The ISME Journal (2011) 5, 1262–1278

www.nature.com/ismej

ORIGINAL ARTICLE

Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential

Christian G Klatt[1], Jason M Wood[1], Douglas B Rusch[2], Mary M Bateson[1], Natsuko Hamamura[1], John F Heidelberg[3], Arthur R Grossman[4], Devaki Bhaya[4], Frederick M Cohan[5], Michael Kühl[6,7], Donald A Bryant[8,9] and David M Ward[1]

[1]*Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT, USA;* [2]*J Craig Venter Institute, Rockville, MD, USA;* [3]*Department of Biological Sciences, Philip K Wrigley Marine Science Center, University of Southern California, Avalon, CA, USA;* [4]*Department of Plant Biology, Carnegie Institution for Science, Stanford University, Stanford, CA, USA;* [5]*Department of Biology, Wesleyan University, Middletown, CT, USA;* [6]*Department of Biology, Marine Biological Section, University of Copenhagen, Helsingør, Denmark;* [7]*Plant Functional Biology Climate Change Cluster, University of Technology, Sydney, New South Wales, Australia;* [8]*Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA and* [9]*Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT, USA*

Phototrophic microbial mat communities from 60 °C and 65 °C regions in the effluent channels of Mushroom and Octopus Springs (Yellowstone National Park, WY, USA) were investigated by shotgun metagenomic sequencing. Analyses of assembled metagenomic sequences resolved six dominant chlorophototrophic populations and permitted the discovery and characterization of undescribed but predominant community members and their physiological potential. Linkage of phylogenetic marker genes and functional genes showed novel chlorophototrophic bacteria belonging to uncharacterized lineages within the order Chlorobiales and within the Kingdom Chloroflexi. The latter is the first chlorophototrophic member of Kingdom Chloroflexi that lies outside the monophyletic group of chlorophototrophs of the Order Chloroflexales. Direct comparison of unassembled metagenomic sequences to genomes of representative isolates showed extensive genetic diversity, genomic rearrangements and novel physiological potential in native populations as compared with genomic references. *Synechococcus* spp. metagenomic sequences showed a high degree of synteny with the reference genomes of *Synechococcus* spp. strains A and B′, but synteny declined with decreasing sequence relatedness to these references. There was evidence of horizontal gene transfer among native populations, but the frequency of these events was inversely proportional to phylogenetic relatedness.

The ISME Journal (2011) **5,** 1262–1278; doi:10.1038/ismej.2011.73; published online 23 June 2011
Subject Category: microbial population and community ecology
Keywords: cyanobacteria; Chloroflexi; community structure and function; metagenomics

Introduction

The cyanobacterial mats of alkaline siliceous hot springs in Yellowstone National Park (Supplementary Figures 1A and 1B) have been studied for several decades as models for understanding the composition, structure and function of microbial communities (Brock, 1978; Ward *et al.*, 1987, 1992, 2002, 2011a). Simple and stable microbial communities containing dense populations of unicellular cyanobacteria (*Synechococcus* spp.) form in effluent channels of these springs between temperatures of 71 and 75 °C (the upper temperature limit of the phototrophic mats), and ~50 °C.

Analysis of environmental 16S ribosomal RNA (rRNA) gene sequences showed a poor relationship between initially cultivated isolates and predominant native populations. For instance, the predominant *Synechococcus* spp. of these mats (A/B lineage) had ⩽92% nucleotide identity at the 16S rRNA locus to the cultivated representatives available at that time (Ward *et al.*, 1990). Similarly, based on cultivation and pigment analyses (Bauld and Brock, 1973; Pierson and Castenholz, 1974), it was once thought that *Chloroflexus* spp., which in culture use bacteriochlorophyll (BChl)-*c* and BChl-*a* to support

Correspondence: CG Klatt, Department of Land Resources and Environmental Sciences, Montana State University, 334 Leon Johnson Hall, PO Box 173120, Bozeman, MT 59717-3120, USA.
E-mail: klatt.christian@gmail.com
Received 15 February 2011; revised 9 May 2011; accepted 9 May 2011; published online 23 June 2011

photo-heterotrophy (Pierson and Castenholz, 1974) or photo-autotrophy (Holo and Sirevåg, 1986; Strauss and Fuchs, 1993), were the dominant anoxygenic phototrophic bacteria in these mats. However, environmental 16S rRNA studies uncovered the importance of *Roseiflexus* spp. (Nübel *et al.*, 2002), organisms that contain BChl-*a* but lack BChl-*c* and grow axenically as photo-heterotrophs (Hanada *et al.*, 2002), although they possess genes encoding the enzymes of the 3-hydroxypropionate autotrophic pathway (Klatt *et al.*, 2007). In these cases, the inference of chlorophototrophic physiologies (that is, Chls are obligately required for phototrophy, in contrast to retinal-mediated proton translocation) could be made because oxygenic chlorophototrophs and anoxygenic chlorophototrophic Chloroflexales comprise monophyletic groups defined by 16S rRNA phylogeny. These predictions were confirmed with more recent cultivation and genomic analyses of *Synechococcus* spp. and *Roseiflexus* spp. isolates closely related to native mat populations (Allewalt *et al.*, 2006; Bhaya *et al.*, 2007; van der Meer *et al.*, 2010).

The inference of functional potential from 16S rRNA phylogeny is more problematic when sequences do not belong to groups that are monophyletic with respect to function. For instance, based on the observation that some 16S rRNA sequences retrieved from the mats fell just outside the monophyletic clade of known Chlorobiales, Ferris and Ward (1997) suggested the possible presence of bacteria closely related to green sulfur bacteria. Targeted analyses of photosynthetic reaction center genes provided evidence in support of this hypothesized functional group (Bryant *et al.*, 2007), but there was no way to associate the functional genes directly with the phylogenetic marker gene. Despite the successful retrieval of Chlorobiales from other thermal environments (Wahlund *et al.*, 1991; Madigan *et al.*, 2005), this organism has to date evaded cultivation. Interestingly, the search for photosynthetic reaction center genes in the mats led to the discovery of the first known chlorophototrophic member of Kingdom Acidobacteria, *Candidatus* Chloracidobacterium thermophilum (*Ca.* C. thermophilum) (Bryant *et al.*, 2007; usage of 'kingdom' for major Domain sub-lineages *sensu* Ward *et al.*, 2011b). The inference of potential chlorophototrophy was based on the discovery of a metagenomic clone containing an insert of mat DNA with both phylogenetic marker and functional genes. Because cultivated acidobacteria were not previously known to be phototrophic, inferences based on 16S rRNA data concerning the potential for phototrophy could not have been made before this discovery. Studies of an enrichment culture of *Ca.* C. thermophilum (Bryant *et al.*, 2007) and its genome (Bryant *et al.*, 2011) confirmed the inferences made from genetic data.

In this study, we used assembly of metagenomic sequences, combined with oligonucleotide frequency distributions and cluster analysis of scaffolds, to identify phylogenetically distinctive populations inhabiting the Octopus Spring and Mushroom Spring mats. Oligonucleotide frequency patterns contain phylogenetic information (Pride *et al.*, 2003; Teeling *et al.*, 2004) and have been used as a tool to determine phylogenetic signatures in metagenomic data from microbial communities (Woyke *et al.*, 2006; Wilmes *et al.*, 2008; Dick *et al.*, 2009; Inskeep *et al.*, 2010). Annotation of open reading frames was used to identify phylogenetically and functionally informative genes in the scaffolds. We used the sequenced genomes of selected organisms, many of which have been cultivated from these or similar hot spring environments, and some of which are close relatives of predominant native populations in these mats, to 'recruit' metagenomic sequences (Supplementary Table 1). This combined approach enabled us to (i) discover new major populations of uncultivated community members; (ii) explore differences in the functional potential of native populations as compared with closely related isolates and (iii) observe differences in genomic content and synteny among closely related populations. This study also created a foundation for a companion study using metatranscriptomics to describe *in situ* gene expression in the chlorophototrophic taxa (Liu *et al.*, 2011), the results of which strongly support our functional inferences and expand upon *in situ* gene expression studies of these mats (Steunou *et al.*, 2006, 2008; Jensen *et al.*, 2011).

Materials and methods

The experimental approaches have been presented in the following sections. Technical details of the methods used have been provided in Supplementary Information Section 3.

Collection, preliminary sequence analysis and metagenomic sequencing

Microbial mats were collected from Mushroom Spring (44.5386°N, 110.7979°W) on 2 October 2003 and from Octopus Spring (44.5340°N, 110.7978°W) on 5 November 2004 (Bhaya *et al.*, 2007) at sites with average temperatures of ~60 and ~65 °C. *Synechococcus* spp. genotypes B′ and A are the respective dominant cyanobacterial 16S rRNA sequences at these temperatures. Samples were collected and sectioned vertically into approximately 1-mm-thick layers, which were frozen and then stored at −80 °C until further analysis. After enzymatic lysis of cells in the top green layer, DNA was extracted and sequences were characterized by PCR amplification of cyanobacterial 16S rRNA genes and subsequent analysis by denaturing gradient-gel electrophoresis to verify the presence of *Synechococcus* A- and B′-like genotypes (Supplementary Figure 3). The extracted

DNA was sheared into ~1–3 and ~10–12-kb fragments, which were used to prepare four metagenomic libraries corresponding to the low- and high-temperature samples from both Octopus and Mushroom Springs. End sequences of cloned inserts were produced by Sanger sequencing at the J Craig Venter Institute (JCVI, Rockville, MD, USA).

Metagenome assembly and annotation
The metagenomic sequences were assembled into scaffolds using the Celera assembler (Miller *et al.*, 2008), with the 'error' (that is, mismatch) rate set to 8% for the purpose of assembling non-identical close relatives, and the utgGenomeSize set to 2 000 000. The phylogenetic and functional marker genes in assemblies were identified using the programs AMPHORA (Wu and Eisen, 2008), the JCVI annotation pipeline (Tanenbaum *et al.*, 2010) or BLAST (Altschul *et al.*, 1990), using known reference sequences as queries. All annotations are inferences based on multiple lines of evidence produced using the tools listed above, but their functions are considered hypotheses for future biochemical characterization.

Clustering and characterization of assemblies
Oligonucleotide patterns were determined to obtain phylogenetic signals (Teeling *et al.*, 2004) by counting the frequencies of all possible tri-, tetra-, penta- and hexa-nucleotide combinations for each scaffold $\geqslant$20 000 bp. Frequency counts were normalized by the length of the respective scaffold and subjected to *k*-means clustering (Kanungo *et al.*, 2002) using the *a priori* value of *k* equal to 8 (see Supplementary Information Section 3 for rationale). Scaffolds that clustered together with $\geqslant$90% of 100 Monte Carlo trials were mapped using Cytoscape (Shannon *et al.*, 2003). Many scaffolds formed associations with core clusters at less stringent thresholds, but, except where noted, these were not included in the cluster analysis described here.

BLASTN recruitment and synteny with reference genomes
Metagenomic sequences were used as queries in a custom BLASTN search to a selected database of 20 genomes from organisms isolated from thermal springs, known to be functionally and/or phylogenetically related to indigenous mat populations and/or processes, or representative of phylogenetic groups not otherwise included (Supplementary Table 1). The percent nucleotide identity (% NT ID) of metagenomic sequences relative to the reference genome that recruited them was used to identify those that could be confidently associated with the reference organism, taking into account the % NT ID between the genomes of strains of named species and genera (approximately >70% NT ID among species of named genera; see Supplementary Information Section 3 and Supplementary Figure 7). The end sequences of a clone were considered 'jointly recruited' if the sequences were recruited by the same genome, or were considered 'disjointly recruited' if their end sequences were recruited by different reference genomes. The end sequences of jointly recruited clones were considered 'syntenous' when the sequences had the same orientation as the reference genome and were separated by a distance that was similar to the size of the DNA fragments used to construct the metagenomic library. Jointly recruited sequences that did not meet both of these criteria were considered 'non-syntenous.' The details of this process are described in Supplementary Information Section 3.

Results

Sanger sequencing of samples from all sites and temperatures yielded 167 Mb of metagenomic sequence data. Assembly resulted in 5769 scaffolds, totaling 33 Mb, which were produced from 67 Mb (40%) of the total sequence data set. Cluster analysis of oligonucleotide frequencies was used to characterize 394 scaffolds that were $\geqslant$20 000 bp in length, totaling 20.2 Mb (Table 1). Prior to assembly, recruitment by reference genomes above the specified % NT ID cutoffs indicated in Table 2 accounted for 102 Mb of the total sequence data set (61%). Scaffold clusters accounted for an additional 13 Mb (7.8%) of the total unassembled metagenomic sequences that were not recruited to reference genomes above % NT ID cutoffs. Thus, we could confidently assign 69% of the total metagenomic

Table 1 Assembly statistics of scaffold clusters $\geqslant$20 000 bp in length

Cluster	*Phylogenetic affiliation*	*Number of scaffolds*	*Mb sequence in cluster*	*Mean scaffold length (kb)*	*Mean depth of coverage (read depth)*
1	*Synechococcus* spp.	68	3.54	52.1	4.7
2	*Roseiflexus* spp.	78	4.26	54.7	3.7
3	*Chloroflexus* spp.	59	1.91	32.4	1.7
4	*Ca.* C. thermophilum-like organisms	10	3.20	319	3.7
5	Chlorobiales-like organisms	32	2.82	88.0	5.7
6	Anaerolineae-like organisms	46	2.31	50.3	3.2
7	Unknown Cluster-1	27	1.42	52.4	2.2
8	Unknown Cluster-2	39	1.62	41.6	4.4

Table 2 Comparison of metagenomic analyses based on genome recruitment and assembly[a]

Reference genome	*Reference genome size (Mb)*	*Mean ± s.d. % NT ID of recruited sequences*	*% NT ID range used in analyses*[b]	*% of individual metagenomic sequences recruited*[c]					*% of total assembled sequences*
				MS low	*MS high*	*OS low*	*OS high*	*Total*	
Synechococcus sp. strain-A	2.93	94.1 ± 10.8	92–100	7.03	22.1	6.36	21.1	11.0	9.78
Synechococcus A′[d]	—	—	83–92	0.63	2.92	1.23	1.14	1.57	1.15
Synechococcus sp. strain-B′	3.04	95.6 ± 7.4	90–100	22.1	1.15	24.9	1.10	17.7	8.75
Roseiflexus sp. strain-RS1	5.80	84.6 ± 16.1	80–100	16.0	9.58	7.52	14.9	9.17	9.42
Chloroflexus sp. strain-396-1	5.2	90.4 ± 6.6	65–100	0.77	16.0	0.99	1.91	4.63	4.62
Ca. C. thermophilum	3.7	78.5 ± 11.0	70–100	9.66	2.15	10.4	6.69	8.10	9.23
C. thalassium	3.29	63.4 ± 5.3	50–100	6.46	2.11	11.6	3.48	8.41	7.65
T. roseum	2.93	64.0 ± 12.7	75–100	0.21	0.66	0.03	0.15	0.21	0.05
T. thermophilus	2.11	73.6 ± 11.6	75–100	0.08	0.80	0.22	0.27	0.35	0.13
T. yellowstonii	2.00	73.4 ± 8.0	75–100	0.46	0.15	0.04	0.12	0.11	0.03

Abbreviation: % NT ID, percent nucleotide identity.
[a]All results met the criterion of having an e-value more significant than 10^{-10} for the WU-BLASTN parameters M = 3 and N = −2, and a database size of 68 Mb.
[b]Relative to the reference genome.
[c]The abbreviations for the four metagenomes are as follows: MS, Mushroom Spring; OS, Octopus Spring; low, 60 °C average; high, 65 °C average.
[d]Recruited by the *Synechococcus* sp. strain-A genome with 83–92% NT ID.

sequences to known taxa or novel phylogenetic clusters by combining these approaches; 31% of the metagenomic sequences are currently of unknown origin. Consistent with the failure to detect 18S rRNA sequences at these temperatures (Liu *et al.*, 2011), no eukaryotic sequences were observed. Aside of a relative underrepresentation of sequences from *Ca.* C. thermophilum (Supplementary Figure 4), pyrosequencing of SSU rDNA amplicons from environmental DNA showed taxonomic profiles that were similar to those for cDNA sequences produced from rRNAs for the meta-transcriptome studies (Liu *et al.*, 2011). Sequences likely originating from archaea were present, but were not in high abundance in the upper photic layer of these mats.

Major populations and their functional potential

Clustering on the basis of oligonucleotide frequency showed eight scaffold clusters (Figure 1 and Table 1). Phylogenetic affiliations of these clusters were inferred from (i) direct co-clustering with reference genomes (Figure 1); (ii) clusters being composed of sequences recruited by a reference genome at high % NT ID (Figure 2 and Table 2 and Supplementary Table 7) and (iii) the presence of phylogenetically informative marker genes within the clusters (Figure 1 and Table 3). The metabolic potentials of organisms associated with these clusters were inferred from the functional genes they contained.

(i) Oxygenic Chlorophototrophs. Cluster-1 contained scaffolds that were strongly associated with the *Synechococcus* spp. strains A and B′ genomes, and included cyanobacterial phylogenetic marker genes and functional genes that were indicative of oxygenic photosynthesis, the Calvin–Benson–Bassham cycle and genes involved in nitrogen and phosphorus acquisition that were described previously (Steunou *et al.*, 2006, 2008; Bhaya *et al.*, 2007). Most (86%) of these metagenomic sequences were jointly recruited and were more closely related to either the *Synechococcus* sp. strain-A or B′ genome (Supplementary Figure 8). The cyanobacterial scaffolds in these bins accounted for 19.7% of the total assembled sequence data (Table 2), which was the largest amount assigned to any particular group of organisms. Differences between these cyanobacterial scaffolds and the *Synechococcus* spp. isolate genomes were found and provide evidence for functional diversity. Scaffolds from native *Synechococcus* sp. strain-A-like populations contained genes encoding *feoAB* (involved in Fe^{2+} transport) and genes homologous to the characterized bacterial enzymes urea carboxylase (*ureA*) and allophanate hydrolase (*atzF*; involved in the degradation of urea into ammonia and CO_2), both of which are not found in the *Synechococcus* sp. strain-A genome (Supplementary Table 9) (Kanamori *et al.*, 2004; Cheng *et al.*, 2005).

(ii) Filamentous Anoxygenic Chlorophototrophs. Cluster-2 scaffolds had similar oligonucleotide frequencies to both *Roseiflexus* sp. strain RS1 and *Roseiflexus castenholzii* genomes, and they predominantly comprised sequences recruited by the *Roseiflexus* sp. strain RS1 genome (98%, with a mean of 95% NT ID; Supplementary Table 7). Many conserved phylogenetic marker genes, with sequences almost identical to homologs in the *Roseiflexus* sp. strain RS1 genome, were found on Cluster-2 scaffolds (Table 3). Most of the Cluster-2 sequences were jointly recruited by the *Roseiflexus* sp. strain RS1 genome with more than 80% NT ID (Figure 2), which was above the mean from a comparison of *Roseiflexus* sp. strain RS1 and *R. castenholzii* homologs (Supplementary information Section 3). This observation implies that a large

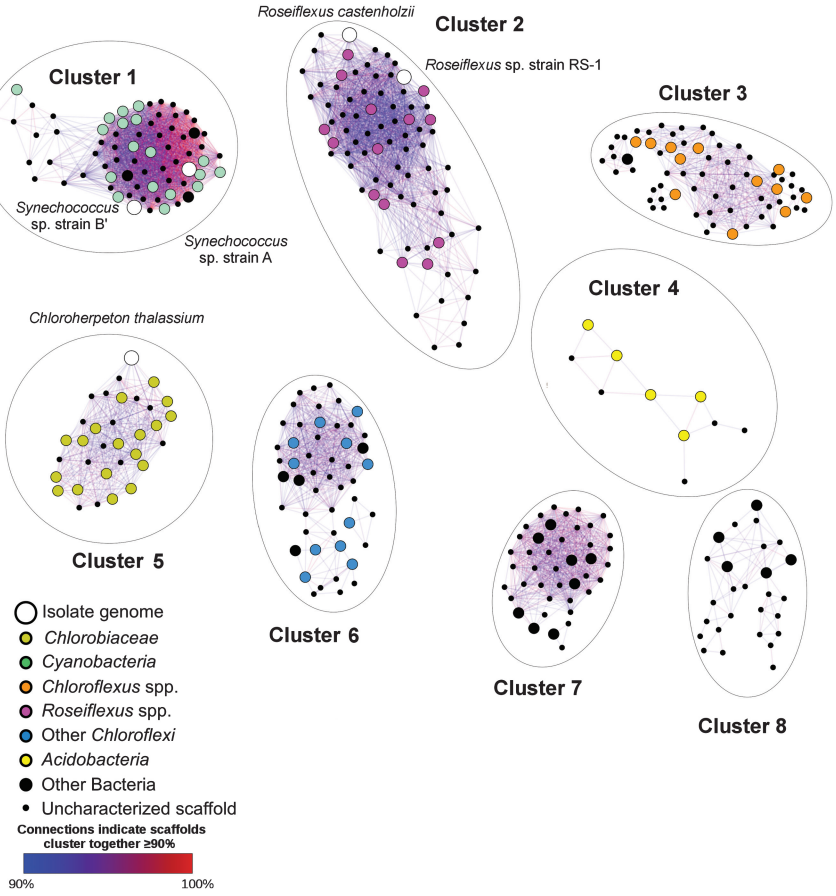

Figure 1 A network map of the core scaffold clusters observed in the Celera assemblies. Scaffolds with similar oligonucleotide frequency profiles that group together in the same cluster are connected by lines colored to indicate the percentage of times they cluster together (in ⩾90% of 100 trials). The isolate genomes included in this analysis are indicated by large white circles, whereas metagenomic scaffolds that contain characterized phylogenetic marker genes are indicated by medium-sized circles colored according to taxonomic grouping. The area of each ellipse is proportional to the amount of metagenomic sequence data contained within each respective scaffold cluster.

proportion of scaffolds are represented by sequences from a diverse assemblage of *Roseiflexus* spp., and is consistent with the diversity of sequences directly recruited by the *Roseiflexus* sp. strain RS1 genome by BLASTN independently of a metagenomic assembly (Figure 2). One scaffold in Cluster-2 contained a diagnostic fused *pufLM* gene that encodes both of the type-2 photosystem reaction center polypeptides (*pufL* and *pufM* are characteristically fused in *Roseiflexus* spp.; Youvan *et al.*, 1984; Yamada *et al.*, 2005) (Figure 3). There were *recA* sequences highly similar to the *Roseiflexus* sp. strain RS1 *recA* in the metagenome (Supplementary Figure 10), but these were not encoded on the large scaffolds included in the cluster analysis. Suggesting that these organisms have the capability to fix inorganic carbon, Cluster-2 also contained eight open reading frames homologous to *Roseiflexus* spp. genes encoding key enzymes in the 3-hydroxypropionate pathway (Klatt *et al.*, 2007). Like *Roseiflexus* sp. strain RS1, *Roseiflexus* spp. native to the mat may have the potential to use H_2 as an electron donor because Cluster-2 scaffolds contain homologs of bidirectional [NiFe]-hydrogenases (*hydAB*) (Table 3; van der Meer *et al.*, 2010). One open reading frame homologous to a *nifH* gene in the *Roseiflexus* sp. strain RS1 genome was also observed.

Oligonucleotide compositions of Cluster-3 scaffolds were not similar to any sequenced isolate genomes above the 90% cutoff; however, the phylogenetic and functional marker genes they contained indicated that these scaffolds were contributed by *Chloroflexus* spp. Most (82%) of the metagenomic sequences comprising these scaffolds were recruited at a high degree of similarity (Supplementary Table 7) by the genome of *Chloroflexus* sp. strain 396-1, which is currently the most representative cultivated organism compared with the native *Chloroflexus* spp. in these mats (van der Meer *et al.*, 2010). Most (85%) of the metagenome sequences recruited by the *Chloroflexus* sp. strain 396-1 genome were jointly recruited sequences that had a mean % NT ID of 91.3 ± 5.3% (Figure 2). One Cluster-3 scaffold contained a *pufC* homolog adjacent to *bchP* and *bchG*, consistent with the

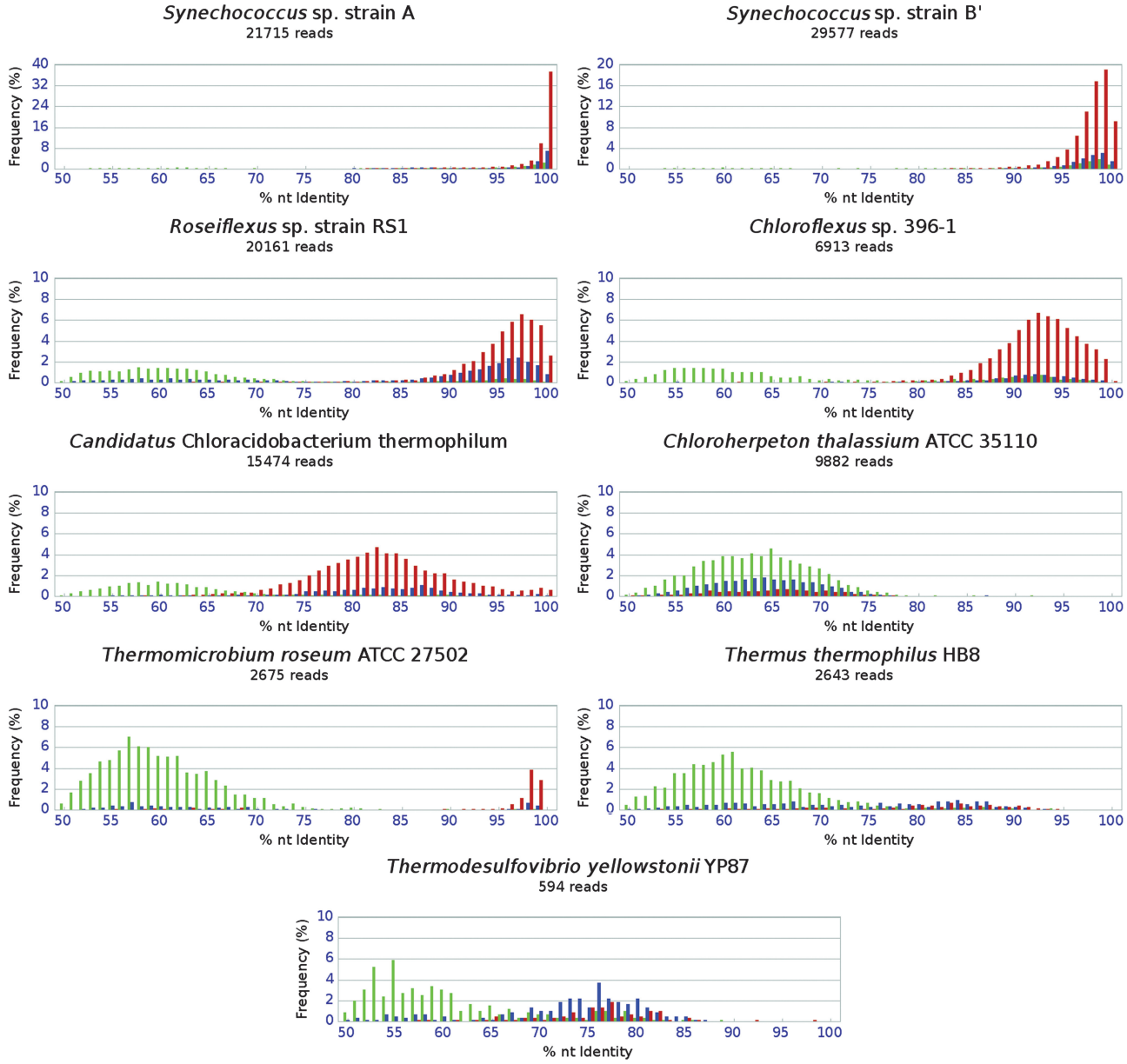

Figure 2 Histograms of disjointly recruited (green), jointly recruited syntenous (red) and jointly recruited, non-syntenous (blue) metagenomic sequences that can be associated confidently with a reference genome presented as a function of their % NT ID relative to the reference genomes that recruited them in the BLASTN analysis. % NT ID, percent nucleotide identity.

Chloroflexus sp. 396-1 genome (93% NT ID, 100% amino-acid identity) (Figure 3). Overlapping metagenome sequences were missing upstream from the *pufC* open reading frame, so it could not be confirmed whether the native *Chloroflexus* spp. have the *pufBAC* operon structure observed in other *Chloroflexus* spp. (Watanabe *et al.*, 1995). However, the colocalized *bchG* and *bchP* genes and high % NT ID to *Chloroflexus* sp. 396-1 are consistent with this inference derived from oligonucleotide clustering (Figure 3). Homologs of genes involved in both BChl-*c* and BChl-*a* biosynthesis were present in Cluster-3, indicating that the native *Chloroflexus* spp. are physiologically similar to known isolates with respect to light-harvesting strategies (Bryant and Frigaard, 2006; Frigaard and Bryant, 2006; Bryant *et al.*, 2011) (Table 3). Sequences encoding two key enzymes in the 3-hydroxypropionate pathway, and most closely related to homologs in the *Chloroflexus* sp. strain 396-1 genome, were present on Cluster-3 scaffolds. This suggests that *Chloroflexus* spp. in the mats may be capable of carbon fixation by the 3-hydroxypropionate pathway. Cluster-3 contained a homolog of sulfide-quinone oxidoreductases (*sqr*) in *Chloroflexus* spp., which suggested that these organisms might oxidize sulfide to polysulfides (Bryant *et al.*, 2011).

Table 3 Phylogenetic marker genes and functional genes in assembly clusters

Cluster: phylogeny	*Phylogenetic genes*[a]	*Functional genes*[b]	*Pathways/functions*
1. *Synechococcus* spp.	16S rRNA	*chlB*, *chlG*, *chlJ*, *psaA*, *psaL*, *psbB*, *psbC*	Oxygenic chlorophototrophy
	pyrG	*cpcABCDEFG*, *apcABC*, *nblA*	
	recA	*rbcX*, *cbbS*, *cbbL,* FBP aldolase, PRK	Calvin–Benson–Bassham cycle
	rplB, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*,	*ctaC*, *ctaD*, *ctaE*	Oxygen respiration
	rplL, *rplN*, *rplM*, *rplP*, *rplT*	*nifH*	Nitrogen fixation
	rpoB	*narB*	Nitrate metabolism
	rpsB, *rpsE*, *rpsI*, *rpsJ*, *rpsK*,	*amtB*	Ammonium transport
	rpsM, *rpsS*	*pstABCS*	Phosphate transport
	smpB	*phnCEGHILJ*	Phosphonate metabolism
	tsf		
2. *Roseiflexus* spp.	16S rRNA	*bchB*, *bchC*, *bchD*, *bchE*, *bchF*, *bchG*, *bchH*-1,	BChl-*a* biosynthesis
	nusA	*bchH*-2, *bchI*, *bchJ*, *bchL*, *bchM*, *bchN*, *bchP*,	
	rplA, *rplB*, *rplC*, *rplD*, *rplE*,	*bchT-like*, *bchX*, *bchY*, *bchZ*	
	rplN, *rplP*, *rplT*	*pufB*, *pufC*, *pufLM*	Anoxygenic chlorophototrophy
	rpsB, *rpsC*, *rpsI*, *rpsS*	*cyoB*, *coxA*, *coxB*	Oxygen respiration
	tsf	*mch*, *mcl*, *mcr*, *mct*, *meh*, *pcs*, *smtA*, *smtB*	3-Hydroxypropionate pathway
		hydAB	Hydrogenase
		amtB	Ammonium transport
3. *Chloroflexus* spp.	*infC*	*bchB*, *bchG*, *bchH*-1, *bchL*, *bchN*, *bchP*, *bchS*,	BChl-*a* and BChl-*c* biosynthesis
	nusA	*bchY, slr1923*-homolog	
	pgk	*mcr*	3-Hydroxypropionate pathway
	recA	*pcs, pufC*	Anoxygenic chlorophototrophy
	rplT	*cyoB*, *coxC*	Oxygen respiration
		sqr	Reduced S oxidation
		hydAB	Hydrogenase
		pstABCS	Phosphate transport
4. *Ca.* C. thermophilum-like organisms	16S rRNA	*bchD*, *bchE*, *bchF*, *bchL*, *bchU*, *acsF*, *bchI*,	BChl-*a* and BChl-*c* biosynthesis
	dnaG	*bchM*, *bchX*, *bchB*, *bchN*, *bchK*, *bchC*, *bchZ*,	
	frr	*bchY*, *bchG*, *bchP*	
	nusA	*csmA*	Chlorosome biosynthesis
	pgk	*amtB*	Ammonium transport
	pyrG		
	recA		
	rplE, *rplM*, *rplT*		
	rpsB, *rpsI*		
	smpB		
5. Chlorobiales-like organisms	16S rRNA	*bchB*, *bchC*, *bchD*, *bchF*, *bchG*-homolog,	BChl-*a* and BChl-*d* biosynthesis
	frr	*bchH*, *bchH*-homolog, *bchI*, *bchK*, *bchL*,	
	rplB, *rplC*, *rplD*, *rplF*, *rplK*, *rplL*,	*bchP*, *bchR*, *bchX*, *bchY*, *bchZ*	
	rplN, *rplP*, *rplS*, *rplT*	*fmoA*	Anoxygenic chlorophototrophy
	rpsE, *rpsI*, *rpsJ*, *rpsK*, *rpsM*	*pscA*, *pscB*, *csmC*	
	smpB	*norE*-like cyt. c oxidase	Oxygen respiration
	tsf	*amtB*	Ammonium transport
6. Anaerolineae-like organisms	16S rRNA	*bchF*, *bchG*, *bchI*, *bchP*, *bchS*-like, *bchX*,	BChl-*a* biosynthesis
	frr	*bchY*, *bchZ*	
	infC	*pufL*, *pufM*, *pufC*	Anoxygenic chlorophototrophy
	pgk	*coxA*, *coxB*	Oxygen respiration
	pyrG	*sqr*	Reduced S oxidation
	recA		
	rplA, *rplK*, *rplL*, *rplM*, *rplT*		
	rpoB		
	rpsI, *rpsM*		
	smpB		
	tsf		
7. Unknown Cluster-1	*dnaG*	*glcD*, *glcE*	Glycolate oxidation
	rplE, *rplN*, *rplP*, *rpsC*, *rpsK*,	*acs*	Acetate metabolism
	rpsM, *rpsS*	*coxA*, *coxB*, *coxC*	Oxygen respiration
8. Unknown Cluster-2	*dnaG*	*coxA*, *coxB*, *coxC*	Oxygen respiration
	infC		
	pgk		
	pyrG		
	recA		
	rplB, *rplC*, *rplK*, *rplL*, *rplM*,		
	rplN, *rplP*		
	rpsC, *rpsE*, *rpsJ*, *rpsM*		

[a]Phylogenetic marker genes identified using AMPHORA and/or phylogenetic analysis.
[b]Functional genes identified using BLAST and annotated using hidden Markov models, genomic context and/or phylogenetic analysis (see Supplementary Information Section 3).

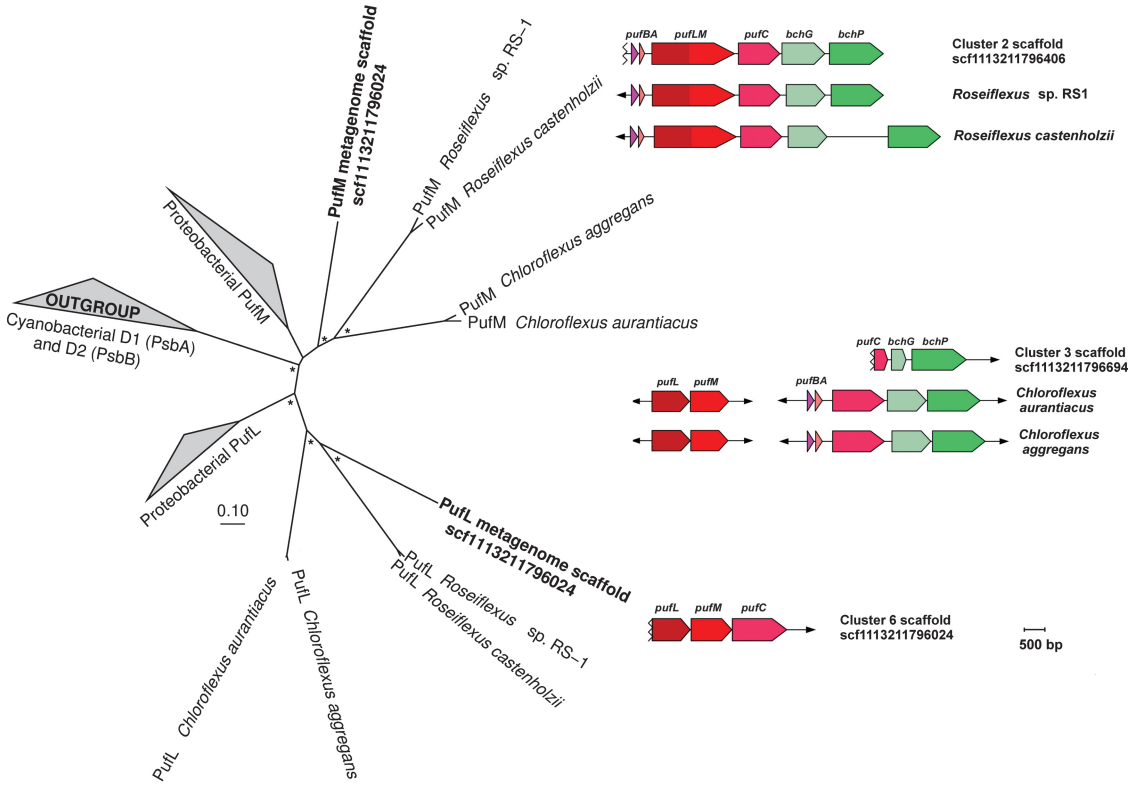

Figure 3 The PufL and PufM phylogeny and genomic context. The neighbor-joining phylogenetic tree of the PufL and PufM sequences from a novel Chloroflexi metagenomic scaffold from Cluster-6 and from sequenced genomes; asterisks at nodes indicate bootstrap support >50% (1000 replications). A more detailed tree is shown in Supplementary Figure 12. The genomic context of the genes encoding the type-2 reaction center and the light-harvesting polypeptides in the metagenomic scaffolds and chromosomes of *Chloroflexus* and *Roseiflexus* isolates is also shown. The jagged lines indicate positions on scaffolds that are interrupted by a lack of overlapping sequence data between contigs.

(iii) Ca. *Chloracidobacterium spp.* Cluster-4 contained five scaffolds containing phylogenetic marker genes with best matches to Kingdom Acidobacteria (including a *recA* sequence labeled 'RecA Cabt' in Supplementary Figure 10). These scaffolds had distinct oligonucleotide frequency patterns as compared with the *Ca.* C. thermophilum genome, of which a detailed analysis will be published separately (A Garcia Costas, Z Liu, L Tomsho, SC Schuster, DM Ward and DA Bryant, submitted), despite the fact that 97% of the sequences from these scaffolds were recruited by this genome with a mean of 82.5% NT ID (Figure 2 and Supplementary Table 7). Genes involved in BChl and chlorosome biosynthesis were observed on these scaffolds, and a gene encoding a type-1 photosynthetic reaction center gene (*pscA*) was observed when the clustering stringency was lowered to 80%. Although the number of Cluster-4 scaffolds was small, these scaffolds were the largest produced by the Celera assembler (the average size was >300 000 bp, and the largest was 1.6 Mb; see Table 1). The *Ca.* C. thermophilum genome recruited 8.1% of all unassembled metagenome sequences, 90.8% of which were jointly recruited (Figure 2). The % NT ID distribution of these sequences suggested that, whereas there are native mat organisms nearly identical to the *Ca.* C. thermophilum isolate at some loci (Figure 2), most Cluster-4 sequences are derived from organisms that are more distantly related to *Ca.* C. thermophilum than any species belonging to the same genera that we investigated (Supplementary information Section 3). The high proportion of syntenous, jointly recruited metagenome sequences from the genome recruitment analysis was evidence for conservation of synteny within this population, which probably contributed in part to the longer-than-average assemblies.

(iv) Chlorobiales-Like Organisms. Cluster-5 scaffolds had oligonucleotide frequency signatures similar to that of the *Chloroherpeton thalassium* genome (Figure 1) and contained phylogenetic marker and functional genes (Table 3) that are typical of members of the Chlorobiales. The genome of *C. thalassium* recruited 8.4% of the metagenomic sequences across all temperature–spring combinations, most of which were from low-temperature samples and were disjointly recruited (Table 2 and Figure 2). Although they were not found on scaffolds >20 kb, many *recA* sequences were recruited that, like the *C. thalassium recA* sequence, form an out-group to the clade that contains the

well-characterized chlorophototrophs in the order Chlorobiales (Supplementary Figure 10). The 63.4% mean NT ID to *C. thalassium* homologs was approximately equal to the % NT ID of homologs belonging to different genera within a kingdom-level lineage (Figure 2 and Supplementary information Section 3). Hence, phylogenetic information alone did not provide high confidence that these sequences were derived from members of the Chlorobiales. Functional genes found on the scaffolds of this cluster clarified the potential physiological properties of this population. In particular, one scaffold contained a gene encoding a homolog of the Fenna–Matthews–Olson protein, which is a BChl-*a*-binding antenna protein involved in anoxygenic photosynthesis and only known to occur in the members of the Chlorobiales and chlorophototrophic Acidobacteria (Bryant *et al.*, 2007, 2011). High-performance liquid chromatographic analysis of pigments extracted from the Mushroom Spring mat will be published elsewhere (M Pagel and DA Bryant, unpublished data), but preliminary results indicated the presence of BChl-*d*, which replaces BChl-*c* as the aggregated pigments within the chlorosomes of 'green' chlorophototrophic Chlorobiales. Other Cluster-5 scaffolds contained homologs of the reaction center subunit gene *pscA* ('OS GSB PscA'; Bryant *et al.*, 2007), *pscB*, *pscD* as well as *csmC*, a gene encoding a chlorosome envelope protein that has no homologs in other chlorosome-containing chlorophototrophs and thus is currently diagnostic for Chlorobiales (Bryant *et al.*, 2011).

(v) A Novel Anaerolineae-Like Chlorophototroph. Cluster-6 scaffolds were not similar in oligonucleotide composition to any isolate genome, but contained phylogenetic marker genes associated with bacteria from Kingdom Chloroflexi (Figure 1). The RDP Bayesian Classifier assigned a full-length 16S rRNA sequence in this cluster to the taxonomic class Anaerolineae with 95% confidence, and this observation was supported by phylogenetic analysis (see Supplementary Figure 11). Furthermore, genes encoding ribosomal proteins and *recA* genes (Table 3) supported this kingdom-level phylogenetic assignment. In particular, a *recA* gene associated with assembly Cluster-6 ('RecA 6'; Supplementary Figure 10) is phylogenetically earlier diverging than the monophyletic clade containing known chlorophototrophic Chloroflexales (for example, *Roseiflexus* and *Chloroflexus* spp.). Several genes involved in anoxygenic chlorophototrophy were encoded on the same scaffold as the 16S rRNA gene in Cluster-6. This cluster also contained *bchXYZ* genes encoding the subunits of the light-independent chlorophyllide reductase, an enzyme required for the biosynthesis of BChl-*a* (Chew and Bryant, 2007), as well as other BChl biosynthesis genes (*bchD*, *bchF*, *bchH* and *bchI*) common to the BChl-*a* and BChl-*c* biosynthetic pathways. A separate scaffold in this cluster contained non-fused *pufL* and *pufM* sequences homologous to Chloroflexi sequences but in a unique genomic context (Figure 3). Phylogenetic analysis of the PufL and PufM sequences showed that, in comparison with those of known filamentous anoxygenic chlorophototrophs in the Chloroflexales, these sequences occupy novel and/or basal positions in a phylogenetic tree (Figure 3 and Supplementary Figure 12). When compared to their closest homologs in the *Chloroflexus* and *Roseiflexus* spp. genomes, these PufL and PufM sequences had amino-acid identities of 48% and 62%, respectively.

Assembly-independent BLASTN analysis showed that the metagenome sequences comprising Cluster-6 scaffolds had low % NT ID (60–66%) to the Chloroflexi genomes. Approximately 33% of the sequences comprising the Cluster-6 scaffolds were not recruited by any reference genome above the established cutoffs, and thus were 'null' bin sequences (see Supplementary Table 7).

(vi) Novel Putatively Chemoorganotrophic Populations. The scaffolds in Clusters 7 and 8 did not have oligonucleotide frequencies similar to any tested isolate genomes, and contained functional and phylogenetic marker genes (including 'RecA 7' in Supplementary Figure 10) with very distant relationships to sequences in currently available public databases. Most metagenomic sequences contained in these scaffolds were not recruited by a reference genome above the specified cutoff and were assigned to the 'null' bin, but some sequences were recruited at low % NT ID by multiple genomes (Supplementary Table 7). Clusters 7 and 8 did not contain any genes homologous to those specific for chlorophototrophy. Both clusters contained genes encoding caa_3-type cytochrome *c* oxidases, which suggested the potential for aerobic oxidative phosphorylation to exist in the organisms contributing these sequences. Cluster-7 additionally included scaffolds encoding glycolate oxidase (*glcD*) and acetyl-CoA synthetase (*acs*) genes (Table 3). Thus, the organisms contributing these sequences may have the potential for aerobic chemoorganotrophy using glycolate and/or acetate as an electron donor.

No assembly clusters corresponded to organisms related to *Thermomicrobium roseum*, *Thermus thermophilus* or *Thermodesulfovibrio yellowstonii*, but the genomes of these isolates recruited sequences above 75% NT ID (Table 2 and Figure 2). All other reference genomes recruited a low number of sequences with low % NT ID values (Supplementary Figure 13). Approximately 20% of metagenomic sequences could not be associated with any reference genome above an e-value cutoff of 10^{-10} using the specified parameters and were assigned to the 'null' bin.

Patterns of metagenomic diversity

(i) Multiple Populations in Recruitment Bins. Recruitment analysis of the metagenomic clones

from the 65 °C Mushroom Spring sample showed at least two populations, one with >92% NT ID and one with 83–92% NT ID, relative to the *Synechococcus* sp. strain-A genome (Figure 4). The more divergent sequences were likely contributed by A′-like *Synechococcus* spp., as they showed >98% NT ID with homologs in a metagenome produced by pyrosequencing from a 68 °C sample from Mushroom Spring, known to be dominated by these genotypes (Supplementary information Section 6 and Supplementary Figures 3 and 14; Ferris *et al*., 2003). These accounted for only 1.57% of the A-like sequences in all metagenomes (Table 2).

(ii) Synteny versus Relatedness. There was a positive relationship between the degree of genetic relatedness and the conservation of synteny in both metagenomic sequences and genomic reference sequences as compared with *Synechococcus* sp. strain-A (Figure 5). Metagenomic sequences originating from A-like organisms (that is, ⩾92% NT ID with the *Synechococcus* sp. strain-A genome) showed greater synteny with respect to the *Synechococcus* sp. strain-A genome than did sequences associated with A′-like organisms (that is, 83–92% NT ID with the *Synechococcus* sp. strain-A genome), which in turn showed higher synteny than did B′-like sequences (that is, comparing sequences that had ⩾90% NT ID to the *Synechococcus* sp. strain-B′ genome with homologs in the *Synechococcus* sp. strain-A genome). To assess synteny with more distantly related isolate genomes, we compared paired-end sequences of simulated metagenomic fragments (comprising sequence fragments from representative cyanobacterial isolate genomes fractionated to reflect the range of sizes and the abundances of our Sanger metagenome clone inserts) with the *Synechococcus* sp. strain-A genome (Supplementary information Section 3). Synteny between the *Synechococcus* sp. strain-A and B′ genomes was nearly identical to that observed empirically, but synteny between the *Synechococcus* sp. strain-A genome and the more distantly related genomes was almost undetectable (Figure 5).

Evidence for homologous recombination. Metagenomic clones, whose disjointly recruited ends can each be confidently associated with different reference genomes, provided evidence for possible past gene exchange between A-like *Synechococcus* spp. and members of the *Synechococcus* A′ and B′ lineages, as well as between these cyanobacteria and filamentous anoxygenic chlorophototrophs or *Ca.* C. thermophilum. The relative percentage of clones, whose end sequences could be confidently associated with *Synechoccoccus* sp. strain-A on one end and with other populations on the other end, decreased from 26% for all A′-like sequences (that is, 83–92% NT ID to *Synechococcus* sp. strain-A; no isolate genome is available from this organism type) to 4.5% for all *Synechococcus* sp. strain-B′-like sequences (that is, >90% NT ID to *Synechococcus* sp. strain-B′), to 1.1% for sequences associated with a more distantly related cyanobacterial reference genome (that is, *Thermosynechococcus elongatus* BP-1) and to 0.2% for sequences associated with yet more distantly related genomes (that is, *Roseiflexus* sp. strain RS1, *Chloroflexus* sp. strain 396-1 or *Ca.* C. thermophilum). Many of these disjointly recruited metagenome sequences encoded CRISPR-associated proteins putatively involved in adaptive responses to phage predation. Some recombination events among cyanobacteria and more distantly related organisms may thus be indicative of phage–host interactions (Supplementary Table 9; Heidelberg *et al*., 2009). Other disjointly recruited cyanobacterial sequences encoded transposases on the linked

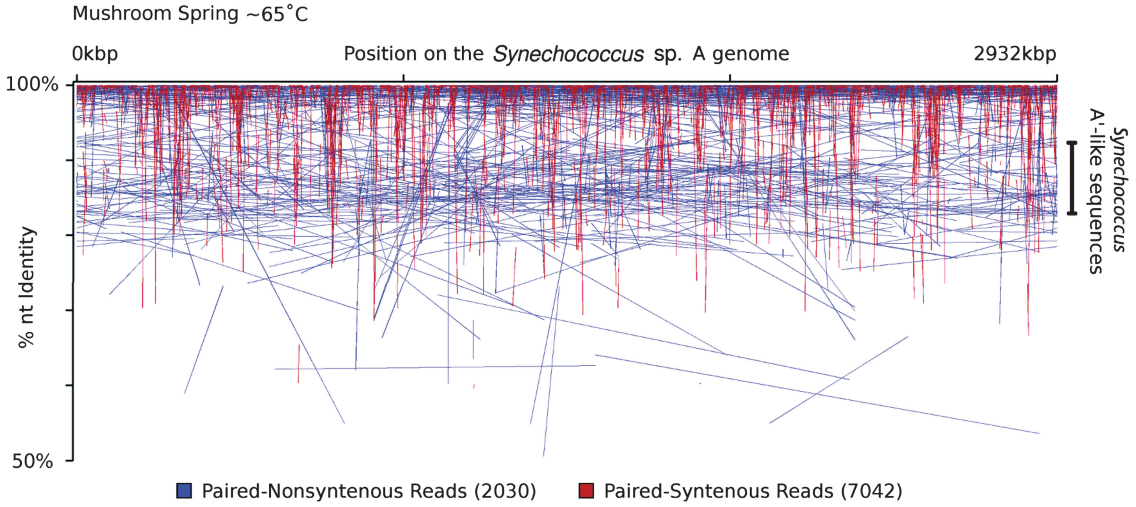

Figure 4 Position of alignments and the corresponding % NT ID to the *Synechococcus* sp. A genome of syntenous (red) and non-syntenous (blue) sequences jointly recruited by the *Synechococcus* sp. A genome from the Mushroom Spring ~65 °C metagenome. Each end sequence is connected by a line to its clone mate. Sequences suspected to originate from *Synechococcus* sp. A′-like populations ranging from 83to 92% NT ID are indicated on the right side of the graph. % NT ID, percent nucleotide identity.

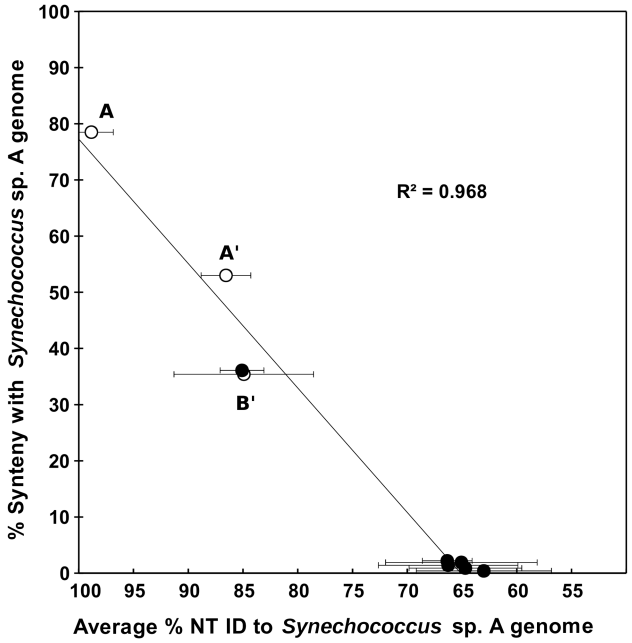

Figure 5 Synteny conservation between the *Synechococcus* sp. strain-A genome and metagenomic sequences and other genomes. The open circles represent alignments of metagenomic sequences relative to the *Synechococcus* sp. strain-A genome. Metagenome sequences were categorized as *Synechococcus* A, A′ or B′ based on % NT ID ranges to the *Synechococcus* spp. strain-A and B′ recruitment bins. The closed circles represent alignments of genome sequences from cultivated cyanobacteria (*T. elongatus*, *Gloeobacter violaceus*, *Synechococcus* sp. strain WH8102, *Nostoc* sp. strain PCC7120) and the out-group organism *Roseiflexus* sp. RS-1 relative to the *Synechococcus* sp. strain-A genome. These genome fragments were generated *in silico* to represent the same proportion of insert sizes as observed in the distribution of metagenome sequences that were recruited by the *Synechococcus* sp. strain-A genome. % NT ID, percent nucleotide identity.

paired-end sequences that were recruited to bacterial genomes other than from cyanobacteria. Such mobile genetic elements may even be transferred across distant lineages (Supplementary Table 9).

Discussion

This 167-Mb metagenome study of the green mat layer of Octopus and Mushroom Springs resulted in depth-of-coverage estimates between ~1.7X and ~5.7X for the eight dominant populations demarcated by scaffold clustering (Table 1). The complexity of this metagenome was relatively limited compared with the metagenome of a non-thermal, hyper-saline phototrophic, microbial mat from Guerrero Negro in Baja California Sur, Mexico (~105 Mb total metagenomic sequence; Kunin *et al.*, 2008), which did not produce assemblies greater than ~8400 bp in length. Metagenomic studies of less complex microbial communities have benefited from the assembly of metagenomic sequence data to identify and characterize the function of novel community members for which reference genomes of closely related organisms are not available (for example, Tyson *et al.*, 2004; Simmons *et al.*, 2008; Dick *et al.*, 2009; Denef *et al.*, 2010; Inskeep *et al.*, 2010). The structure of the Octopus and Mushroom Spring mat communities enabled us to use similar strategies to link community composition and potential function in these mats by resolving the phylogenetic and genomic context of individual functional genes, which led to the assignment of metabolic characteristics for microorganisms previously known only by the presence of 16S rRNA sequences.

Linkage between community composition and potential community function

The observation of assembly clusters with genes that indicated metabolic properties consistent with *Synechococcus* spp., *Roseiflexus* spp., *Chloroflexus* spp. and *Ca.* C. thermophilum was expected. However, the ability to associate functional potential with phylogeny also enabled us to link genes indicative of anoxygenic chlorophototrophy with a Chlorobiales-like population, and thus to confirm suspicions based on 16S rRNA sequence data that were not definitive and on a *pscA* sequence that previously could not be linked to phylogenetic markers. The ability to link functional and phylogenetic markers through assembly also enabled the discovery of three new predominant populations of organisms in this mat, which is remarkable because this system has been studied by numerous microbiologists over many decades.

One newly discovered population (Cluster-6), which has the functional potential for anoxygenic chlorophototrophy, is most closely related to cultured chemoorganotrophic bacteria isolated from thermal environments belonging to the classes Anaerolineae and Caldilineae within Kingdom Chloroflexi (Sekiguchi *et al.*, 2003; Hugenholtz and Stackebrandt, 2004; Yamada *et al.*, 2006, 2007). We detected the 16S rRNA sequences of these populations (Supplementary Figure 4 and Liu *et al.*, 2011) but were unable to infer from them a phototrophic phenotype, as these lineages of Kingdom Chloroflexi had not been known previously to contain phototrophic organisms. The novel population forms an out-group to the currently known filamentous anoxygenic chlorophototrophs within Order Chloroflexales and sequences of non-phototrophic Chloroflexi (Supplementary Figure 11). Before this discovery, chlorophototrophy in Chloroflexi was thought to be restricted to the Chloroflexales, which seemed to have evolved from a chemoorganotrophic common ancestor of this group and the non-phototrophic organisms in Order Herpetosiphonales. The discovery of chlorophototrophy in another deeply rooted branch of Kingdom Chloroflexi suggests that it is plausible that chlorophototrophy was an ancestral trait in Kingdom Chloroflexi that

was subsequently lost in some descendant lineages. Possible ancestral traits in Kingdom Chloroflexi can be inferred from properties shared between the newly discovered Anaerolineae-like chlorophototroph and members of Chloroflexales. All contain genes needed for BChl-*a* synthesis and type-2 photosynthetic reaction centers similar to those of *Proteobacteria*, but some members (for example, *Chloroflexus* spp.) also have chlorosomes, a trait shared with Chlorobiales and one member of the Acidobacteria (Bryant *et al.*, 2011). It is not yet known whether the newly discovered chlorophototroph has the capability of producing BChl-*c* and chlorosomes.

Genes indicating chlorophototrophic metabolism were not found on metagenomic scaffolds of two other newly discovered populations corresponding to Clusters 7 and 8, yet these scaffolds provide an estimated depth of coverage that is greater than that of *Chloroflexus* spp. represented by Cluster-3 in which nine phototrophy genes were observed. Genes for oxidation of reduced inorganic compounds were not observed, but these organisms apparently possess genes that encode enzymes involved in aerobic respiratory metabolism. One of these populations has the genes necessary for oxidation of glycolate and acetate, which are known to be produced and excreted by mat cyanobacteria and can be metabolized by other community members (Bateson and Ward, 1988; Nold and Ward, 1996; van der Meer *et al.*, 2005).

Pyrosequencing of cDNA from reverse-transcribed rRNA (Liu *et al.*, 2011) showed that most rRNAs (~88%) dominating the upper green layer of the mat are derived from the same eight phylogenetic groups identified in the metagenome. The linkage of rRNA sequences matching those in the scaffold clusters from the shotgun metagenomic data contributed to the assignment of functional roles for five of the eight predominant populations in the upper green layer of the Octopus and Mushroom Springs.

Description of functional guilds

Our analysis of the attributes of eight distinct assembly clusters (Table 3) provided evidence for the functions of major taxa, which we assigned to functional guilds according to their partitioning of environmental resources and conditions (Table 4). Cyanobacteria perform oxygenic photosynthesis using the visible light spectrum, but other chlorophototrophic groups have the potential to harvest near infrared light. For instance, *Roseiflexus* spp. have the genes to produce BChl-*a* harvesting 850- to 900-nm light. Two phylogenetic groups share the potential to produce BChl-*c*, whereas the Chlorobiales-like organisms likely produce BChl-*d*. The Chlorobiales-like population also contained genes essential for producing chlorosomes, which are also known to occur in *Chloroflexus* spp. and *Ca.* C. thermophilum isolates (Pierson and Castenholz, 1974; Bryant *et al.*, 2007). These observations suggest that these three populations harvest primarily 700- to 750-nm light, and finer niche differentiation within this group is precedented by the BChl-*d*-containing Chlorobiales. Such niche partitioning among chlorosome-containing chlorophototrophs in natural environments has been shown by the

Table 4 Relationship between predominant phylogenetic groups, functional potential and functional guilds

Phylogenetic group	*Chlorophylls*	*Carbon metabolism*	*Relative temperature distribution*	*Possible electron donor utilization*	*Possible electron acceptor utilization*	*Functional guild*[a]
A-like and B′-like *Synechococcus* spp.	Chl-*a*	Autotrophy	60 and 65 °C[b]	H_2O	O_2	600- to 700-nm oxygenic phototrophs
Roseiflexus-like FAPs	BChl-*a*	Mixotrophy[c]	60 and 65 °C	H_2	O_2	850- to 950-nm mixotrophs
Chloroflexus-like FAPs	Major: BChl-*c*; Minor: BChl-*a*	Mixotrophy	60<65 °C	H_2, HS^-, $S_2O_3^{2-}$, S^0	O_2	700- to 750-nm mixotrophs
C. thermophilum-like spp.	Major: BChl-*c*; Minor: BChl-*a*, Chl-*a*	Heterotrophy	60>65 °C	ND	O_2	700- to 750-nm heterotrophs
Chlorobiales-like spp.	Major: BChl-*d*; Minor: BChl-*a*, Chl-*a*	Mixotrophy	60>65 °C	ND	O_2	700- to 750-nm mixotrophs
Anaerolineae-like spp.	BChl-*a*[d]	Unknown	60 and 65 °C	HS^-, $S_2O_3^{2-}$, S^0	O_2	BChl-*a*/N-IR mixotrophs
Cluster-7 spp.	ND	Heterotrophy	60 and 65 °C	Glycolate, acetate	O_2	Aerobic chemoorganoheterotrophs
Cluster-8 spp.	ND	Heterotrophy	60 < 65 °C	????	O_2	Aerobic chemoorganoheterotrophs

[a]Ranges in which the absorption maxima of the light-harvesting systems of these guilds are maximal in the red to near-IR of the electromagnetic spectrum.
[b]B′-like sequences were much more predominant at 60 °C than at 65 °C. A-like sequences were observed at 60 °C and were predominant at 65 °C.
[c]Mixotrophy refers to both heterotrophic and autotrophic growth, perhaps simultaneously (Bryant *et al.*, 2011).
[d]Insufficient evidence currently exists to determine whether this organism can synthesize other chlorophylls, and to know its principal absorption range in the near-IR.

vertical stratification of BChl-*c*-, BChl-*d*- and BChl-*e* containing organisms in lakes (see discussion in reference Maresca *et al.*, 2004).

Further niche partitioning undoubtedly explains the coexistence of different types of phototrophs using similar light wavelengths. One possibility is that the different members of a functional guild differ in terms of carbon metabolism. For instance, among phototrophs using 700- to 750-nm light, native *Chloroflexus* spp. have the genetic potential for carbon fixation through the 3-hydroxypropionate pathway (Klatt *et al.*, 2007; Bryant *et al.*, 2011), but most *Chloroflexus aurantiacus* strains achieve higher growth rates in culture with photo-heterotrophic metabolism (Madigan *et al.*, 1974; Pierson and Castenholz, 1974) and may conduct mixotrophic rather than autotrophic carbon metabolism *in situ* (Bryant *et al.*, 2011). However, *Ca.* C. thermophilum and the Chlorobiales population do not appear capable of autotrophic metabolism and are more likely heterotrophic.

Another possible explanation for niche differentiation among these phototrophs is temperature adaptation. *Chloroflexus* spp. sequences were relatively more abundant in the 65 °C metagenome, whereas the *Ca.* C. thermophilum and Chlorobiales-like organisms were relatively more abundant in the 60 °C metagenome (Table 2 and Supplementary Figure 6). At this time, *Ca.* C. thermophilum spp. and Chlorobiales-like organisms cannot be placed into separate functional guilds on the basis of carbon metabolism or temperature preference. Differences in electron donor utilization could also be involved in niche partitioning, but deeper metagenomic sequencing, coupled with genetic and physiological studies, will be required to test this hypothesis. Differences in the timing of gene expression provide additional clues to explain the coexistence of populations, which cannot be separated based on putative physiological differences inferred from gene content (Liu *et al.*, 2011).

Differences between metagenomes and isolate genomes

The taxonomic resolution of the phylogenetic groups defined by scaffold clustering in this study is approximately at the level of named genera. However, population genetics studies of uncultivated *Synechococcus* spp. from Octopus and Mushroom Spring have indicated the presence of numerous, genetically distinct ecotypes within the A-like and B′-like lineages that occupy discrete positions along environmental gradients (for example, light and temperature (Melendrez *et al.*, 2011; ED Becraft, FM Cohan, M Kühl, S Jensen and DM Ward, unpublished) and show different metabolic regulation over the diel cycle (Liu *et al.*, 2011). Consistent with these findings, the genetic and functional differences in metagenomic *Synechococcus* spp. populations in comparison with the two cyanobacterial isolate genomes showed ecological heterogeneity within closely related phylogenetic groups. The discovery of ferrous iron transporter homologs in *Synechococcus* sp. A-like populations (this study), and in B′-like populations (Bhaya *et al.*, 2007), as well as the presence of these genes in the *Roseiflexus* sp. strain RS1 genome (van der Meer *et al.*, 2010), suggests that the ability to use Fe^{2+} might be a common adaptation among the mat community members. The presence of genes for an alternative pathway for urea metabolism in the metagenomic A-like *Synechococcus* provides additional evidence that urea may be an important nitrogen-containing nutrient in these mats (Bhaya *et al.*, 2007).

Overall, there were few examples of functional genes present in native populations but absent in the genomes of sequenced isolates; however, it is clear that ecological diversification also occurs through mechanisms other than differences in gene content. For example, adaptations to temperature (Miller and Castenholz, 2000; Allewalt *et al.*, 2006) may be based on adaptive nucleotide substitutions (Miller, 2003; Ward *et al.*, 2011a). The metagenomic diversity with respect to the *Roseiflexus* sp. RS1 genome likely encompasses multiple ecologically distinct *Roseiflexus* spp., such as those showing different distributions along the flow path in these mats (Ferris and Ward, 1997; Nübel *et al.*, 2002; Ward *et al.*, 2006).

Insights into genome evolution

Comparisons of metagenomic sequences and genomes of representative mat isolates also yielded insights into genome diversity among closely related populations. Cyanobacterial genomes are less syntenous with each other at a given degree of sequence divergence compared with other taxonomic groups (Rocha, 2006; Frangeul *et al.*, 2008). The number of translocations and transpositions evident when comparing the genomes of *Synechococcus* spp. strains A and B′ has caused nearly a complete lack of synteny between them (Bhaya *et al.*, 2007), yet it is apparent that *Synechococcus* spp. more closely related to either *Synechococcus* sp. strain-A or B′ are more syntenous to their respective closest relative. Both synteny and the number of disjointly recruited metagenomic clones, which might document past recombination events, decrease as the genetic relatedness between two organisms decreases. The latter trend is consistent with empirical findings in *Escherichia*, *Bacillus* and *Streptococcus* spp., which showed that recombination rates declined as the genetic distances between organisms increased (Roberts and Cohan, 1993; Vulić *et al.*, 1997; Majewski *et al.*, 2000). Our results suggested that homologous recombination between populations as divergent as *Synechococcus* spp. strains A and B′ has generally been uncommon (~5% of the total number of sequences recruited by either *Synechococcus* sp. strain-A or B′). Comparative genomic studies have

shown that, although gene transfer among cyanobacteria is evident (Zhaxybayeva *et al.*, 2006), these events have been infrequent to the degree that they do not obscure inferences about the phylogenetic relationships in this kingdom (Kettler *et al.*, 2007; Swingley *et al.*, 2008; Zhaxybayeva *et al.*, 2009; Melendrez *et al.*, 2011; Popa *et al.*, 2011).

Conclusion

This metagenomic study showed that the chlorophototrophic communities inhabiting the effluent channels of Octopus and Mushroom Springs were more phylogenetically and physiologically diverse than was known on the basis of light microscopy, traditional cultivation methods and previous 16S rRNA surveys. The combination of depth of coverage and limited diversity enabled metagenomic assemblies leading to (i) the confirmation of a novel chlorophototrophic member of Chlorobiales in these mats and (ii) the discovery of several novel populations, including a chlorophototroph in a novel lineage of Chloroflexi and two types of putatively chemoorganotrophic community members more representative of the native populations than the currently cultivated chemoorganotrophic isolates. This effectively doubled the number of predominant populations known to inhabit the mat. Deeper-coverage metagenomes are in production that will further enhance our understanding of the physiological potential of the dominant members of this microbial mat community. The availability of genomes of isolates closely related to native populations enabled (i) the discovery of functions not represented by the isolates and (ii) the observation that breakdown of synteny and exchange of genetic information are functions of how much populations have diverged. Finally, the results of these analyses provide the foundation for interpreting the metatranscriptome of the Mushroom Spring mat over a portion of the diel cycle in an accompanying study (Liu *et al.*, 2011).

Acknowledgements

This research was supported by the National Science Foundation Frontiers in Integrative Biology Research Program (EF-0328698) and the IGERT Program in Geobiological Systems (DGE 0654336), the National Aeronautics and Space Administration Exobiology Program (NAG5-8824, NAG5-8807 and NX09AM87G) and the US Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Genomic Science Program 395 (GSP) (This contribution originates from the GSP Foundational Scientific Focus Area (FSFA) at the Pacific Northwest National Laboratory (PNNL); contract no. 112443). We appreciate the support and assistance of the National Park Service personnel at Yellowstone National Park. We thank Marcus B Jones at the J Craig Venter Institute for help in using the PCT Barocycler for cell lysis. DAB additionally and gratefully acknowledges the support from the National Science Foundation (MCB-0523100), the Department of Energy (DE-FG02-94ER20137), and the Joint Genome Institute for support in obtaining the genomic sequences mentioned here.

References

Allewalt JP, Bateson MM, Revsbech NP, Slack K, Ward DM. (2006). Effect of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in the Octopus Spring microbial mat community of Yellowstone National Park. *Appl Environ Microbiol* **72**: 544–550.

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

Bateson MM, Ward DM. (1988). Photoexcretion and fate of glycolate in a hot spring cyanobacterial mat. *Appl Environ Microbiol* **54**: 1738–1743.

Bauld J, Brock TD. (1973). Ecological studies of *Chloroflexis*, a gliding photosynthetic bacterium. *Arch Microbiol* **92**: 267–284.

Bhaya D, Grossman AR, Steunou A-S, 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.

Brock TD. (1978). *Thermophilic Microorganisms and Life at High Temperatures*. Springer Verlag: New York.

Bryant DA, Frigaard N-U. (2006). Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* **14**: 488–496.

Bryant DA, Garcia Costas AM, Maresca JA, Chew AGM, Klatt CG, Bateson MM *et al.* (2007). *Candidatus Chloracidobacterium thermophilum*: an aerobic phototrophic acidobacterium. *Science* **317**: 523–526.

Bryant DA, Klatt CG, Frigaard N-U, Liu Z, Li T, Zhao F *et al.* (2011). Comparative and functional genomics of anoxygenic green bacteria from the taxa *Chlorobi*, *Chloroflexi*, and *Acidobacteria*. In: Burnap RL, Vermaas W (eds). *Advances in Photosynthesis and Respiration, Vol. 33, Functional Genomics and Evolution of Photosynthetic Systems*. Springer: Dordrecht, The Netherlands (in press).

Cheng G, Shapir N, Sadowsky MJ, Wackett LP. (2005). Allophanate hydrolase, not urease, functions in bacterial cyanuric acid metabolism. *Appl Environ Microbiol* **71**: 4437–4445.

Chew AGM, Bryant DA. (2007). Chlorophyll biosynthesis in bacteria: the origins of structural and functional diversity. *Annu Rev Microbiol* **61**: 113–129.

Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC *et al.* (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci USA* **107**: 2383–2390.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Ferris MJ, Kühl M, Wieland A, Ward DM. (2003). Cyanobacterial ecotypes in different optical microenvironments of a 68 °C hot spring mat community

revealed by 16S-23S rRNA internal transcribed spacer region variation. *Appl Environ Microbiol* **69**: 2893–2898.

Ferris MJ, Ward DM. (1997). Seasonal distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **63**: 1375–1381.

Frangeul L, Quillardet P, Castets A, Humbert J, Matthijs H, Cortez D *et al.* (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9**: 274.

Frigaard N-U, Bryant DA. (2006). Chlorosomes: antenna organelles in green photosynthetic bacteria. In: Shively, JM (ed). *Complex Intracellular Structures in Prokaryotes, Microbiology Monographs.*, Springer: Berlin, Germany, pp 79–114.

Hanada S, Takaichi S, Matsuura K, Nakamura K. (2002). *Roseiflexus castenholzii* gen. nov., sp. nov., a thermophilic, filamentous, photosynthetic bacterium that lacks chlorosomes. *Int J Syst Evol Microbiol* **52**: 187–193.

Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **4**: e4169.

Holo H, Sirevåg R. (1986). Autotrophic growth and CO_2 fixation of *Chloroflexus aurantiacus*. *Arch Microbiol* **145**: 173–180.

Hugenholtz P, Stackebrandt E. (2004). Reclassification of *Sphaerobacter thermophilus* from the subclass *Sphaerobacteridae* in the phylum *Actinobacteria* to the class *Thermomicrobia* (emended description) in the phylum *Chloroflexi* (emended description). *Int J Syst Evol Microbiol* **54**: 2049–2051.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA *et al.* (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5**: e9773.

Jensen SI, Steunou A-S, Bhaya, D, Kühl M, Grossman AR. (2011). *In situ* dynamics of O_2, pH and cyanobacterial transcripts associated with CCM, photosynthesis and detoxification of ROS. *ISME J* **5**: 317–328.

Kanamori T, Kanou N, Atomi H, Imanaka T. (2004). Enzymatic characterization of a prokaryotic urea carboxylase. *J Bacteriol* **186**: 2532–2539.

Kanungo T, Mount D, Netanyahu N, Piatko C, Silverman R, Wu A. (2002). An efficient *k*-means clustering algorithm: analysis and implementation. *IEEE Transact Patt Anal Mach Intell* **24**: 881–892.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**: e231.

Klatt CG, Bryant DA, Ward DM. (2007). Comparative genomics provides evidence for the 3-hydroxypropionate autotrophic pathway in filamentous anoxygenic phototrophic bacteria and in hot spring microbial mats. *Environ Microbiol* **9**: 2067–2078.

Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N *et al.* (2008). Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4**: 198.

Liu Z, Klatt CG, Wood JM, Rusch DB, Ludwig M, Wittekindt N *et al.* (2011). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J* **5**: 1279–1290.

Madigan MT, Petersen SR, Brock TD. (1974). Nutritional studies on *Chloroflexus*, a filamentous photosynthetic, gliding bacterium. *Arch Microbiol* **100**: 97–103.

Madigan MT, Jung DO, Karr EA, Sattley WM, Achenbach LA, van der Meer MTJ. (2005). Diversity of anoxygenic phototrophs in contrasting extreme environments. In: Inskeep WP, McDermott TR (eds). *Geothermal Biology and Geochemistry in Yellowstone National Park*. Montana State University Publications: Bozeman, pp 203–219.

Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG. (2000). Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**: 1016–1023.

Maresca JA, Chew AGM, Ponsatí MR, Frigaard, N-U, Ormerod JG *et al.* (2004). The *bchU* gene of *Chlorobium tepidum* encodes the C-20 methyltransferase in bacteriochlorophyll c biosynthesis. *J Bacteriol* **186**: 2558–2566.

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 Microbiol* **77**: 1359–1367.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A *et al.* (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**: 2818–2824.

Miller SR. (2003). Evidence for the adaptive evolution of the carbon fixation gene *rbcL* during diversification in temperature tolerance of a clade of hot spring cyanobacteria. *Mol Ecol* **12**: 1237–1246.

Miller SR, Castenholz RW. (2000). Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechecoccus*. *Appl Environ Microbiol* **66**: 4222–4229.

Nold SC, Ward DM. (1996). Photosynthate partitioning and fermentation in hot spring microbial mat communities. *Appl Environ Microbiol* **62**: 4598–4607.

Nübel U, Bateson MM, Vandieken V, Wieland A, Kühl M, Ward DM. (2002). Microscopic examination of distribution and phenotypic properties of phylogenetically diverse *Chloroflexaceae*-related bacteria in hot spring microbial mats. *Appl Environ Microbiol* **68**: 4593–4603.

Pierson BK, Castenholz RW. (1974). A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen. and sp. nov.. *Arch Microbiol* **100**: 5–24.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**: 599–609.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**: 145–158.

Roberts MS, Cohan FM. (1993). The effect of DNA sequence divergence on sexual isolation in *Bacillus*. *Genetics* **134**: 401–408.

Rocha EPC. (2006). Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* **23**: 513–522.

Sekiguchi Y, Yamada T, Hanada S, Ohashi A, Harada H, Kamagata Y. (2003). *Anaerolinea thermophila* gen. nov., sp. nov. and *Caldilinea aerophila* gen. nov., sp. nov., novel filamentous thermophiles that represent a previously uncultured lineage of the domain Bacteria at the subphylum level. *Int J Syst Evol Microbiol* **53**: 1843–1851.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Simmons SL, DiBartolo G, Denef VJ, Goltsman DSA, Thelen MP, Banfield JF. (2008). Population genomic analysis of strain variation in *Leptospirillum* Group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6**: e177.

Steunou A-S, Bhaya D, Bateson MM, Melendrez MC, Ward DM, Brecht E *et al.* (2006). *In situ* analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. *Proc Natl Acad Sci USA* **103**: 2398–2403.

Steunou A-S, Jensen SI, Brecht E, Becraft ED, Bateson MM, Kilian O *et al.* (2008). Regulation of *nif* gene expression and the energetics of N_2 fixation over the diel cycle in a hot spring microbial mat. *ISME J* **2**: 364–378.

Strauss G, Fuchs G. (1993). Enzymes of a novel autotrophic CO_2 fixation pathway in the phototrophic bacterium *Chloroflexus aurantiacus*, the 3-hydroxypropionate cycle. *Eur J Biochem* **215**: 633–643.

Swingley WD, Blankenship RE, Raymond J. (2008). Integrating Markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol Biol Evol* **25**: 643–654.

Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, Thiagarajan M *et al.* (2010). The JCVI standard operating procedure for prokaryotic metagenomics shotgun sequencing data processing. *Stand Genomic Sci* **2**: 2.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**: 938–947.

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 der Meer MTJ, Klatt CG, Wood J, Bryant DA, Bateson MM, Lammerts L *et al.* (2010). Cultivation and genomic, nutritional, and lipid biomarker characterization of *Roseiflexus* strains closely related to predominant *in situ* populations inhabiting Yellowstone hot spring microbial mats. *J Bacteriol* **192**: 3033–3042.

van der Meer MTJ, Schouten S, Bateson MM, Nübel U, Wieland A, Kühl M *et al.* (2005). Diel variations in carbon metabolism by green nonsulfur-like bacteria in alkaline siliceous hot spring microbial mats from Yellowstone National Park. *Appl Environ Microbiol* **71**: 3978–3986.

Vulić M, Dionisio F, Taddei F, Radman M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* **94**: 9763–9767.

Wahlund TM, Woese CR, Castenholz RW, Madigan MT. (1991). A thermophilic green sulfur bacterium from New Zealand hot springs, *Chlorobium tepidum* sp. nov.. *Arch Microbiol* **156**: 81–90.

Ward DM, Bateson MM, Ferris MJ, Kühl M, Wieland A, Koeppel A *et al.* (2006). Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units linking microbial community composition, structure and function. *Philos Trans R Soc Lond B Biol Sci* **361**: 1997–2008.

Ward DM, Bauld J, Castenholz RW, Pierson BK. (1992). Modern phototrophic microbial mats: anoxygenic, intermittantly oxygenic/anoxygenic, thermal, eukaryotic, and terrestrial. In: Schopf JW, Klein C (eds). *The Proterozoic Biosphere: a Multidisciplinary Study*. Cambridge University Press: Cambridge, pp 309–324.

Ward DM, Castenholz RW, Miller SR. (2011a). Cyanobacteria in geothermal habitats. In: Whitton BA (ed). *Ecology of Cyanobacteria*, 2nd edn. Springer: Dordrecht, the Netherlands (in press).

Ward DM, Klatt CG, Wood J, Cohan FM, Bryant DA. (2011b). Functional genomics in an ecological and evolutionary context: maximizing the value of genomes in systems biology. In: Burnap RL, Vermaas W (eds). *Advances in Photosynthesis and Respiration, Vol. 33, Functional Genomics and Evolution of Photosynthetic Systems*. Springer: Dordrecht, the Netherlands (in press).

Ward DM, Papke RT, Nübel U, McKitrick MC. (2002). Natural history of microorganisms inhabiting hot spring microbial mat communities: clues to the origin of microbial diversity and implications for microbiology and macrobiology. In: Staley JT, Reysenbach A-L (eds). *Biodiversity of Microbial Life: Foundations of Earth's Biosphere*. John Wiley and Sons: New York, pp 27–48.

Ward DM, Tayne TA, Anderson KL, Bateson MM. (1987). Community structure and interactions among community members in hot spring cyanobacterial mats. *Symp Soc Gen Microbiol* **41**: 179–210.

Ward DM, Weller R, Bateson MM. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**: 63–65.

Watanabe Y, Feick RG, Shiozawa JA. (1995). Cloning and sequencing of the genes encoding the light-harvesting B806-866 polypeptides and initial studies on the transcriptional organization of *puf2B*, *puf2A* and *puf2C* in *Chloroflexus aurantiacus*. *Arch Microbiol* **163**: 124–130.

Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B *et al.* (2008). Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* **2**: 853–864.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Wu M, Eisen J. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Yamada M, Zhang H, Hanada S, Nagashima KVP, Shimada K, Matsuura K. (2005). Structural and spectroscopic properties of a reaction center complex from the chlorosome-lacking filamentous anoxygenic phototrophic bacterium *Roseiflexus castenholzii*. *J Bacteriol* **187**: 1702–1709.

Yamada T, Sekiguchi Y, Hanada S, Imachi H, Ohashi A, Harada H *et al.* (2006). *Anaerolinea thermolimosa* sp.

nov., *Levilinea saccharolytica* gen. nov., sp. nov. and *Leptolinea tardivitalis* gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes *Anaerolineae* classis nov. and *Caldilineae* classis nov. in the bacterial phylum *Chloroflexi*. *Int J Syst Evol Microbiol* **56**: 1331–1340.

Yamada T, Imachi H, Ohashi A, Harada H, Hanada S, Kamagata Y *et al*. (2007). *Bellilinea caldifistulae* gen. nov., sp. nov. and *Longilinea arvoryzae* gen. nov., sp. nov., strictly anaerobic, filamentous bacteria of the phylum *Chloroflexi* isolated from methanogenic propionate-degrading consortia. *Int J Syst Evol Microbiol* **57**: 2299–2306.

Youvan DC, Bylina EJ, Alberti M, Begusch H, Hearst JE. (1984). Nucleotide and deduced polypeptide sequences of the photosynthetic reaction-center, B870 antenna, and flanking polypeptides from *R. capsulata*. *Cell* **37**: 949–957.

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. (2009). Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Biol Evol* **2009**: 325–339.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**: 1099–1108.

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