 859 926 bp/L50 = 9778 bp/N50 = 107) and 23 490 contigs (21 080 885 bp/L50 = 897 bp/N50 = 4674), respectively. For each assembly, we selected contigs >3 kb that were plotted against GC content and read depth (Figure 1). To assemble *Chl. luteolum* CIII specifically, we selected contigs with ~57% GC (equivalent to the reported 57.33% GC content of *Chl. luteolum* DSM 273^T) and a read depth of ~29. Using these criteria, we selected 45 contigs from the Newbler assembly (average length, 47 kb, average read depth 29.9 ± 2.8 and mol% GC = 56.7%) and 75 contigs from the CLC assembly (average length 28 kb, average read depth 29.1 ± 3.2 and mol% GC = 56.7%). We then used phred/phrap/consed package (Gordon *et al.*, 1998) to combine these assemblies to produce an assembly of 41 contigs that were reduced to 40 after closing the gap between contig 31 and 41. The final assembly was 2 154 228 bp with an average mol% GC = 56.73%. The sequence of the assembly has been deposited at DDBJ/EMBL/GenBank under the accession number LVWG01000000, assembly accession GCA_001622165.1

Contigs were ordered and oriented according to the reference genome, *Chl. luteolum* DSM 273^T, and visualized for a synteny comparison using Genome Matcher (Ohtsubo *et al.*, 2008). The genome encoded 2023 open reading frames (ORFs) that were annotated by the NCBI Prokaryotic Genome Annotation Pipeline (released 2013; https://www.ncbi.nlm.nih.gov/genome/annotation_prok/) with rigorous manual curation. We checked for genomic completeness by searching for a set of 110 universally occurring marker genes, very rarely duplicated, essential for cellular life, and believed to be very ancient (Dupont *et al.*, 2012). All 110 of these genes were present in CIII genome, and all were present as single-copy genes.

We used DNAPlotter (Carver *et al.*, 2009) for the visualization of different traits such as the global genome, the mol% GC, the GC skew and all ORFs. For a visual comparison with the reference genome, we used the Artemis Comparison Tool (Carver *et al.*, 2012). Perl scripts were run to obtain a list of the ORFs that were classified as orthologs with the reference genome using a whole-genome reciprocal BlastP analysis in order to establish differences in protein coding between strains.

For a global comparison of similarity between genomes and to assess the average nucleotide identity (ANI) of CIII genome with sequenced strains of GSB, we used JSpecies V1.2.1 with the average clustering method (Richter and Rossello-Mora, 2009). An ANI value above ~95–96% is used as a standard for the prokaryotic species definition (Richter and Rossello-Mora, 2009). Hierarchical clustering analysis of the resulting all versus all ANI similarity matrix obtained with JSpecies was performed in R (R Core Team, 2014).

As a result of the assemblies with Newbler and CLC, we detected the presence of large contigs (>3 kb) with an unusually high-read depth (~79–325 ×; Figure 1). Based on their isolation from the cellular size fraction, these sequences were eventually assigned to a potential infecting phage population. We re-assembled these contigs with the same procedure followed for *Chl. luteolum* CIII. A long contig (65 kb) with read depth 34 and mol% GC = 34.8 was also selected because of the presence of genes coding for phage-related proteins. We ended with five contigs designated as putatively phage-derived sequences. These contigs were annotated with the JCVI viral annotation pipeline (Lorenzi *et al.*, 2011). One of the putative phage-derived contigs had similarity to a region of the *Chl. luteolum* CIII genome. This region was visualized for synteny using the R package genoPlotR (Guy *et al.*, 2010).

16S rRNA gene phylogenetic analysis

A comprehensive phylogenetic tree of the 16S rRNA gene was generated with reference sequences from the phylum *Chlorobi*, from the assembled genome and from previous studies in the area (Figueras *et al.*, 1997; Casamayor *et al.*, 2000). Sequences were aligned with SINA aligner (Pruesse *et al.*, 2012), and phylogenetically compared by maximum likelihood with the general time-reversible model from RAxML v7.3.0 (Stamatakis, 2006) and *Bacteroidetes fragilis* as outgroup.

BChl e phylogenetic analysis

Genes encoding proteins associated with BChl *e* biosynthesis (for example, BchF3 and BciD (Harada *et al.*, 2013)) and isorenieratene biosynthesis (CruB (Maresca *et al.*, 2008b)) were found on two different CIII contigs. We designed primers for each contig end in order to confirm by PCR amplification and DNA sequencing that the genes were contiguous and formed a cluster in one genomic locale in the natural population. The concatenated protein sequences of these genes were used to construct a maximum likelihood tree to assess the phylogenetic relationships of the Bchl

e cluster inserted in strain *Chl. luteolum* CIII with the other sequenced brown-colored GSB species (*Chl. phaeobacteroides* DSM 266, *Chl. clathratiforme* DSM 5477, *Prosthecochloris phaeum* CIB 2401, *Cba. limnaeum* DSM 1677 and *Ptc. phaeobacteroides* BS1). PartitionFinderProtein v1.0.1 (Lanfear *et al*., 2012) was used to determine the best substitution model for each partition, and RAxML v7.3.0 (Stamatakis, 2006) was used to generate the maximum likelihood tree. We used as outgroup a combination of distantly related sequences for each of four concatenated proteins: SDR (short-chain dehydrogenase/reductase enzyme; AGA91907 from *Thioflavicoccus mobilis* 8321), CruB (ACF12554 from *Chloroherpeton thalassium* ATCC 35110), RSAM (radical *S*-adenosylmethionine protein; ACF01393 from *Rhodopseudomonas palustris* TIE-1) and BchF3 (ABB27675 from *Chlorobium chlorochromatii* CaD3). A visual syntenic analysis of the region containing the BChl *e* cluster in all sequenced genomes was carried out using the R package genoPlotR (Guy *et al*., 2010).

FeoB, metallophospherase and vrl *locus analyses*
Both FeoB and metallophosphoesterase protein trees were generated as follows. Reference sequences were collected from the non-redundant NCBI database using BlastP and aligned using MUSCLE (Edgar, 2004). Aligned sequences were cleaned with Gblocks (Castresana, 2000), and a maximum likelihood tree for each protein alignment was generated using RAxML v7.3.0 (Stamatakis, 2006). SyntTax (Oberto, 2013) was used to explore the genomic context of FeoB in other *Chlorobium* spp. genomes and *vrl* locus in additional genomes.

CRISPR identification
Clustered regularly interspaced short palindromic repeats (CRISPRs) were identified using CRISPRfinder (Grissa *et al*., 2007). One CRISPR was found with a direct repeat of 32 bp containing 12 spacers. The spacer regions were used as a BLAST query against a database created from all metagenomic reads. The genome annotation was generated using CLC workbench.

Viral DNA polymerase B phylogeny
Reference viral sequences were identified with a Hidden Markov Model search (PF00136). The sequences were then aligned with MUSCLE and a phylogenetic tree representation constructed using PhyML (Guindon and Gascuel, 2003) and Archaeopteryx (Han and Zmasek, 2009).

Results

Genome identification and 16S rRNA gene phylogenetic analyses
A hierarchical clustering analysis of *Chlorobi* genomes after all versus all ANI values comparison showed that the closest genome to Banyoles CIII population was *Chl. luteolum* DSM 273^{T} (ANI value 91.7%) which together with *Chl. phaeovibrioides* DSM 265 formed a separate phylogenetic clade (Figure 2). This was confirmed after a global syntenic visualization with the closest GSB sequenced genomes, and by the phylogenetic tree of the 16S rRNA gene sequences (99.71% identity, Supplementary Figure S1). Interestingly, we noticed that the current 16S rRNA gene *Chlorobium* sequence was identical to VIBAC-6 sequence, which was collected from Lake Vilar in 1996 (Casamayor *et al*., 2000), a neighboring lake connected to Lake Banyoles, but different from the *Chlorobium* CIBAC-3 present in Lake Ciso, just <1 km away and connected by the same groundwater-fed karstic system (Casamayor *et al*., 2000).

In order to establish the relationship between 16S rRNA gene identity values and whole-genome similarities within the *Chlorobi* group, we plotted the ribosomal identities versus the ANI values for all genomes (Figure 3). Most genomes were within 66–74% ANI and 91.9–97.7% 16S rRNA gene identity, but two *Chlorobi* species were distantly related to any other: *Ignavibacterium album* and *Chl. thalassium* ATCC 35110. The three species of *Chlorobaculum* clustered together, and the CIII population was distantly related to the brown green sulfur bacterium *Chl. phaeobacteroides*.

Comparative genomic analyses with Chl. luteolum *CIII population*
We explored in detail which clusters of genes were not present in the strain CIII as compared with DSM 273^{T}. For example, DSM 273^{T} has a gas vesicle gene cluster encoding eighteen proteins (YP374609 to YP374627), most of which have best hits to *Chl. clathratiforme* DSM 5477, that was not detected in the CIII population. More relevant was the

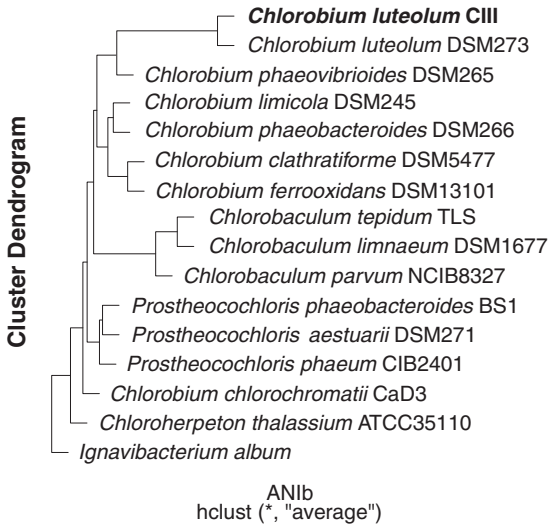

Figure 2 JSpecies hierarchical clustering analysis carried out on the ANI similarity matrix obtained from the available *Chlorobi* genomes.

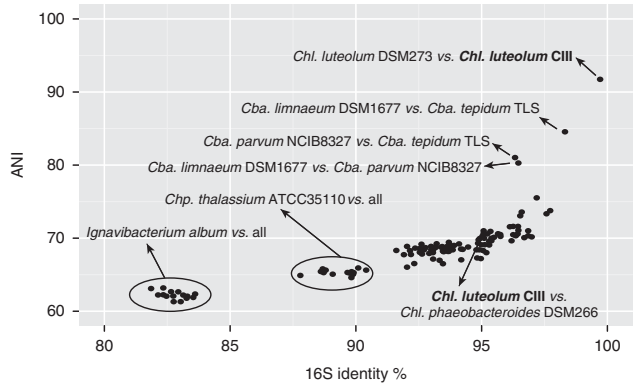

Figure 3 Pairwise comparison values of ANI versus 16S rRNA gene identity for the different available *Chlorobi* genomes.

absence of an ATP synthase operon (YP374968 to YP374975), which included the eight genes required for the synthesis of the ATP synthase complex (*atpA* (F_1), *atpD* (F_1), *atpG* (F_1), *atpH* (F_1), *atpC* (F_1), *atpE* (F_0), *atpB* (F_0) and *atpF* (F_0)). We checked for the presence of additional ATP synthase genes, as they are essential for cell viability, and found them interspersed across the genome (both in DSM 273^T and CIII). Apparently, the operon found in *Chl. luteolum* DSM 273^T and missing in CIII strain had homology with the Na^+-dependent F_1F_0-ATP synthase found in the halotolerant cyanobacterium *Aphanothece halophytica* (Soontharapirakkul *et al.*, 2011), indicating that it could be related to salt tolerance, which is not required in fresh waters.

Conversely, 286 genes in the *Chl. luteolum* CIII genome were more closely related to genes in organisms other than *Chl. luteolum* DSM 273^T and thus might have been acquired by HGT (Figure 4). Among them, eight ORFs (from A3K90_03400 to A3K90_03435) were detected in contig 51 with the same structure and best protein similarity scores (95–99%) to the brown-pigmented *Chl. phaeovibrioides* DSM 265. The products of these genes were identified as two different copies of FeoA, FeoB, flavodoxin, a ferritin-DPS family member and three hypothetical proteins. A phylogenetic tree of the FeoB proteins showed that the two different variants of FeoB are encoded in *Chlorobium* spp. genomes (Supplementary Figure S2). The first form of the FeoB is predicted to be a protein of 712 amino acids, and it is present in most *Chlorobium* spp. genomes, including *Chl. luteolum* DSM 273^T. The second form yields a protein of 790 amino acids with homologs only found in some GSB, including *Cba. tepidum* TLS, *Cba. parvum* NCIB 8327, *Chl. limicola* DSM 245, *Ptc. phaeobacteroides* BS1, *Chl. phaeobacteroides* DSM 266 and *Chl. phaeovibrioides* DSM 265.

The region including ORFs A3K90_09535 to A3K90_09620 showed greatest similarity (80–99%) with *Chl. phaeobacteroides* DSM 266 and *Chl. clathratiforme* DSM 5477 proteins. A closer inspection of this region identified BChl *e* and isorenieratene biosynthesis genes, which are linked pathways. The region with the BChl *e* genes was initially split into two contigs connected by a 1312 nucleotide linking sequence, which contained a transposase of the IS4 family with best hit with *Chl. phaeobacteroides* DSM 266 (YP912276).

The phylogenetic tree of the putative proteins involved in BChl *e* biosynthesis showed a high similarity between *Chl. luteolum* CIII and *Chl. clathratiforme* DSM 5477 and *Chl. phaeobacteroides* DSM 266, respectively (Figure 5). A synteny analysis of the region showed that gene positions were always conserved with the exception of an inversion in the SDR gene, a putative dehydrogenase/oxidoreductase. It also showed that *Chl. clathratiforme* DSM 5477, *Chl. phaeobacteroides* DSM 266 and *Chl. luteolum* CIII had additional genes in this cluster that apparently were not related to BChl *e* biosynthesis, possibly explaining a different phylogenetic history and also showing the genomic flexibility of this region. A detailed analysis of the transition in mol% GC along this region with *Chl. phaeobacteroides* DSM 266 (Supplementary Figure S3) showed the same GC pattern in the incorporated fragment suggesting a highly probable HGT episode.

Phage-related HGT

Interestingly, we detected the insertion in CIII population of a cluster of six genes (A3K90_02865 to A3K90_02955 on contigs 44 and 20) with homology with the virulence-related locus (*vrl* locus) of *Dichelobacter nodosus* (Haring *et al.*, 1995) and not present in the genome of *Chl. luteolum* DSM 273^T. Specifically, the *vrlJKLOPQ* usually found in distantly related microbes such as *Acidothermus cellulolyticus*, *Thermoanaerobacter ethanolicus*, *Nitrosococcus mobilis, N. oceani* (Knaust *et al.*, 2007) and also *Desulfovibrio aespoeensis* Aspo2 and *Methanosalsum zhilinae* DSM 4017. The presence of phage-related proteins at the end of the region suggested a possible phage-related HGT episode.

A couple of 'promiscuous regions' characterized by multiple recombination events were detected in the *Chl. luteolum* CIII genome with many genes related to mobile elements (integrases, recombinases, transposases, among others), and some of them annotated as phage-related proteins. Moreover, additional indicators of past phage infection were identified such as CRISPRs, genomic regions containing multiple and short repeats with interspersed spacer DNA, acquired from previous viral encounters (Westra *et al.*, 2014). One CRISPR locus was identified with direct repeat sequence of 32 bp and 12 spacer regions (Supplementary Table S1). Although no spacer sequences were identified within the larger data, it suggests this CIII population may use a CRISPR/Cas-like mechanism to evade phage infection.

Promiscuous regions, CRISPRs and *vrl* locus, altogether provided evidence for an ongoing association and genomic exchange between *Chl. luteolum* CIII and phages. In addition, up to 37% of the identified proteins with non-reciprocal BLAST hits to *Chl. luteolum* DSM 273^T (and, therefore, the

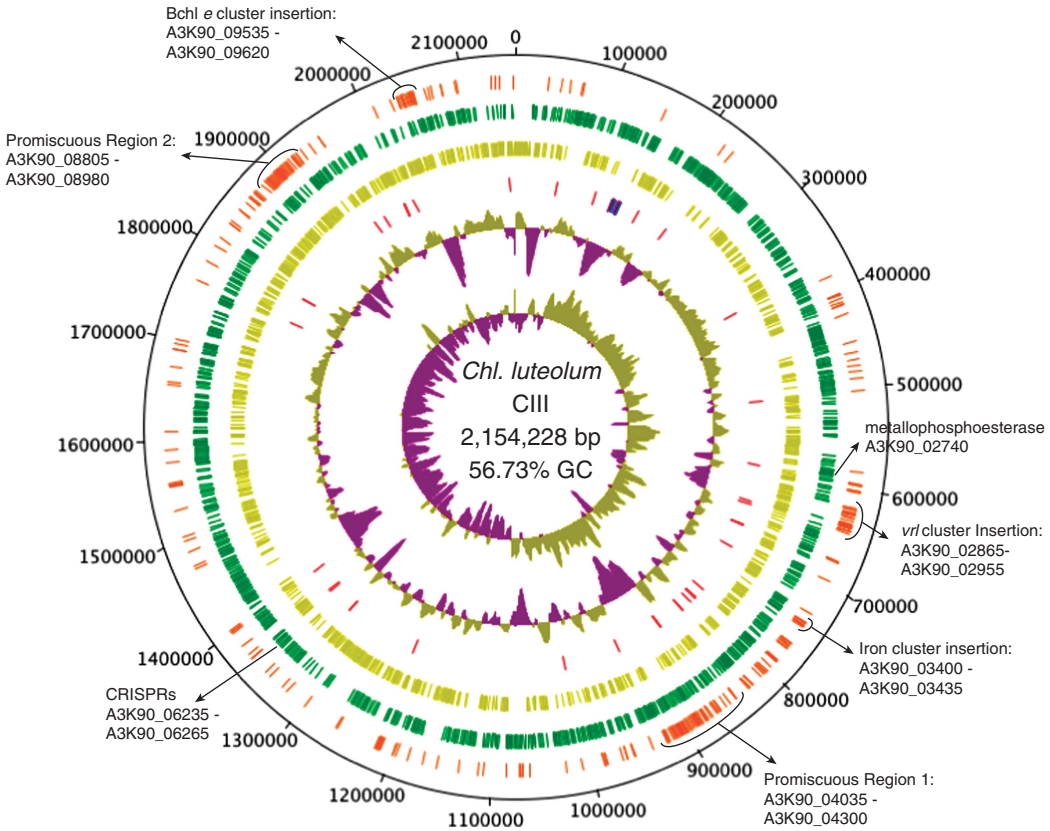

Figure 4 Circular map of *Chl. luteolum* CIII genome using DNAplotter. The first two central circles show GC skew and G+C content, respectively. Baseline on the G+C plot represents 56.7% average value. The remaining four circles show tRNA and rRNA (red and blue labels, respectively), all reverse strand ORFs (light green label), all forward strand ORFs (green label) and all the ORFs more closely related to genes in organisms other than *Chl. luteolum* DSM 273^{T} (orange label). Special features are highlighted, see main text.

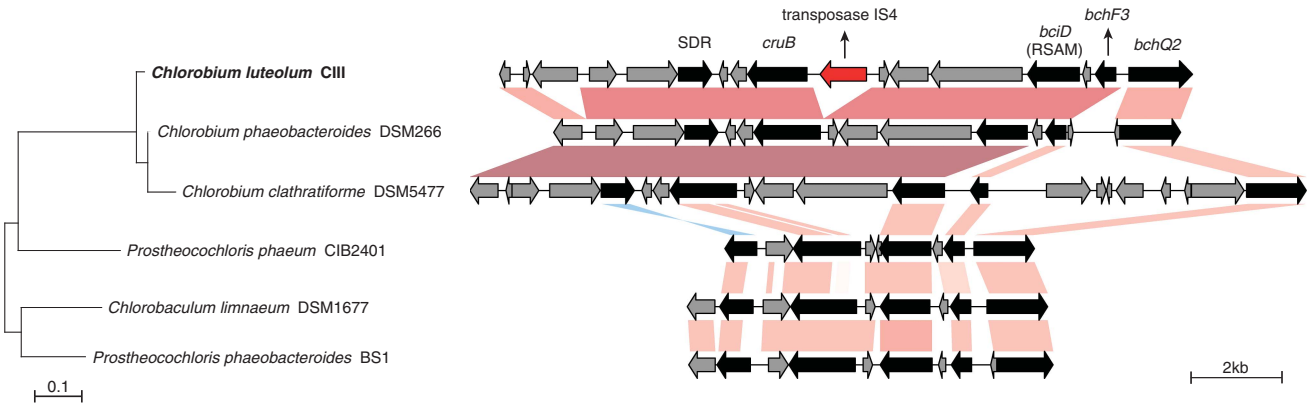

Figure 5 Maximum likelihood phylogeny with RAxML of five concatenated proteins (SDR, CruB, BciD (RSAM), BchF3 and BchQ2) found in all BChl *e* synthesizing *Chlorobi*. A combination of distantly related sequences for each of the four concatenated genes were used as outgroup. The genome synteny of the region is shown next to the phylogenetic tree. The bit score of the BLAST alignment is represented with shades of red (same orientation) and blue (inverse orientation).

potentially acquired proteins) were related to mobile elements or phages (Supplementary Figure S4). The 39% of the proteins gained by *Chl. luteolum* CIII (Supplementary Figure S5) had best hits with other GSB, but a striking 33% were most closely related to a large variety of species, mostly within *Proteobacteria*. Overall, these results illustrate the plasticity of the CIII population genome and the substantial gene flow that may appear *in situ*, mostly between closely related species but also between phylogenetically distant organisms.

Putative GSB phage

To date, phages infecting *Chlorobi* have not been reported (Frigaard and Bryant, 2008), but genomic

evidence supports the existence of GSB viruses. In addition, in the assembly analysis we detected a pool (CLC 16 contigs, length 165 187 nt; Newbler 12 contigs, length 168 317 nt) of large contigs (>3 kbp) with a very high-read depth (Figure 1). We re-assembled these contigs following the same procedure used for *Chl. luteolum* CIII, and obtained five contigs (Supplementary Table S2) that were identified as belonging to putative bacteriophages (see below). Because we collected a plankton size fraction >0.8 µm, these putative phages were likely infecting the *Chlorobi* CIII population at the time of sampling, either through a lysogenic or a lytic infection process. Putative phage contigs 2 and 4 could not be classified to any known bacteriophage and only encoded three phage-related genes. Based on the homology of multiple ORFs to an N4-like phage, contig 1 was probably derived from a Podovirus. Finally, according to the GC content, read depth (Supplementary Table S2) and predicted host-acquired auxiliary metabolic genes, contigs 3 and 5 (Supplementary Table S3) were potentially derived from a putative lytic *Myoviridae* phage. Contig 5 and its predicted proteins indicated a mixed homology to known bacteriophage (Supplementary Figure S6). The read depth of contigs 3 and 5 was ~6-fold higher (that is, 174×) than the average for *Chl. luteolum* CIII, which strongly suggests that *Myoviridae* was responsible for an active, lytic infection process *in situ*. A phylogenetic reconstruction of the phage-conserved DNA polymerase provided further evidence that the predicted sequences form a distinctive lineage, which are related to DNA polymerases found in other bacteriophage (Supplementary Figure S7).

Finally, a BlastN analysis of *Chl. luteolum* CIII versus the putative phage returned a high-identity region of 51 nucleotides next to the *vrl* locus in *Chl. luteolum* CIII that matched two separate parts of the putative phage (Figure 6). One was 20-nt long with 100% identity, and the other was 33-nt long but contained two mismatches. The two regions were separated by 2169 nt, in which the coding sequence for a hypothetical protein from *Sinorhizobium* phage PBC5 (Contig5_11) was found (Figure 6 and Supplementary Table S3). Interestingly, next to this similarity region, and as part of the *vrl* locus insertion, we found a homolog of bacteriophage P4 integrase (A3K90_02945) and a phage transcriptional regulator *alpA* (A3K90_02940). About 50 kb away from this region, a putative auxiliary metabolic genes within a region of conserved phage genes was predicted to be a metallophosphoesterase, a Ser/Thr protein phosphatase (Contig5_33), with strong sequence homology with A3K90_02740 from *Chl. luteolum* CIII genome (Figure 6). The phylogenetic tree of metallophosphoesterase sequences showed that the protein encoded on phage Contig5_33 was clearly derived from GSB, and also had close sequence homology with other phage metallophosphoesterases (Supplementary Figure S8).

Discussion

HGT has been widely assumed to be a major mechanism for bacterial innovation and ecological diversification (Wiedenbeck and Cohan, 2011), and several HGT studies have been focused on the exchange of virulence-associated genes in both human and animal pathogens (for example, Saunders *et al*., 2005), and of antibiotic resistance genes (Summers, 2006). In the environment, genomic investigations pointing to a role for HGT in the origin of ecological diversification are however more scarce and difficult to achieve. The available studies carried out mostly in soils (for example, *Pseudomonas putida*, Wu *et al*., 2011), microbial mats (*Synechococcus* spp., Melendrez *et al*., 2011) and the ocean (*Prochlorococcus marinus*, Martiny *et al*., 2009) have used isolated strains. In our investigation, we unveiled the ecological success of a natural GSB population *in situ* through HGT without previous culturing using massive sequencing and genome reconstruction *in silico*. The monoclonal nature of GSB blooms has been previously reported (Gregersen *et al*., 2009) and we assembled promiscuous regions with a coverage even higher to that of the rest of the genome. This fact suggests that there is a nearly clonal population in Lake Banyoles, in sharp contrast to the oceanic photoautotroph where we could not close these promiscuous regions after metagenomic assembly (Rusch *et al*., 2010). The reason could be that marine *Prochlorococcus* contains a high diversity of strains for the same species with the same genomic backbone but variations in the promiscuous regions, also called

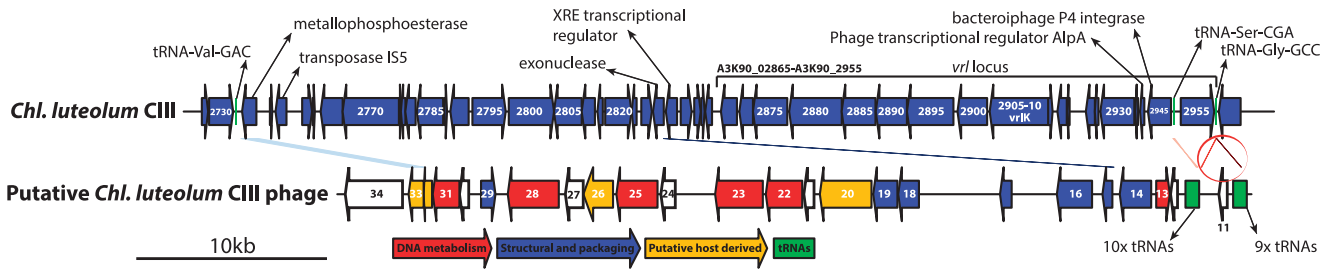

Figure 6 Syntenic analysis showing the relationship of the putative GSB phage with the *Chl. luteolum* CIII genome sequence. A blue shade indicates the homology region of the metallophosphoesterase gene. A red circle indicates the region with the 51 nucleotides identity region. The types of genes identified in the putative *Chl. luteolum* CIII phage are shown in colors according to DNA metabolism (red), structural and packaging (blue), putative host-derived (yellow) and tRNAs (green). Special features on both sequences are highlighted with arrows.

hypervariable islands, that leads to a high 'collective diversity' even when one species dominates (Biller *et al.*, 2015). The ecological implications or evolutionary reasons are unknown, but the continuous nature of the marine environment may promote collective diversity, whereas the endemic nature of isolated lakes (Barberan and Casamayor, 2011) may promote clonal similarity.

The ecophysiology of phototrophic sulfur bacteria has been widely studied not only from an ecological perspective (Montesinos *et al.*, 1983; Van Gemerden and Mas, 1995) but also from molecular and genomic points of view (Frigaard and Bryant, 2008; Habicht *et al.*, 2011; Bryant *et al.*, 2012). These previous findings supplied consistent background to unveil the ecological consequences of HGT, in addition to the direct genome comparison with a closely related cultured counterpart. Interestingly, *Chl. luteolum* CIII and DSM 273^T exhibited 99.7% identity in 16S rRNA gene sequence and only 91.7% in the ANI value. This could simply reflect lack of enough evolution time for diversification, indicating that the differences found in the genome are relatively very recent and that the high degree of genetic exchange between these two *Chlorobi* populations is not yet reflected in the most conserved genes, but significantly captured through a whole-genome evaluation. This supports the idea that lateral exchange of genetic material within GSB occurs at a very high rate, as previously suggested (Nakamura *et al.*, 2004). Therefore, most probably GSB could be very good candidates as model organisms for HGT-guided evolution studies.

GSB are strict photoautotrophs that strongly depend on light availability and light harvesting for growth. In order to exploit the light wavelengths that reach the anoxic deep water layers where GSB reside, these organisms synthesize specific carotenoids and specialized light-harvesting antenna organelles, the chlorosomes, which are the most efficient light-harvesting structures known (Frigaard and Bryant, 2006). The type of bacteriochlorophylls preferentially found in the chlorosome is one of the key factors that explain the ecological success of GSB (Montesinos *et al.*, 1983; Van Gemerden and Mas, 1995; Bryant *et al.*, 2012). Interestingly, brown-pigmented blooms of GSB were already reported from Lake Banyoles basin CIII in 1978 (Montesinos *et al.*, 1983) and the population sampled in 2010 had the same 16S rRNA gene signature that the dominant population from 1996 (VIBAC-6, Casamayor *et al.*, 2000). The BChl *e* cluster acquired by *Chl. luteolum* CIII confers some advantages that are crucial from an ecological point of view. First, the absorption peak of the BChl shifts from 746 nm in BChl *c* to 714 nm in BChl *e* (Harada *et al.*, 2013) allowing it to cover a different range of wavelengths. However, more importantly, there is a large increase in absorption in the blue near 520 nm, which overlaps strongly with those light wavelengths that penetrate most deeply in the water column. Furthermore, the BChl *e* cluster includes the *cruB* gene, which is responsible for the biosynthesis of β-carotene and thus enables the production of isorenieratene and β-isorenieratene, which are almost universally associated with organisms that synthesize BChl *e* (Maresca *et al.*, 2008a). These carotenoids are important elements to broaden and increase the absorption of brown-colored species between 480 and 550 nm and expand the photoadaptation range (Hirabayashi *et al.*, 2004). These differences are of great significance in terms of competition in an ecological environment where light is one of the limiting factors (Van Gemerden and Mas, 1995). HGT of iron-processing genes may also provide ecological success. GSB have numerous proteins with Fe/S clusters in the reaction centers and are therefore highly dependent upon iron for growth. Soluble Fe^{2+} is abundant under low-oxygen conditions as in basin CIII but the reaction with hydrogen sulfide reduces its biological availability. The incorporation of an iron transport genes cluster might confer both higher affinity and higher iron storage capacity to *Chl. luteolum* CIII. Thus, FeoAB proteins can be advantageously used for Fe^{2+} uptake (Kammler *et al.*, 1993), flavodoxin as a low-potential electron donor that replaces ferredoxin under iron limitation (Chauhan *et al.*, 2011) and ferritin-DPS as an iron storage protein (Andrews *et al.*, 2003).

Phages may influence the history and evolution of their hosts through HGT (Rohwer and Thurber, 2009) and control the abundance of bacterial blooms (for example, Deng and Hayes, 2008). Intriguingly, any phage capable of infecting a GSB has been reported so far (Frigaard and Bryant, 2008) although genomic evidence exists that members of the Chlorobiaceae have been exposed and responded to phage infection. In this investigation, we observed the presence of phage-related contigs with very high-read depth, which directly points to the possibility of an ongoing lytic infection of the dominant *Chl. luteolum* CIII population. The presence of CRISPRs and the high proportion of mobile elements and phage-related proteins in the *Chl. luteolum* CIII genome provides additional evidence for previous phage infection, and a close relationship between GSB and phages. Although definitive studies should still be carried out, additional evidences supported this hypothesis. First, the presence in the phage sequence assembly of a metallophosphoesterase that has close phylogenetic identity to a gene associated with the putative GSB host. Metallophosphoesterases represent a functionally diverse superfamily of enzymes with Ser/Thr protein phosphatases as important components of various regulatory mechanisms (Lohse *et al.*, 1995; Villafranca *et al.*, 1996), and previously identified in phages such as PBECO4 (Kim *et al.*, 2013) or bacteriophage λ, suggesting they may mediate the dephosphorylation of certain proteins to allow more effective production of phage or regulate viral transcription (Cohen and Cohen, 1989). Second, the similarity found in the region next to the insertion of the *vrl* locus, which has previously been related to virulence factors (Billington *et al.*, 1999). Interestingly, proteins encoded by the *vrl* locus have best BlastP hits

with distantly related organisms, suggesting that the transfer of these genes might occur through virus-mediated mechanisms (Haring *et al*., 1995; Billington *et al*., 1999; Fuhrman, 1999; Knaust *et al*., 2007). Finally, the phylogenetic reconstruction of the phage-conserved DNA polymerase showed an association with a distinct lineage within *Myoviridae*. Considering that no phage has yet been described for any GSB, the lack of close relatives is not surprising. Additional mechanisms that can drive HGT such as transformation and conjugation cannot be however completely ruled out.

Overall, we were able to provide consistent genomic explanation for the success of the GSB blooming in Lake Banyoles. The ecological implications of acquiring genes for BChl *e* synthesis and Fe transport are substantial, and they could confer upon *Chl. luteolum* CIII, a clear advantage over green-colored GSB. In addition, our results unequivocally show that pigments such as BChl *e* are not reliable phylogenetic markers for GSB, in agreement with the fact that they are not monophyletic traits (Overmann and Tuschak, 1997). Interestingly, the presence of a transposase IS4 in the cluster of genes thought to mediate the synthesis of BChl *e* may partly explain the mobility of this region and may additionally be related to recent biological transformations and horizontal transfer mechanisms, which is the most reasonable explanation for the pigment-phylogeny incongruences. The influence of phages in the environment has been shown to be larger than previously thought (Fuhrman, 1999; Sharon *et al*., 2009), and we show here initial evidence for a putative phage that may infect GSB and could control the blooming population and act as a HGT vector for these ancient photoautotrophic microorganisms.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This research was funded by grant DARKNESS CGL2012-32747 from the Spanish Office of Science (MINECO) to EOC and by the Global Ocean Sampling Project supported by the Beyster Family Foundation Fund of the San Diego Foundation and the Life Technology Foundation (to JCVI). Work on BChl *e* biosynthesis and the genomics of GSB in the laboratory of DAB was supported by the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the U.S. Department of Energy through Grant DE-FG02-94ER20137.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen K L, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.

Aminov RI. (2011). Horizontal gene exchange in environmental microbiota. *Front Microbiol* **2**: 158.

Andersson A, Banfield JF. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **230**: 1047–1050.

Andrews SC, Robinson AK, Rodriguez-Quinones F. (2003). Bacterial iron homeostasis. *FEMS Microbiol Rev* **27**: 215–237.

Barberan A, Casamayor EO. (2011). Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microb Ecol* **61**: 465–472.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N *et al.* (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* **1**: 703–713.

Biller SJ, Berube PM, Lindell D, Chisholm SW. (2015). *Prochlorococcus*: The structure and function of collective diversity. *Nat Rev Microbiol* **13**: 13–27.

Billington SJ, Huggins AS, Johanesen PA, Crellin PK, Cheung JK, Katz ME *et al.* (1999). Complete nucleotide sequence of the 27-kilobase virulence related locus (vrl) of *Dichelobacter nodosus*: evidence for extrachromosomal origin. *Infect Immun* **67**: 1277–1286.

Bryant D, Liu Z, Li T, Zhao F, Costas AG, Klatt C *et al.* (2012). Comparative and functional genomics of anoxygenic green bacteria from the taxa *Chlorobi*, *Chloroflexi*, and *Acidobacteria*. In: Burnap R, Vermaas W (eds). *Functional Genomics and Evolution of Photosynthetic Systems*. Springer: Netherlands, pp 47–102.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**: 464–469.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25**: 119–120.

Casamayor EO, Schafer H, Baneras L, Pedros-Alio C, Muyzer G. (2000). Identification of and spatio-temporal differences between microbial assemblages from two neighboring sulfurous lakes: comparison by microscopy and denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **66**: 499–508.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Chauhan D, Folea IM, Jolley CC, Kouril R, Lubner CE, Lin S *et al.* (2011). A novel photosynthetic strategy for adaptation to low-iron aquatic environments. *Biochemistry* **50**: 686–692.

Cohan FM, Koeppel AF. (2008). The origins of ecological diversity in prokaryotes. *Curr Biol* **18**: R1024–U1017.

Cohen PT, Cohen P. (1989). Discovery of a protein phosphatase activity encoded in the genome of bacteriophage lambda. Probable identity with open reading frame 221. *Biochem J* **260**: 931–934.

Deng L, Hayes PK. (2008). Evidence for cyanophages active against bloom-forming freshwater cyanobacteria. *Freshwater Biol* **53**: 1240–1252.

Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Richter RA, Valas R *et al.* (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* **6**: 1186–1199.

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

Figueras JB, Garcia-Gil LJ, Abella CA. (1997). Phylogeny of the genus *Chlorobium* based on 16S rDNA sequence. *FEMS Microbiol Lett* **152**: 31–36.

Frigaard N-U, Bryant D. (2008). Genomic insights into the sulfur metabolism of phototrophic green sulfur bacteria. In: Hell R, Dahl C, Knaff D, Leustek T (eds). *Sulfur Metabolism in Phototrophic Organisms.* Springer: Netherlands, pp 337–355.

Frigaard N-U, Bryant DA. (2006). Chlorosomes: antenna organelles in photosynthetic green bacteria. *Complex Intracellular Structures in Prokaryotes*. Springer: Netherlands, pp 79–114.

Frigaard NU, Bryant DA. (2001). Chromosomal gene inactivation in the green sulfur bacterium *Chlorobium tepidum* by natural transformation. *Appl Environ Microbiol* **67**: 2538–2544.

Fuhrman JA. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**: 195–202.

Grissa I, Vergnaud G, Pourcel C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.

Gregersen LH, Habicht KS, Peduzzi S, Tonolla M, Canfield D E, Miller M *et al.* (2009). Dominance of a clonal green sulfur bacterial population in a stratified lake. *FEMS Microbiol Ecol* **70**: 30–41.

Guy L, Kultima JR, Andersson SG. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**: 2334–2335.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Habicht KS, Miller M, Cox RP, Frigaard NU, Tonolla M, Peduzzi S *et al.* (2011). Comparative proteomics and activity of a green sulfur bacterium through the water column of Lake Cadagno, Switzerland. *Environ Microbiol* **13**: 203–215.

Han MV, Zmasek CM. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**: 356.

Harada J, Mizoguchi T, Satoh S, Tsukatani Y, Yokono M, Noguchi M *et al.* (2013). Specific gene bciD for C7-Methyl oxidation in bacteriochlorophyll e biosynthesis of brown-colored green sulfur bacteria. *PLoS One* **8**: e60026.

Haring V, Billington SJ, Wright CL, Huggins AS, Katz ME, Rood JI. (1995). Delineation of the virulence-related locus (Vrl) of *Dichelobacter nodosus*. *Microbiology* **141**: 2081–2089.

Hirabayashi H, Ishii T, Takaichi S, Inoue K, Uehara K. (2004). The role of carotenoids in the photoadaptation of the brown-colored sulfur bacterium *Chlorobium phaeobacteroides*. *Photochem Photobiol* **79**: 280–285.

Kammler M, Schon C, Hantke K. (1993). Characterization of the ferrous iron uptake system of *Escherichia coli*. *J Bacteriol* **175**: 6212–6219.

Kim MS, Hong SS, Park K, Myung H. (2013). Genomic analysis of bacteriophage PBECO4 infecting *Escherichia coli* O157:H7. *Arch Virol* **158**: 2399–2403.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Hamamura N, Heidelberg JF *et al.* (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J* **5**: 1262–1278.

Knaust F, Kube M, Reinhardt R, Rabus R. (2007). Analyses of the vrl gene cluster in *Desulfococcus multivorans*: homologous to the virulence-associated locus of the ovine footrot pathogen *Dichelobacter nodosus* strain A198. *J Mol Microbiol Biotechnol* **13**: 156–164.

Lanfear R, Calcott B, Ho SY, Guindon S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* **29**: 1695–1701.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101**: 11013–11018.

Llorens-Mares T, Yooseph S, Goll J, Hoffman J, Vila-Costa M, Borrego CM *et al.* (2015). Connecting biodiversity and potential functional role in modern euxinic environments by microbial metagenomics. *ISME J* **9**: 1648–1661.

Llorens-Marès T, Triadó-Margarit X, Borrego CM, Dupont CL, Casamayor EO. (2016). High bacterial diversity and phylogenetic novelty in dark euxinic freshwaters analyzed by 16S tag community profiling. *Microb Ecol* **71**: 566–574.

Lohse DL, Denu JM, Dixon JE. (1995). Insights derived from the structures of the Ser/Thr phosphatases calcineurin and protein phosphatase 1. *Structure* **3**: 987–990.

Lorenzi HA, Hoover J, Inman J, Safford T, Murphy S, Kagan L *et al.* (2011). TheViral Metagenome Annotation Pipeline (VMGAP): an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand Genomic Sci* **4**: 418–429.

Luo CW, Tsementzi D, Kyrpides NC, Konstantinidis KT. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* **6**: 898–901.

Maresca JA, Graham JE, Bryant DA. (2008a). The biochemical basis for structural diversity in the carotenoids of chlorophototrophic bacteria. *Photosynth Res* **97**: 121–140.

Maresca JA, Romberger SP, Bryant DA. (2008b). Isorenieratene biosynthesis in green sulfur bacteria requires the cooperative actions of two carotenoid cyclases. *J Bacteriol* **190**: 6384–6391.

Martin HG, Ivanova N, Kunin V, Warnecke F, Barry K W, McHardy A C *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.

Martiny AC, Huang Y, Li WZ. (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340–1347.

Melendrez MC, Lange RK, Cohan FM, Ward DM. (2011). Influence of molecular resolution on sequence-based discovery of ecological diversity among *Synechococcus* populations in an alkaline siliceous hot spring microbial mat. *Appl Environ Microb* **77**: 1359–1367.

Montesinos E, Guerrero R, Abella C, Esteve I. (1983). Ecology and physiology of the competition for light between *Chlorobium limicola* and *Chlorobium phaeobacteroides* in natural habitats. *Appl Environ Microbiol* **46**: 1007–1016.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**: 760–766.

Oberto J. (2013). SyntTax: a web server linking synteny to prokaryotic taxonomy. *BMC Bioinformatics* **14**: 4.
Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.
Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. (2008). GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* **9**: 376.
Overmann J, Tuschak C. (1997). Phylogeny and molecular fingerprinting of green sulfur bacteria. *Arch Microbiol* **167**: 302–309.
Palenik B, Ren Q, Tai V, Paulsen IT. (2009). Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ Microbiol* **11**: 349–359.
Polz MF, Alm EJ, Hanage WP. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* **29**: 170–175.
Pruesse E, Peplies J, Glockner FO. (2012). SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.
R Core Team. (2014). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria.
Richter M, Rossello-Mora R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* **106**: 19126–19131.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.
Rohwer F, Thurber RV. (2009). Viruses manipulate the marine environment. *Nature* **459**: 207–212.
Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of Prochlorococcus clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* **107**: 16184–16189.
Saunders NJ, Boonmee P, Peden JF, Jarvis SA. (2005). Inter-species horizontal transfer resulting in core-genome and niche-adaptive variation within *Helicobacter pylori*. *BMC Genomics* **6**: 9.
Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N *et al.* (2009). Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**: 258–262.
Soontharapirakkul K, Promden W, Yamada N, Kageyama H, Incharoensakdi A, Iwamoto-Kihara A *et al.* (2011). Halotolerant cyanobacterium *Aphanothece halophytica* contains an Na+-dependent F1F0-ATP synthase with a potential role in salt-stress tolerance. *J Biol Chem* **286**: 10169–10176.
Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.
Summers AO. (2006). Genetic linkage and horizontal gene transfer, the roots of the antibiotic multi-resistance problem. *Anim Biotechnol* **17**: 125–135.
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.
Van Gemerden H, Mas J. (1995). Ecology of phototrophic sulfur bacteria. In: Blankenship R, Madigan M, Bauer C (eds). *Anoxygenic Photosynthetic Bacteria*. Springer: Netherlands, pp 49–85.
Villafranca JE, Kissinger CR, Parge HE. (1996). Protein serine/threonine phosphatases. *Curr Opin Biotechnol* **7**: 397–402.
Westra ER, Buckling A, Fineran PC. (2014). CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* **12**: 317–326.
Wiedenbeck J, Cohan FM. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* **35**: 957–976.
Wu XA, Monchy S, Taghavi S, Zhu W, Ramos J, van der Lelie D. (2011). Comparative genomics and functional analysis of niche-specific adaptation in *Pseudomonas putida*. *FEMS Microbiol Rev* **35**: 299–323.
Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC *et al.* (2013). Abundant SAR11 viruses in the ocean. *Nature* **494**: 357–360.

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