The ISME Journal (2007) 1, 703–713

www.nature.com/ismej

ORIGINAL ARTICLE

Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses

Devaki Bhaya[1], Arthur R Grossman[1], Anne-Soisig Steunou[1], Natalia Khuri[1,2], Frederick M Cohan[3], Natsuko Hamamura[4], Melanie C Melendrez[4], Mary M Bateson[4], David M Ward[4] and John F Heidelberg[5,6]

[1]*Department of Plant Biology, Carnegie Institution, Stanford, CA, USA;* [2]*Department of Computer Sciences, San Jose State University, San Jose, CA, USA;* [3]*Department of Biology, Wesleyan University, Middletown, CT, USA;* [4]*Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT, USA;* [5]*Department of Biological Sciences, Philip K Wrigley Marine Science Center, University of Southern California, Avalon, CA, USA and* [6]*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD, USA*

In microbial mat communities of Yellowstone hot springs, ribosomal RNA (rRNA) sequence diversity patterns indicate the presence of closely related bacterial populations along environmental gradients of temperature and light. To identify the functional bases for adaptation, we sequenced the genomes of two cyanobacterial (*Synechococcus* OS-A and OS-B′) isolates representing ecologically distinct populations that dominate at different temperatures and are major primary producers in the mat. There was a marked lack of conserved large-scale gene order between the two *Synechococcus* genomes, indicative of extensive genomic rearrangements. Comparative genomic analyses showed that the isolates shared a large fraction of their gene content at high identity, yet, differences in phosphate and nitrogen utilization pathways indicated that they have adapted differentially to nutrient fluxes, possibly by the acquisition of genes by lateral gene transfer or their loss in certain populations. Comparisons of the *Synechococcus* genomes to metagenomic sequences derived from mats where these *Synechococcus* stains were originally isolated, revealed new facets of microbial diversity. First, *Synechococcus* populations at the lower temperature regions of the mat showed greater sequence diversity than those at high temperatures, consistent with a greater number of ecologically distinct populations at the lower temperature. Second, we found evidence of a specialized population that is apparently very closely related to *Synechococcus* OS-B′, but contains genes that function in the uptake of reduced ferrous iron. *In situ* expression studies demonstrated that these genes are differentially expressed over the diel cycle, with highest expression when the mats are anoxic and iron may be in the reduced state. Genomic information from these mat-specific isolates and metagenomic information can be coupled to detect naturally occurring populations that are associated with different functionalities, not always represented by isolates, but which may nevertheless be important for niche partitioning and the establishment of microbial community structure.

The ISME Journal (2007) **1,** 703–713; doi:10.1038/ismej.2007.46; published online 25 October 2007
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: cyanobacteria; synteny; ecotype; iron; microbial mat; *Synechococcus*

Introduction

Microbial mats have played a crucial role in the evolution of life on earth, and modern mats are considered to be analogs of the extensive Precambrian microbial communities, which significantly altered the atmosphere and the oxidation/reduction balance of the earth (Hoehler *et al.*, 2001; Teske and Stahl, 2002). Microbial mats thrive in many extreme

Correspondence: D Bhaya, Department of Plant Biology, Carnegie Institution, 260 Panama Street, Stanford, CA 94305, USA.
E-mail: dbhaya@stanford.edu
Received 5 February 2007; revised 7 May 2007; accepted 14 May 2007; published online 25 October 2007

environments such as hypersaline bays, permanent-ice lakes and hot springs, where competition from other organisms is limited (Paerl *et al*., 2000). The mats represent model systems for discovering the principles of microbial community ecology and have been studied intensively for decades (Ward *et al*., 1998, 2006; Stal, 2000).

The dense microbial communities growing in the hot spring effluent channels of Yellowstone National Park are predominantly prokaryotic in nature, affording a unique opportunity to study interactions among different prokaryotes. The mats are relatively stable, simple communities, typical of other mat communities in the sense that organic matter is produced in the uppermost layers by phototrophs and subsequently recycled and used in the deeper layers by aerobic and anaerobic decomposition metabolisms (Ward *et al*., 1998; Ward and Castenholz, 2000). The mats contain a diversity of microorganisms ranging from phototrophs such as cyanobacteria (predominantly *Synechococcus* spp.) and green non-sulfur-like bacteria (GNSLB), such as *Roseiflexus* spp. and *Chloroflexus* spp., to less well-characterized heterotrophic anaerobic and aerobic bacteria (Brock, 1978; Ward *et al*., 1998; Ward and Castenholz, 2000). The metabolic activities of and interactions among, the mat-building organisms create impressive stratified structures in which microorganisms may experience steep gradients of light, oxygen and nutrients (Ward *et al*., 2006).

Early work based primarily on microscopy and cultivation approaches led to the belief that the mat was primarily comprised of a single morphologically recognizable *Synechococcus* species and that the physical cohesiveness of the mat was provided by the GNSLB, *Chloroflexus aurantiacus*. With the advent of molecular techniques such as direct sequencing of 16S rRNA from mat clones and denaturing gradient gel electrophoresis (DGGE), it became obvious that the mat contained many unique 16S rRNA sequences, most of which were not identical to the 16S rRNA sequences of isolates cultivated from the mat. Cyanobacterial and GNSLB 16S rRNA sequences detected by molecular analyses were found to be distantly related to those of readily cultivated *Synechococcus* and *Chloroflexus* isolates and to occur as sets of closely related sequences differing by $<5\%$ in 16S rRNA sequence (Ward *et al*., 1990; Ferris and Ward, 1997; Nübel *et al*., 2002). DGGE surveys also revealed that the 16S rRNA gene distributions of the closely related cyanobacterial and GNSLB 16S rRNA populations (we use the term populations to mean sets of related individuals and use adjectives to specify how the individuals within populations are related (for example, cyanobacterial 16S rRNA populations, *Synechococcus* populations, functionally specialized populations, and so on)) varied along well-defined temperature gradients that exist in the effluent channels. For instance, closely related *Synechococcus* 16S rRNA sequences designated A″, A′, A, B′ and B, were detected at progressively lower temperatures from about 70°C to about 50°C, respectively (Ferris and Ward, 1997; Ward *et al*., 1998). Furthermore, recent experiments with *Synechococcus* isolates representing some of these 16S rRNA sequences suggest that they are physiologically adapted to temperature ranges and light intensities experienced *in situ (*Allewalt *et al*., 2006; Kilian *et al*., 2007). There is also evidence of vertical stratification of *Synechococcus* populations within the microbial mats. On the basis of fluorescence characteristics and DGGE analysis, physiologically and genetically distinct *Synechococcus* populations were identified at different depth intervals within the 1-mm thick upper green layer of these mats (Ramsing *et al*., 2000). Many parameters such as light intensity and quality, UV penetration, oxygen tension, and nutrient fluxes vary as a function of mat depth and over a diel cycle (Ward *et al*., 2006), although it is not obvious which of these parameters (one or more) lead to the stratification of *Synechococcus* populations within the mat. This correspondence between sequence variation and microgeographic distribution matches the prediction of some evolutionary ecology models of bacterial speciation, where natural selection acts upon variation to yield distinct phylogenetic clusters of individuals, with each cluster representing an ecologically distinct population (or 'ecotype', defined as a clade of individuals related by common ecological characteristics) (Ward, 1998; Ward and Cohan, 2005; Cohan and Perry, 2007). These models provide a framework for using sequence data to discover ecologically distinct populations, which can be considered the fundamental species-like units of which communities are comprised and whose ecological distinctness can be elucidated by genomic and physiological analyses (Cohan, 2006; Ward, 2006; Cohan and Perry, 2007; Ward *et al*., 2007).

Although a combination of genetic, biochemical and physiological data collected over decades has led to the hypothesis that there may be multiple specialized populations within these stratified mats, the underlying functional and genetic basis for this diversity has not yet been fully appreciated (Ward *et al*., 1998, 2006). Because the mat *Synechococcus* populations have been extensively characterized at the level of rRNA diversity and physiology, and are available as isolates, they provide optimal starting material for high-resolution molecular analyses to establish links between physiological differences, niche partitioning strategies and microbial diversity within the mat community (Steunou *et al*., 2006; Kilian *et al*., 2007). Here, we present our approach in which we sequenced the genomes of two isolates of *Synechococcus* derived from different temperature regions of the mat. This analysis has provided us with 'anchor genomes' from two differently temperature-adapted *Synechococcus* isolates whose 16S rRNA sequences correspond to dominant mat

sequences, which can be directly compared to each other, and also compared to metagenomic sequences derived from regions of the mat from which these isolates were originally isolated.

Materials and methods

Source of DNA

Synechococcus JA-2-3Aa {genotype designation A-NACy05a} (hereafter *Synechococcus* OS-A) was isolated by filter cultivation from samples derived from a region of the mat in Octopus Spring exposed to temperatures ranging from 58°C to 65°C; *Synechococcus JA-2-3B′a(2-13)* {genotype designation B′NACy10o} (hereafter *Synechococcus* OS-B′) was isolated from Octopus Spring mat samples collected at temperatures ranging from 51 to 61°C (Allewalt *et al.*, 2006). We also generated metagenomic sequences from recombinant libraries from total DNA isolated from the top green layer (upper ~1 mm) of the microbial mats in Octopus Spring and from Mushroom Spring, a nearby spring with similar physicochemical characteristics (Papke *et al.*, 2003). The samples used to generate the metagenomic sequences were collected from two different temperature-defined sites averaging ~60 and ~65°C (Materials and methods and Supplementary Table S5).

Sequencing and annotation of Synechococcus *OS-A and OS-B′ genomes*

Plasmid libraries with small (2–3 kbp) and large (10–12 kbp) inserts were constructed in pUC-derived vectors following random mechanical shearing (nebulization) of genomic DNA. The plasmid sequences were assembled using the Celera Assembler. The coverage criteria were (a) every position required at least double-clone coverage (or sequence from a PCR product amplified from genomic DNA) and (b) sequences were generated from both strands. The sequence was edited manually, and additional PCR and sequencing reactions were performed to close gaps, improve coverage and resolve sequence ambiguities. Sequences of regions of the DNA that were repeated on the genome were verified by PCR amplification across the repeat followed by sequence analysis of the amplification product. The final assemblies of the *Synechococcus* OS-A and OS-B′ genomes were based on 51 866 and 48 038 sequences, respectively.

Genome analysis

An initial set of open reading frames likely to encode proteins (coding sequences (CDSs)) were predicted as described previously (Heidelberg *et al.*, 2004). All predicted proteins larger than 30 amino acids were searched against a non-redundant protein database (Heidelberg *et al.*, 2004). Frameshifts and point mutations were detected and corrected where appropriate. The remaining frameshifts and point mutations were considered to be authentic. Protein membrane-spanning domains were identified by TopPred (Claros and von Heijne, 1994). Each CDS was inspected to define the initiation codon, the position of the ribosomal binding site and the transcription termination site. Two sets of hidden Markov models were used to determine CDS membership in families and superfamilies: Pfam (Bateman *et al.*, 2004) and TIGRFAMs (Haft *et al.*, 2003). Pfam hidden Markov models were also used with a constraint of a minimum of two hits to find repeated domains within proteins and to mask them. Phylogenomic analysis was used to assist with functional predictions, and comparative genome analyses were performed using the Comprehensive Microbial Resource (Peterson *et al.*, 2001).

Comparative genome analysis

Putative orthologs of the *Synechococcus* OS-A and OS-B genomes were identified by reciprocal best-match BLAST comparisons. Those CDSs that did not fit this criterion were identified as being isolate-specific (Supplementary Table S3).

Metagenomic analyses

Metagenomic sequences were generated from high and low temperature samples of the microbial mats of Octopus and Mushroom Springs. The *Synechococcus* metagenomic sequences were classified into *Synechococcus* 'OS-A-like' or 'OS-B′-like' based on BLASTN analysis of the end reads against the complete genome sequences of *Synechococcus* OS-A and OS-B′. Classification into the OS-A or OS-B′-like bins required meeting the following criteria: (a) both ends of the clone insert had the best BLASTN score to the same genome (either to *Synechococcus* OS-A or OS-B′); (b) matches of both end sequences had a nucleotide identity >92% over at least 300 bp; (c) the distance between end-paired reads, based on their known genomic positions, could not be >25% of the average clone insert size; and (d) the paired reads had to come from opposite strands of the DNA (Supplementary Table S6). These sequences represent ~17% of total metagenomic reads. By assembling the clones onto the anchor genomes, large metagenome scaffolds were created that spanned almost the entire genomes of *Synechococcus* OS-A and OS-B′ (86% with an average gap size of 931 bp and 89% with an average gap size of 617 bp for the *Synechococcus* OS-A and OS-B′ genomes, respectively). It is important to note that the genomic reconstruction resulting from this 'stringent' assembly is derived from sequenced genomic fragments from individual cells, and therefore represents a 'virtual genome', rather than the entire genome from an individual *Synechococcus* cell. The assemblies generated were annotated in the same way as for the completed genomes.

In the process of building assemblies from metagenomic sequences, we identified numerous clones with at least one end that matched the *Synechococcus* OS-A or OS-B′ anchor genomes, but that did not fulfill at least one of the four criteria described above. We refer to these as 'illegal clones'. The inserts in the illegal clones (a) had a match to an anchor genome at one end of the clone, but not at the other end (this feature might represent a gene deletion or insertion, or a region that has undergone extensive sequence divergence); (b) had a match to an anchor genome at both ends, but the distance between the paired end reads was greater than expected (this feature might represent a rearrangement of the genome); (c) had a match to an anchor genome at both ends, but the sequences were both from the same strand (this feature might represent a transversion and recombination event); or (d) had a match to an anchor genome at both ends, but the match on one of the ends only covered a very short part of the read.

Genes 'missing' in the metagenome sequence

The region encoding urease subunits and accessory proteins (*ureEFG*), present in the *Synechococcus* OS-A genome appeared to be absent in the metagenome (see below, Figure 2). Clones that would be predicted to contain the urease accessory protein region were identified and sequenced to completion by transposon mutagenesis and transposon sequencing. The completely sequenced clone inserts were annotated as described above.

Transcript analysis and irradiance measurements

The level of the *feoB* gene transcript was monitored by quantitative PCR (qPCR) over the diel cycle using mat samples from Mushroom Spring (60°C site) (Steunou *et al.*, 2006). The primers used for qPCR were feoB-F (5′-CGGGTTTGGTGATGAAAAGC) and feoB-R (5′-CCCACCGAATTTAACAAGCC). Irradiance measurements and the method for qPCR have been described previously (Steunou *et al.*, 2006).

Results and discussion

Genome sequences of two Synechococcus *isolates*

The genomes of *Synechococcus* OS-A (2.9 Mbp) and OS-B′ (3.0 Mbp) have a relatively high G+C% content of 60.3 and 58.5 and include 2892 and 2933 predicted CDSs, respectively (Supplementary Table S1). There was no evidence of plasmids or phage-like genes on the chromosome of either of the *Synechococcus* isolates. Both genomes have two rRNA operons that are identical within each isolate (Figure 1). The 16S rRNA and 16S–23S rRNA internal transcribed spacer (ITS) sequences are identical to sequences previously characterized from mat samples, confirming that we had sequenced isolates from *Synechococcus* populations

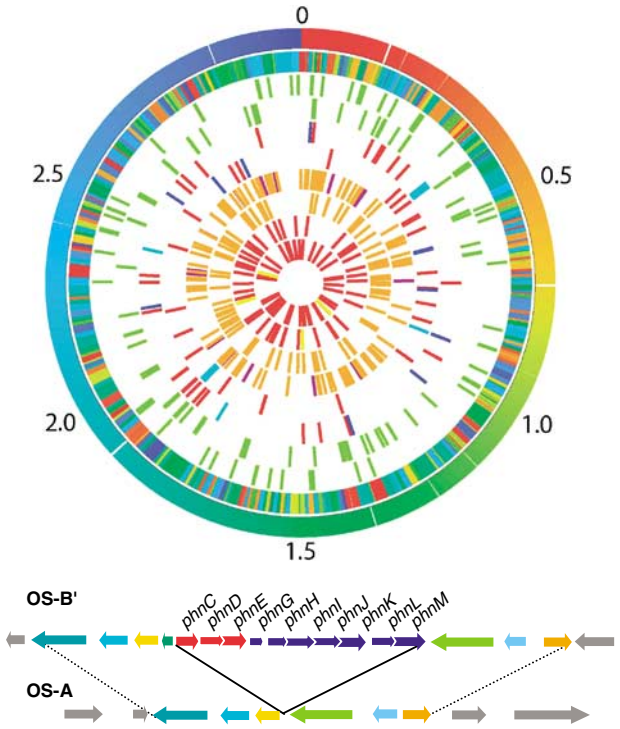


Figure 1 **Top**: genomes of *Synechococcus* OS-A and OS-B′. Circle 1 (outermost); *Synechococcus* OS-B′ genes have been assigned pseudo-spectrum colors based on relative position along the length of the genome, with genes nearest to the putative origin of replication colored red and genes most distal from this origin colored blue. Circle 2; the putative homologs in *Synechococcus* OS-A are assigned the same color as is presented in circle 1 for the *Synechococcus* OS-B′ genes; the scrambling of colors in circle 2 is a reflection of the striking lack of relative conserved large-scale gene order between the genomes of *Synechococcus* OS-A and OS-B′. Circles 3 and 4; photosynthesis genes (green). Circles 5 and 6; nitrogen fixation genes (light blue), urease genes (dark blue) and fermentation genes (red). Circles 7 and 8; transposons and integrase genes (orange) and IS4 transposases (purple). Circles 9 and 10; tRNAs (red) and rRNAs (yellow). Circles 3, 5, 7 and 9 show the relative gene positions on the genome of *Synechococcus* OS-B′. Circles 4, 6, 8 and 10 show the relative gene positions on the genome of *Synechococcus* OS-A. **Bottom**: phosphonate gene cluster in *Synechococcus* OS-B′. Regions of the *Synechococcus* OS-B′ genome displaying the *phn* genes that may be involved in phosphonate transport (*phnC, -D* and *-E*, CYB_0159 to CYB_0161, red arrows) and in the synthesis of C-P lyase (*phnG–M*, CYB_0162 to CYB_0168, blue arrows) shown in the top line, with the comparable regions of the *Synechococcus* OS-A genome in the bottom line. The *phn* cluster spans 8 kbp and the region of synteny between the two genomes that flanks this cluster is positioned between the solid and dotted lines; in *Synechococcus* OS-B′ this region is located between 169.4 and 184.3 kbp and in *Synechococcus* OS-A it is between 518.2 and 525.4 kbp. The arrows representing genes are drawn to scale (note that several of the *phn* genes overlap, as shown by overlapping arrows); potential homologs are shown in the same color; genes shown by gray arrows indicate where synteny between the genomes is not maintained.

that are dominant in the mat (Papke *et al.*, 2003). The *Synechococcus* OS-A and OS-B′ 16S and 23S rRNA sequences are 96.4% and 96.3% identical to each other, respectively, suggestive of significant divergence between the isolates, and yet the genomes of these isolates share a large fraction (~83%

based on bidirectional best BLAST scores) of their CDSs, and the CDSs exhibit a high identity between putative orthologs (~87% amino-acid identity, ~92% amino-acid similarity; ~86% nucleotide identity, on average). A comparison of the *Synechococcus* OS-A and OS-B′ genomes revealed a marked lack of conserved, large-scale gene order, indicating an extensive history of rearrangement events (Figure 1). Regions of colinearity between the *Synechococcus* OS-A and OS-B′ genomes are short (~70% of either genome is composed of conserved blocks of 1–2 kbp in length containing one or two genes), with the largest region of conserved gene order between the two isolates being ~32 kbp and containing the genes involved in nitrogen fixation (Delcher *et al.*, 2002; Steunou *et al.*, 2006). In contrast, a comparison of the complete genomes of closely related bacteria usually reveals extensive conservation of large-scale genome architecture or synteny; even genomes that are not so closely related can show significant synteny (Rocap *et al.*, 2003; Bentley and Parkhill, 2004). The lack of congruence between the measures of 16S rRNA identity, genome-wide identity between putative homologs, and genomic synteny of the two *Synechococcus* isolates implies that any one of these criteria by itself may not accurately reflect the level of relatedness between the *Synechococcus* isolates (Konstantinidis and Tiedje, 2005; Ward *et al.*, 2006).

Genome rearrangements and recombination events are often mediated by transposons or phage (Bennett, 2004). Both *Synechococcus* OS-A and OS-B′ genomes contain many transposon-like or insertion sequence (IS) elements (109 and 98, respectively), but several of these are presumed non-functional as evidenced by truncations or putative frameshift mutations. The IS elements are not always located at the borders of rearranged regions, so their role in the large-scale gene rearrangements cannot be easily assigned (Parkhill *et al.*, 2003). However, *Synechococcus* OS-B′ does contain 17 identical copies of an IS4 family of transposase genes (ISSoc13 or Interpro ID 002559), which are absent in the *Synechococcus* OS-A genome and which may still be active within the *Synechococcus* OS-B′-like organisms (Figure 1, circle 7, purple bars). Furthermore, both the *Synechococcus* OS-A and OS-B′ genomes contain a high frequency of an octameric, palindromic repeat (GGGATCCC), but the functional significance of this repeat is unclear. The presence of an internal GATC sequence within the repeat raises the possibility that they may be part of a DNA modification system that has been implicated in a number of processes ranging from restriction modification to DNA repair and transcriptional regulation (Wion and Casadesus, 2006) (Supplementary Table S2).

As expected from previous studies, both *Synechococcus* OS-A and OS-B′ contain genes encoding proteins required for photosynthesis, the biosynthesis of glycolate (Bateson and Ward, 1988), glycogen (Konopka, 1992; Nold and Ward, 1996), sulfolipids (Ward *et al.*, 1994), and fermentative and respiratory metabolisms (Nold and Ward, 1996). We also identified genes required for the biosynthesis of Type IV pili and photoreceptors associated with phototaxis, which fits with earlier reports of motility of *Synechococcus* cells in the mats (Ramsing *et al.*, 1997; Bhaya, 2004). However, the presence and activity of a functional pathway for nitrogen fixation (based on *in situ* expression *nif* genes and measurements of nitrogenase activity) in *Synechococcus* OS-A and OS-B′ was unexpected, as previous attempts had failed to measure nitrogen fixation in the mats (Steunou *et al.*, 2006).

Identification of genes specific to Synechococcus *OS-A or OS-B′*

To identify functional differences between *Synechococcus* OS-A and OS-B′, we examined subsets of genes unique to each of these isolates. There are 393 and 503 isolate-specific CDSs in *Synechococcus* OS-A and OS-B′, respectively, but approximately half of these CDSs are annotated as either 'hypothetical' (172 and 218 in *Synechococcus* OS-A and OS-B′, respectively) or 'conserved hypothetical' (50 and 46 in *Synechococcus* OS-A and OS-B′, respectively), with other CDSs potentially encoding transposases/resolvases (91 and 89 in *Synechococcus* OS-A and OS-B′, respectively) (Supplementary Table S3). However, we were able to identify examples of genes encoding proteins with known functions that are present on only one of the genomes. For instance, only the *Synechococcus* OS-B′ genome harbors an 8 kbp region containing 10 *phn*-like genes (CYB_0159 to CYB_0168), which might enable the organism to transport (*phnCDE*) and metabolize (*phnGHIJKLM*) phosphonates. This may allow *Synechococcus* OS-B′ to utilized phosphonates as a source of phosphorus in addition to phosphate, the prevalent source of phosphorus in most environments. The region flanking the *phn* cluster is syntenic between the *Synechococcus* OS-A and OS-B′ genomes, indicating that the *phn* gene cluster was either recently acquired by *Synechococcus* OS-B′ or lost in *Synechococcus* OS-A (Figure 1). There is evidence suggesting that operons required for phosphonate uptake and utilization may be acquired through lateral gene transfer events in prokaryotes (Huang *et al.*, 2005). Recently, genes for phosphonate utilization have been identified in metagenomic studies of marine, oxygenic photosynthetic prokaryotes, but the *phn* operon is not universally found in cyanobacteria, perhaps reflecting the different availability of phosphonates in various environments (Palenik *et al.*, 2003; Dyhrman *et al.*, 2006). A search for *phn*-like genes within the metagenome sequences identified 92 sequences that covered this region with high identity (that is, >92% NAID). Almost all originated from the recombinant libraries generated from Octopus or

Mushroom Spring low-temperature regions of the mat, which is consistent with the observation that *Synechococcus* OS-B′-like organisms are more prominent in the lower temperature regions (Supplementary Table S4). There may be species unrelated to *Synechococcus* that can utilize phosphonate in the mats because there are other *phn*-like genes in the metagenome dataset that have low identity (~55% AAID) to the *phn* genes of *Synechococcus* OS-B′. The origins, levels and importance of this and other phosphorus sources in the hot springs are currently being investigated.

Other genes present on the *Synechococcus* OS-B′ but not the *Synechococcus* OS-A genome are those encoding cyanophycin synthetase (CYB_0911) and cyanophycinase (CYB_2043). Cyanophycin is a nitrogen-rich reserve polymer synthesized non-ribosomally from aspartate and arginine by the enzyme cyanophycin synthetase; the polymer is degraded by cyanophycinase (Simon, 1987). Cyanophycin levels vary with growth conditions, but can be high in stationary-phase cultures or under conditions in which the growth potential of the cell decreases because of a limitation for nutrients such as sulfate or phosphate (Stevens and Poane, 1981). Cyanophycin has also been implicated in the integration of carbon and nitrogen metabolism in unicellular and filamentous cyanobacteria (Mackerras *et al.*, 1990). The presence of cyanophycin synthetase and cyanophycinase on the *Synechococcus* OS-B′ genome suggests that this organism experiences fluctuating nitrogen levels; conditions of excess nitrogen may trigger cyanophycin storage, whereas periods of nitrogen limitation may result in cyanophycin degradation. Further experimental evidence in which gene expression in *Synechococcus* OS-B′ is measured under various nutrient regimes under defined laboratory conditions or *in situ* over the diel cycle, will allow us to address these questions.

Evidence of recent acquisition or loss of nutrient utilization pathways

We noted differences between the *Synechococcus* OS-A and OS-B′ isolates for genes required for the utilization of urea. *Synechococcus* OS-A has one genomic region encoding urease (*ureA1B1C*) and accessory factors (*ureEFG1D1*) (cluster 1 urease in Figure 2, top left), which is not present in *Synechococcus* OS-B′. Additionally, there are other regions in both isolates with remnants of genes encoding a second urease that is likely to be non-functional (Figure 2, bottom left). The genes of cluster 1 urease are flanked by transposons. The regions bordering the urease cluster (CYA_0598 and CYB_0030; CYA_0609 and CYB_0031; CYA_0610 and CYB_0033) are syntenic between *Synechococcus* OS-A and OS-B′, suggestive of a relatively recent gain of genes by *Synechococcus* OS-A, or loss of

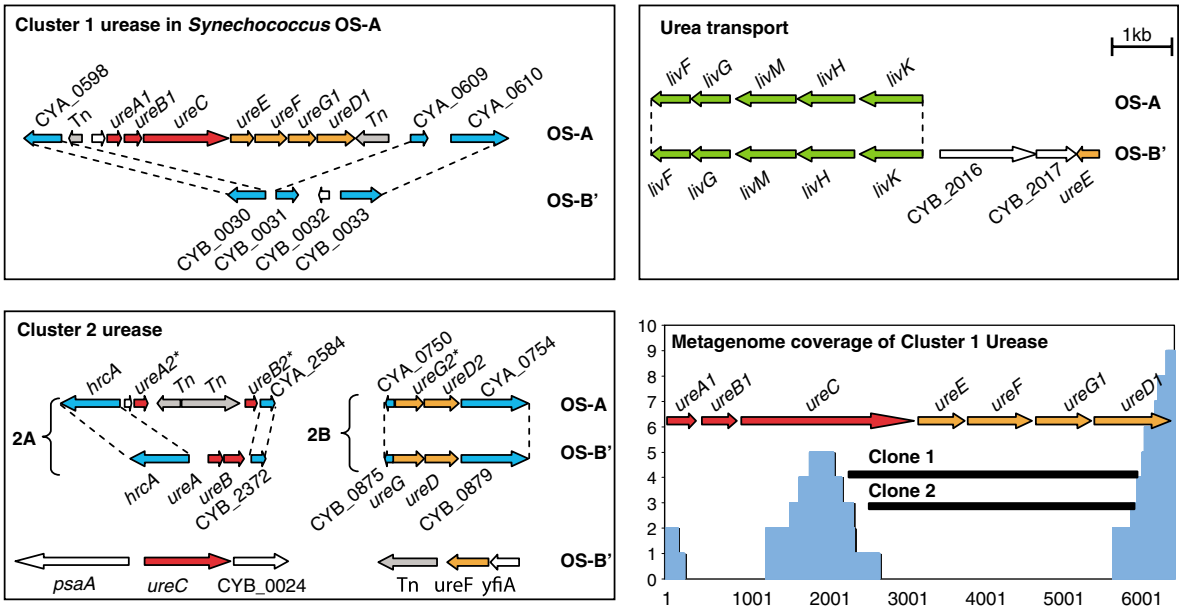


Figure 2 Urease genes in *Synechococcus* OS-A and OS-B′. Top left: top line represents the cluster 1 urease and flanking genes of *Synechococcus* OS-A; the bottom line shows homologs of the flanking genes in *Synechococcus* OS-B′. Bottom left: locations of the cluster 2 urease genes in *Synechococcus* OS-A and OS-B′. In *Synechococcus* OS-A, *ureB2* is inactivated by a transposon; *ureA* and *ureG2* contain frameshift mutations (inactive genes are designated by * symbols). Top right: the urea transporter genes (*livF* to *livK*) of *Synechococcus* OS-A and OS-B′. In *Synechococcus* OS-B′, *ureE* is located close to the transporter genes. Bottom right: metagenome sequence redundancy (*y* axis) in the region of the cluster 1 urease of *Synechococcus* OS-A (*x* axis). Note that no metagenome sequences covered the region encoding *ureE, ureF* and *ureG1*; clones identified that covered this region are shown as black bars. Genes encoding urease (*ureA, -B* and *-C*) are shown in red; accessory factor genes are shown in orange; transporter genes in green; putative transposons are shown in gray; flanking gene homologs in blue and other genes in white. Syntenic regions are indicated by dashed lines.

genes from *Synechococcus* OS-B′. The second set of urease-like genes in *Synechococcus* OS-A (*ureA2, ureB2,* and *ureG2, ureD2*, making up cluster 2A and 2B, respectively) are at two different genomic locations. The gene *ureB2* (CYA_2585) is truncated by the integration of a transposon, whereas *ureA2* (CYA_2588) and *ureG2* (CYA_0751) contain frameshift mutations. A second *ureC* and *ureF* were not identified on the *Synechococcus* OS-A genome. In contrast, the *Synechococcus* OS-B′ genome contains a single copy of the *ure* genes that are located at five different regions of the genome. These genes have higher identity with the genes of *Synechococcus* OS-A cluster 2 than those of cluster 1, and in some cases regions flanking the cluster 2 urease genes in *Synechococcus* OS-A are syntenic with those flanking the homologous genes in *Synechococcus* OS-B′ (Figure 2, bottom left). In *Synechococcus* OS-B′ the cluster 2 urease genes have neither insertions nor frameshift mutations, and are therefore likely to represent functional genes. Also both *Synechococcus* OS-A and OS-B′ genomes contain genes encoding urea transport polypeptides (*livFGMHK*), which have very high identity between the isolates (~99% amino-acid identity) (Figure 2, top right).

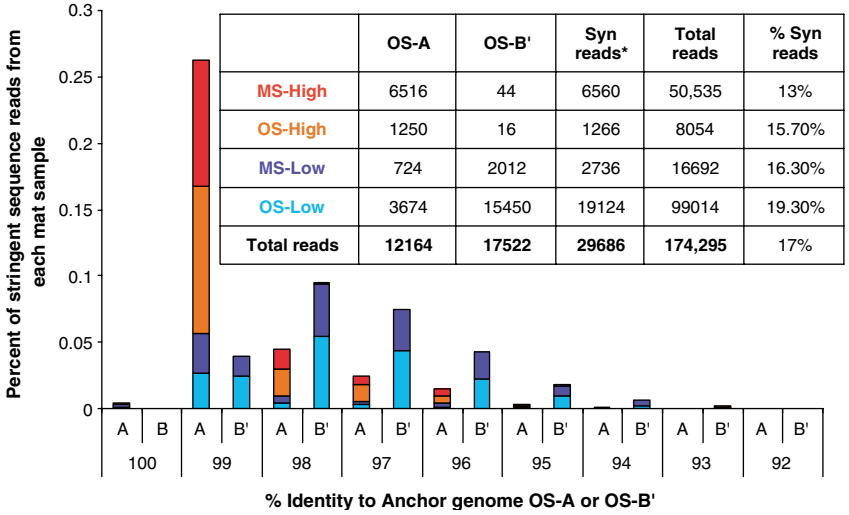
The functionality of the different urease genes has yet to be experimentally verified, but these obvious differences in gene content underscore the possibility that both gene gain and loss are ongoing events in the mat *Synechococcus* populations. The mechanisms underlying this genome fluidity and their evolutionary implications remain to be explored, but we speculate that a relatively recent acquisition of urease cluster 1 (flanked by transposons) by *Synechococcus* OS-A may have led to the progressive loss of functionality of the second urease cluster (still maintained by, and likely to be functional in *Synechococcus* OS-B′). We have recently developed an oligonucleotide microarray that will be used to query the isolate transcriptomes under different nutrient conditions, which will allow us to check the expression status of the urease gene clusters in *Synechococcus* OS-A and OS-B′.

Initial analysis of the metagenome sequence dataset indicated that sequences representing the *ureEFG1D1* of the cluster 1 urease were absent (Figure 2, bottom right), raising the possibility that the *Synechococcus* OS-A anchor genome sequence may not be representative of the predominant *Synechococcus* OS-A-like populations in the mat. To clarify this, we identified and sequenced clones from the metagenome dataset that were predicted to span the region containing urease cluster 1 genes, based on clone-end read information. These clones (GYSA072 and YMBDB52) did contain the urease genes, so the apparent absence of these sequences in the metagenome dataset is possibly a reflection of a region that did not get sequenced at the same frequency as flanking regions. Thus, correlating sequence frequency data from metagenomic datasets with actual frequencies of genes in populations may require close attention to methodological constraints and statistical analysis of the data, followed by further verification before strong conclusions can be drawn about underrepresented genes in a metagenome database (Tringe *et al*., 2005).

Identification of functional diversity in the metagenome

The variations in nutrient acquisition and utilization capabilities suggested by differences in genome content between the two sequenced *Synechococcus* isolates raise the possibility that a detailed examination of *Synechococcus*-like sequences in the metagenome dataset might provide further examples of functionally specialized populations. To accomplish this, we characterized metagenomic sequences derived from both Mushroom and Octopus Springs at two different temperatures (Materials and methods and Supplementary Table S5), on the basis of identity to the *Synechococcus* anchor genomes. We found a significantly higher degree of identity of *Synechococcus* OS-A-like sequences in the metagenomic samples with the anchor genome (85% of the sequences with a best BLASTN score to the OS-A genome were within 2% divergence of the *Synechococcus* OS-A anchor genome) compared to that of the *Synechococcus* OS-B′-like sequences (only 50% of the sequences with a best BLASTN score to the *Synechococcus* OS-B′ genome were within 2% divergence of the *Synechococcus* OS-B′ anchor genome) (Figure 3). These results correlate well with the finding that there is a greater depth of rRNA divergence within the *Synechococcus* OS-B′-like population than within the *Synechococcus* OS-A-like population, and that there is generally less sequence variation among members of the *Synechococcus* OS-A-like population (Ward *et al*., 1998). The greater sequence diversity in the *Synechococcus* OS-B′-like than in the *Synechococcus* OS-A-like metagenome is also consistent with a greater diversity of ecotypes among the *Synechococcus* OS-B′-like organisms (Ward *et al*., 2006), as based on a simulation of ecotype evolution (Cohan and Perry, 2007).

Furthermore, identification of metagenomic sequences based on their source (for example, Octopus or Mushroom Springs, low or high temperatures) yielded a picture supporting the notion that temperature and spring characteristics represent important factors in shaping community diversity (Figure 3 and Supplementary Table S5). In accord with this idea, few *Synechococcus* OS-B′-like sequences were found in libraries from the higher temperature regions of Octopus or Mushroom Spring mats, whereas *Synechococcus* OS-A-like sequences were present in all four of the libraries, but represented a higher proportion of sequences in the those libraries constructed with DNA derived from mat samples from the higher temperature regions (Figure 3, and inset Table). These findings



	OS-A	OS-B'	Syn reads*	Total reads	% Syn reads
MS-High	6516	44	6560	50,535	13%
OS-High	1250	16	1266	8054	15.70%
MS-Low	724	2012	2736	16692	16.30%
OS-Low	3674	15450	19124	99014	19.30%
Total reads	**12164**	**17522**	**29686**	**174,295**	17%

Figure 3 Distributions of metagenome reads compared to the anchor *Synechococcus* OS-A and OS-B′ genomes. The reads that met the criteria for inclusion into 'stringent' *Synechococcus* OS-A and OS-B′ metagenome assemblies were subdivided based on the percentage nucleotide identity to the anchor genome (on the *x* axis, from 100 to 92%). Each of these sequence reads was also identified based on the location from which it originated: Mushroom Spring high temperature (MS-High-*Red*) or Mushroom Spring low temperature (MS-Low-*Dark Blue*) or Octopus Spring high temperature (OS-High-*Orange*) or Octopus Spring low temperature (OS-Low-*Light Blue*). Because the total number of sequences from each mat sample varied, the data were normalized to the total reads in the sample from which the reads were derived. Thus, the *y* axis values are shown as the percentage of the stringent sequence reads from each of the four mat locations. The inset table presents the number of sequence reads from each spring, at each of the two temperatures and places them into the stringent *Synechococcus* OS-A (OS-A) or OS-B′ (OS-B′) bin. Syn refers to *Synechococcus* OS-A and OS-B′ and 'Total reads' refers to all reads from a particular spring and temperature.

raise a number of interesting questions regarding the range of sequence variation within the *Synechococcus* OS-A and OS-B′ populations, which can be addressed by further analysis of the metagenomic sequence dataset or by using a targeted approach to examine particular genes or regions of interest.

Using stringent assembly criteria (see Materials and methods and Supplementary Table S6), 12 450 and 17 521 metagenome reads were assembled into scaffolds that were anchored to the *Synechococcus* OS-A and OS-B′ genome, respectively. These reads covered 86 and 89% of the *Synechococcus* OS-A and OS-B′ anchor genomes, respectively. 21 674 reads matched either the *Synechococcus* OS-A or OS-B′ genome, but the clones from which the read was derived could not be assembled onto the anchor genome because they failed to meet one of the assembly criteria ('illegal clones'), and we surmised that they might represent regions in which recombination, transversion or gene gain or loss may have occurred. Hence, these clones represent a potentially rich source for analysis of genetic variation (Materials and methods and Supplementary Table S6).

We investigated a category of illegal clones in which one end of the clone had high sequence identity (>95% NAID) to a specific region on the *Synechococcus* OS-B′ genome, whereas the paired end sequence did not match any sequences in the *Synechococcus* OS-B′ genome. These clones could represent a *Synechococcus* population closely related to *Synechococcus* OS-B′ which contains additional sequences that are absent in the anchor genome. We identified several individual metagenome clones in which only one end of the clone matched the *Synechococcus* OS-B′ genomic region with high identity in the neighborhood of ~0.565–0.577 Mb on the genome. To resolve how these 'illegal' clones differed from the anchor genome, we identified and sequenced a 7.6 kbp clone that matched the *Synechococcus* OS-B′ genome at both ends but contained an extra 5.5 kbp region relative to the *Synechococcus* OS-B′ anchor genome. In the *Synechococcus* OS-B′ genome, this region contains four genes, CYB_0562, CYB_0563, CYB_0564 and CYB_0565 (Figure 4, top). The metagenomic clone insert, however, contains seven additional CDSs, and is flanked by CYB_0562 (99.72% NAID) on the left and CYB_0565 (99.76% NAID) on the right (Figure 4a), whereas CYB_0563, CYB_0564 genes are absent. Of the genes present in the 5.5 kbp region two exhibited significant identity to the *feoA* (46% AAID to tlr1739) and *feoB* (64% AAID to tlr1740) genes of the unicellular thermophilic cyanobacterium, *Thermosynechococcus elongatus.*

The presence of the *feoA* and *feoB* genes in a *Synechococcus* OS-B′-like population is interesting because these genes are present in several bacteria where they encode proteins required for ferrous ion transport (Andrews *et al.*, 2003). Iron is present in the environment as both ferrous and ferric forms, but the ferrous form predominates under conditions of low oxygen and is relatively soluble. Neither the *Synechococcus* OS-A nor OS-B′ anchor genomes contain *feoA* or *feoB*-like sequences, although they have several genes that function in ferric ion uptake

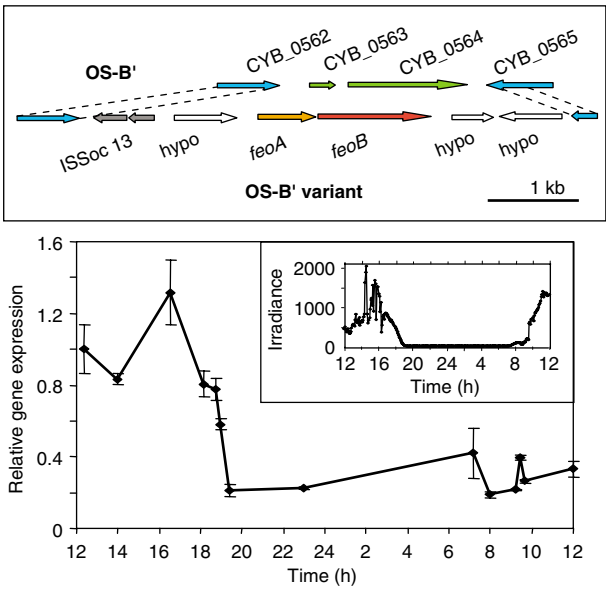


Figure 4 *Feo* containing variant of *Synechococcus* OS-B′. Top panel shows the region containing CYB_0562 (SpoIID protein), CYB_0563 (hypothetical), CYB_0564 (helicase, UvrD/REP family) and CYB_0565 (*menB*, napthoate synthase) from *Synechococcus* OS-B′. Below is the variant clone containing the *feo* genes plus seven additional CDSs; from left, a CDS with 99% AAID to CYB_0562, OrfB and OrfA of ISSoc13 transposase which appears to be *Synechococcus* OS-B′-specific, a 221-amino-acid hypothetical CDS with similarity to a response regulator, *feoA* (158 amino acids), *feoB* (763 amino acids), a 74-amino-acid hypothetical CDS, a 157-amino-acid hypothetical CDS with high identity to CYA_0664 and CYB_0618, and a partial CDS with high identity to CYB_0565. Note that this clone does not contain the entire CYB_0565 gene as indicated by the position of the dotted line. Bottom panel shows *in situ feoB* transcript abundance measured by qPCR in the Mushroom Spring microbial mats (60°C) in September 2005, over a 24 h diel cycle. The inset shows incident photon irradiance (small black circles) in μmol photon $m^{-2} s^{-1}$ as a function of the time of day.

and assimilation. The presence of *feo* genes in an organism that may be closely related to *Synechococcus* OS-B′ is interesting because it suggests the presence of functionally specialized populations with differential specificity for the different forms of iron. The functionality of the *feo* genes in the mat is supported by *in situ* analysis of *feoB* transcript levels (quantified using gene-specific primers for qPCR), which appear to peak in the afternoon, about 2 h before darkness; transcript levels stay low during the night and remain low as the cells are exposed to light in the morning (Figure 4, bottom). Thus, there may be an accumulation of the ferrous transport system as the mat becomes anoxic during the night, allowing some *Synechococcus* OS-B′-like organisms to scavenge the reduced ferrous ions that should become more plentiful as oxygen levels decrease in the mat.

These results provide an example of a comparison of metagenomic sequences with fully sequenced anchor genomes that has helped identify specialized functional populations that potentially have a different physiological capability and ecology. It is worth noting that so far, we have analyzed metagenomic clones that are relatively small (between 2 and 15 kbp), so we cannot easily place them into larger scaffolds with a high degree of certainty. Consequently, we are unable to decipher the degree of similarity between the organism from which this clone was derived and the *Synechococcus* OS-B′ anchor genome. However, screening of BAC libraries that contain much larger clones (>100 kbp) will allow us to identify large contiguous regions of genomes that contain these sequences. This will allow us to assess how similar these metagenomic clones are to the sequenced *Synechococcus* anchor genomes. The hypothesis that these *feo* containing organisms correspond to ecotypes different from that represented by *Synechococcus* OS-B′ would be supported if the *feo*-containing organisms were found to be members of distinct sequence clusters for widely shared genes (for example, *recA*, *rpoB* and *lepA*) (Cohan and Perry, 2007). Further analyses of the other classes of illegal clones will also allow us to create a more detailed picture of the level and kinds of functional variation that exists within the *Synechococcus* populations.

In summary, we have generated two 'anchor' genome sequences from *Synechococcus* OS-A and OS-B′, which represent dominant high- and low-temperature adapted *in situ* mat populations, respectively. Comparison of these two genomes revealed a striking lack of conserved large-scale genome architecture, suggestive of extensive genome rearrangements. Many obvious differences between these genomes appear to be related to the assimilation and storage of different nutrients such as nitrogen, phosphorus and iron. Further, population genetics and *in situ* distribution and expression analyses coupled with detailed physicochemical measurements will be required to ascertain whether the differences in gene complement observed constitute niche-adaptation strategies in the hot springs. The implications of the observations makes it important to define nutrient gradients and fluxes within the microbial mat communities, and to elucidate the ways in which these gradients impact the evolution of microbes in this environment.

We also analyzed metagenome sequences from the microbial mat community and compared them to the *Synechococcus* OS-A and OS-B′ anchor genomes. The spatial distribution and degree of diversity of the metagenomic variation is consistent with that for *Synechococcus* OS-A-like and OS-B′-like populations previously observed by using 16S rRNA and ITS sequences (Ward *et al.*, 1998, 2006). Through comparative genomic analyses we have demonstrated that the genomes of these *Synechococcus* OS-A and OS-B′ isolates are not representative of all native *Synechococcus* OS-A and OS-B′ populations inhabiting the mat community that were sampled in our metagenomic sequences. Our analyses have revealed examples of functionally specialized populations that may have been derived from gene

exchange and may define previously unknown *Synechococcus* OS-A-like or B′-like functional populations (that is, ecotypes). More extensive analysis of these genomic and metagenomic databases may reveal whether or not genomic fluidity is a more general phenomenon associated with hot spring mat communities, where organisms exist in close proximity and where DNA transfer through processes such as natural transformation or phage infections might occur with some frequency. To adequately address the question of niche adaptation, we will need to correlate functional differences (based on gene content and enzyme activity *in situ*) with ecological differences (based on microgeographic distribution) and sequence clustering of shared genes. As appreciation of, and information about, the astounding genetic diversity of microbes increases; it is also becoming obvious that population genetics, systematics, genomics and metabolic potential must be integrated to resolve complex issues about microbial communities.

Acknowledgements

The research was funded by the Frontiers in Integrative Biology (FIBR) program at NSF (Grant EF-0328698). Natalia Khuri acknowledges a NSF-ROA Grant. The sequence of the chromosome of *Synechococcus* OS-A and *Synechococcus* OS-B′ have been deposited at GenBank (CP000239 and CP000240). Sequences of clones covering the *feo* region (EU189023); the urease region (EU189024 and EU189025) and metagenome sequences (Project IDs 20717, 20719, 20721,20723, 20725 and 20727) are also available at Genbank. Updates and further details on this project will be maintained at the following web sites: http://landresources.montana.edu/FIBR/ and http://fumarole.stanford.edu. We thank the Yellowstone National Park authorities for their excellent support. We also acknowledge the discussions with members of the Annual FIBR workshops held at the University of Montana, Bozeman.

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.

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

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res* **32**: D138–D141.

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

Bennett PM. (2004). Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods Mol Biol* **266**: 71–113.

Bentley SD, Parkhill J. (2004). Comparative genomic structure of prokaryotes. *Annu Rev Genet* **38**: 771–792.

Bhaya D. (2004). Light matters: phototaxis and signal transduction in unicellular cyanobacteria. *Mol Microbiol* **53**: 745–754.

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

Claros MG, von Heijne G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* **10**: 685–686.

Cohan F. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans Roy Soc B* **361**: 1985–2006.

Cohan FM, Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373–R386.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**: 2478–2483.

Dyhrman ST, Chappell PD, Haley ST, Moffett JW, Orchard ED, Waterbury JB *et al.* (2006). Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*. *Nature* **439**: 68–71.

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.

Haft DH, Selengut JD, White O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**: 371–373.

Heidelberg JF, Seshadri R, Haveman SA, Hemme CL, Paulsen IT, Kolonay JF *et al.* (2004). The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nat Biotechnol* **22**: 554–559.

Hoehler TM, Bebout BM, Des Marais DJ. (2001). The role of microbial mats in the production of reduced gases on the early Earth. *Nature* **412**: 324.

Huang J, Su Z, Xu Y. (2005). The evolution of microbial phosphonate degradative pathways. *J Mol Evol* **61**: 682–690.

Kilian O, Steunou AS, Fazeli F, Bailey S, Bhaya D, Grossman AR. (2007). Responses of a thermophilic *Synechococcus* isolate from the microbial mat of Octopus Spring to light. *Appl Environ Microbiol* **73**: 4268–4278.

Konopka A. (1992). Accumulation and utilization of polysaccharide by hot-spring phototrophs during a light–dark transition. *FEMS Microbiol Lett* **102**: 27–32.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.

Mackerras AH, Youens BN, Wier RC, Smith GD. (1990). Is cyanophycin involved in the integration of nitrogen and carbon metabolism in the cyanobacteria *Anabaena cylindrica* and *Gleothece* grown on light/dark cycles? *J Gen Microbiol* **136**: 2049–2056.

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, Kuhl 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.

Paerl HW, Pinckney JL, Steppe TF. (2000). Cyanobacterial-bacterial mat consortia: examining the functional unit

of microbial survival and growth in extreme environments. *Environ Microbiol* **2**: 11–26.

Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P *et al.* (2003). The genome of a motile marine *Synechococcus*. *Nature* **424**: 1037–1042.

Papke RT, Ramsing NB, Bateson MM, Ward DM. (2003). Geographical isolation in hot spring cyanobacteria. *Environ Microbiol* **5**: 650–659.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE *et al.* (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**: 32–40.

Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. (2001). The comprehensive microbial resource. *Nucleic Acids Res* **29**: 123–125.

Ramsing NB, Ferris MJ, Ward DM. (2000). Highly ordered vertical structure of *Synechococcus* populations within the one-millimeter-thick photic zone of a hot spring cyanobacterial mat. *Appl Environ Microbiol* **66**: 1038–1049.

Ramsing NB, Ferris MJ, Ward DM. (1997). Light-induced motility of thermophilic *Synechococcus* isolates from Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **63**: 2347–2354.

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.

Simon RD. (1987). Inclusion bodies in the cyanobacteria: cyanophycin, polyphosphate, polyhedral bodies. In: Fay P, Van Baalen C (eds). *The Cyanobacteria*. Elsevier Science Publishers B.V.:Amsterdam, pp 199–225.

Stal LJ. (2000). Cyanobacterial mats and stromatolites. In: Potts M, Whitton BA (eds). *The Ecology of Cyanobacteria*. Kluwer Academic Publishers: Norwell, MA, pp 61–120.

Steunou AS, 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.

Stevens SE, Poane DAM. (1981). Accumulation of cyanophycin granules as a result of phosphate limitation in *Agmenellum quadruplicatum*. *Plant Physiol* **67**: 716–719.

Teske A, Stahl D. (2002). Microbial mats and biofilms: evolution, structure, and function of fixed microbial communities. In: Staley JT, Reysenbach AL (eds). *Biodiversity of Microbial Life*. Wiley-Liss: New York, pp 49–100.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Ward DM. (1998). A natural species concept for prokaryotes. *Curr Opin Microbiol* **1**: 271–277.

Ward DM, Bateson MM, Ferris MJ, Kuhl 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, Castenholz RW. (2000). Cyanobacteria in geothermal habitats. In: Whitton BA, Potts M (eds). *The Ecology of Cyanobacteria*. Kluwer Academic Publishers: Norwell, MA, pp 37–59.

Ward DM, Cohan FM. (2005). Microbial diversity in hot spring cyanobacterial mats: pattern and prediction. In: Inskeep B, McDermott T (eds). *Geothermal Biology and Geochemistry in Yellowstone National Park*. Thermal Biology Institute, Montana State University: Bozeman, MT, pp 185–201.

Ward DM, Cohan FM, Heidelberg J, Bhaya D, Kuhl M, Grossman AR. (2007). Genomics, environmental genomics and the issue of microbial species. *Heredity* (6 June 2007; E-pub ahead of print).

Ward DM, Ferris MJ, Nold SC, Bateson MM. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62**: 1353–1370.

Ward DM, Panke S, Kloppel KD, Christ R, Fredrickson H. (1994). Complex polar lipids of a hot spring cyanobacterial mat and its cultivated inhabitants. *Appl Environ Microbiol* **60**: 3358–3367.

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

Ward DM. (2006). A macrobiological perspective on microbial species. *Microbe* **1**: 269–278.

Wion D, Casadesus J. (2006). N6-methyl-adenine: an epigenetic signal for DNA–protein interactions. *Nat Rev Microbiol* **4**: 183–192.

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