Supplemental Materials

Supplemental Materials and Methods, Figures S1 and S2, and

Tables S1, S2 and S3

for

Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat

Zhenfeng Liu[1], Christian G. Klatt[2], Jason M. Wood[2], Douglas B. Rusch[3], Marcus Ludwig[1], Nicola Wittekindt[1], Lynn P. Tomsho[1], Stephan C. Schuster[1], David M. Ward[2], and Donald A. Bryant[1]

[1]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802 USA; [2]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717-3120 USA; [3]The J. Craig. Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 USA

Supplemental Materials and Methods

RNA extraction. Diethyl pyrocarbonate (DEPC) -treated 10 mM sodium acetate, pH 4.5, (250 µl) and 500 mM Na_2-EDTA, pH 8.0, (37.5 µl) were added to tubes containing mat core samples, and the samples were homogenized by bead-beating with a velocity of 6.5 m s^{-1} for 10 s (Fastprep FP120, Savant Instruments Inc., Farmingdale, NY). DEPC-treated lysis buffer (375 µl) containing 10 mM sodium acetate and 10% (w/v) sodium dodecyl sulfate (pH 4.5) was added to the cells, which were lysed during incubation at 65 °C for 3 min. Acidic phenol equilibrated with DEPC-treated H_2O (700 µl) was added, and the samples were incubated at 65 °C for an additional 3 min. Two subsequent extractions with equal parts of Tris-HCl-equilibrated phenol (pH 8) and chloroform (1:1) were performed. Nucleic acids were precipitated by adding 0.1 volume of 10 M $LiCl_2$ and 2.5 volumes of absolute ethanol; after a 30-min incubation at -20 °C, the solutions were centrifuged at 17,000 × *g* for 30 min at 0 °C. The resulting pellets were resuspended in DEPC-treated H_2O (88 µl), and two successive DNase treatments were performed using Ambion Turbo DNAse^{TM} (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. A final extraction with chloroform:isoamyl alcohol (24:1, v/v) was performed on the DNAse-treated solution to remove protein and residual phenol, and RNA was precipitated from the aqueous phase with 10 M $LiCl_2$ and absolute ethanol as described above. The RNA was pelleted by centrifugation, washed, and resuspended in DEPC-treated H_2O (60 µl). RNA concentrations and purity were estimated by absorbance spectrophotometry at 260 nm and 280 nm with a NanoDrop Spectrophotometer ND-1000 (Thermo Fisher Scientific, Wilmington DE), and RNA integrity was verified by analyzing aliquots on an RNA NanoChip with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Samples with RNA integrity numbers of at least ~7 (range 6.9 to 7.5) were considered to be acceptable for further analyses (Schroeder et al., 2006).

cDNA preparation and sequencing. Small RNA molecules were removed from the total RNA samples using MEGAclear^{TM} (Ambion, Foster City, CA), and the resulting RNA samples were dissolved in doubly distilled H_2O. The RNA concentration was determined using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington DE). For pyrosequencing, cDNA was prepared from 5.0 µg of template RNA using the Just cDNA Double-Stranded cDNA Synthesis Kit (Agilent Technologies, Santa Clara, CA). Libraries were constructed from these cDNA samples, and pyrosequencing was performed in the laboratory of Stephan C. Schuster at The Pennsylvania State University on a Roche GS FLX system (454 Life Science, Branford, CT). Construction of cDNA libraries and SOLiD^{TM}-3 sequencing was performed in the Genomics Core Facility at The Pennsylvania State University (University Park, PA). The cDNA libraries were constructed from 0.5 µg RNA samples according to the "Whole Transcriptome Library Preparation for SOLiD Sequencing" protocol (Applied Biosystems, Foster City, CA), and samples were barcoded using multiplexing barcode set B (Applied Biosystems, Foster

City, CA). The SOLiD ePCR and SOLiD Bead Enrichment Kits (Applied Biosystems, Foster City, CA) were used for processing the samples for sequencing, and the SOLiD^{TM}-3 System (Applied Biosystems, Foster City, CA) was used for sequencing.

Databases used in data analysis. Our strategy was to associate metatranscriptomic sequences with each of the major phylogenetic groups that, as suggested by Klatt et al. (2011), might also represent physiological groups (*i. e.*, guilds). Two *Synechococcus* spp. strains A and B′ were very closely related in both sequence and physiology, and thus cDNA sequences aligned to the reference genomes/metagenomes of these two strains have been grouped in the analyses presented here unless otherwise specified (*e. g.*, Suppl. Table 2).

i) **rRNA database.** A curated rRNA database (Urich et al., 2008), which includes a small subunit (16S rRNA, SSU) database (SSUrdb) and a large subunit (23S rRNA, LSU) database (LSUrdb), was downloaded from http://www.bioinfo.no/services/community-profiling. The SSUrdb was updated with bacterial sequences from RDP-II release 10.10 (Cole et al., 2007), and only SSU sequences for which the taxonomic affiliations (NCBI taxonomy) were clear to at least the class level were included. The LSUrdb was updated in similar manner with bacterial LSU sequences longer than 1900 bp from the SILVA project release 98 (Pruesse et al., 2007). The two updated databases were aligned against each other using BLASTN, and those sequence regions that aligned to any sequences from the other database with a bit score of 71 or higher were cropped to ensure that there was no overlap between the two reference databases. For in-depth analyses of the 16S rRNA sequences of the four major taxa containing chlorophototrophs (*i. e., Cyanobacteria, Acidobacteria, Chloroflexi,* and *Chlorobi*), sequences in these phyla from RDP-II release 10.10, for which the taxonomic affiliation was clear at least to the genus level were retrieved and used as databases for higher phylogenetic resolution. Complete or partial 16S rRNA sequences of unclassified members of the *Chlorobi* and *Chloroflexi,* respectively*,* were extracted from the mat metagenome sequences (Klatt et al. 2011) and used as additional references in both the overview and the in-depth rRNA analyses.

ii) **Database for analysis of mRNA sequences.** Nucleotide sequences of all assembled metagenome scaffolds (Klatt et al., 2011) were used as the reference database for this analysis. In addition to scaffolds clustered and assigned to different taxa in the metagenome study, all unclustered scaffolds were manually inspected, and some of these scaffolds were then assigned to the major taxa in the community based on the similarity of their gene contents and sequences to complete reference genome sequences. Scaffolds 2kb or larger were assigned to *Roseiflexus* spp., *Synechococcus* spp., *Cab. thermophilum* and *Chloroflexus* spp. when >90% of their sequences could be aligned to the complete *Roseiflexus* sp. RS-1, *Synechococcus* sp. Strain A and strain B′, *Cab. thermophilum* and draft *Chloroflexus* sp. 396-1 genomes with >80% sequence identity. Eleven additional scaffolds originating from *Chloroflexus* sp. were identified that contained genes putatively involved in photosynthesis and bacteriochlorophyll biosynthesis and that also had >80% nucleotide identity with the *Chloroflexus* sp. 396-1 draft genome. A scaffold that contained *pufLM* genes, which encode the subunits of a type-2 photochemical reaction center and which were highly divergent from the known FAP sequences in phylogenetic analysis, was included with the scaffolds originating from the uncultivated organisms belonging to the kingdom *Chloroflexi* (Klatt et al., 2011). This uncultivated *Chloroflexi* cluster, related to *Anaerolineae*-like organisms, contained genes putatively involved in bacteriochlorophyll biosynthesis; and thus provided evidence that these scaffolds originate from a currently undescribed and uncultured chlorophototroph. A list of best BLASTP hits in the nr database was also inspected for each putative gene for scaffolds that were 5 kb or larger. Scaffolds were assigned to *Chlorobiales* when more than half of the genes in the scaffold had best hits to genes in any *Chlorobiales* genome, which in most cases was that of *Chp. thalassium*. In total, 209, 181, 57, 493, 29 and 80 scaffolds were assigned or clustered to *Roseiflexus* spp., *Chloroflexus* spp., uncultivated *Chloroflexi*, *Synechococcus* spp., *Cab. thermophilum*, and *Chlorobiales* spp., respectively.

Analysis of pyrosequencing datasets. Sequences generated by pyrosequencing were processed through a pipeline similar to one that was employed in a previous metatranscriptome analysis (Urich et al., 2008). This pipeline was designed to identify and remove rRNA sequences first, followed by mRNA sequences of the several phototrophs that are the major components of the microbial community (see Klatt et al., 2011 and Results), and then other possible mRNA sequences.

i) **Analysis of rRNA sequences.** All sequences were aligned to SSUrdb and LSUrdb using BLASTN. Alignments were analyzed using MEGAN (Huson et al., 2007) to infer community structure. The absolute minimum alignment score cutoff was set to 71 and the relative cutoff was set to within 10% of the top score. These settings were chosen because they were tested and found to assign mRNA sequences with similar lengths accurately to the kingdom level (Urich et al., 2008). (Note: There is currently disagreement as to whether major sub-domain lineages for Bacteria should be considered kingdoms or phyla (see Ward et al., 2008), and here we have chosen to use the term "kingdom.") Changing the absolute cutoff to 80 or the relative cutoff to 5% or 20% did not significantly alter the perception of community structure at the kingdom level. Because previous studies (*e. g.,* Urich et al., 2008) and initial analyses showed that there was no significant difference between the compositional information deduced for the 16S and 23S rRNAs, the data for both were summed and reported here. For an in-depth survey of the taxonomic composition of the four major kingdoms, (specifically, *Chloroflexi*, *Cyanobacteria*, *Chlorobi* and *Acidobacteria*), SSU rRNA sequences assigned to these four taxa were extracted and aligned to their respective high resolution databases using BLASTN. Alignments were analyzed in MEGAN in a similar manner except that the relative cutoff was set to 0%, which means that only the top-scoring alignments were taken into consideration. This analysis provided a general picture of the taxonomic composition of the major kingdoms within the mat community.

ii) **Analysis of mRNA sequences.** Sequences that did not have an alignment score of 71 or higher in the previous step were aligned to the metagenome with BLASTN. A table of all putative genes was obtained from the annotation of the assembled metagenome scaffolds and the metagenomic sequence content of these scaffolds in comparison to isolate genomes (Klatt et al., 2011). A sequence was assigned to a gene when the gene was its best hit and its alignment had at least 90% nucleic acid sequence identity. For each gene in the database, the number of mRNA sequences assigned to it was counted for each time-point RNA sample. All unassigned sequences were aligned to peptide sequences from the NCBI nr database downloaded in April 2009 using BLASTX. Sequences with a bit score of 40 or higher were considered as potential mRNA sequences from other sources. All remaining sequences were assigned as "unidentified." Based upon a manual inspection of the data, a bit score of 40 equaled an e-value of ~1 × 10^{-4} in the search and should be a reasonable cutoff for potential mRNA sequences, given that some may represent novel genes without good references in current databases. Increasing the cutoff to 50 or 60 had almost no effect on the counts of assigned mRNA sequences, because these sequences were usually aligned to a reference with very high scores anyway.

Analysis of SOLiD™ datasets. Sequencing results were first converted into fastq format using the PERL script supplied in the bwa software package (Li & Durbin, 2009). Sequences in fastq were aligned to databases in color space using bwa with options "-c -n 5" allowing a maximum of 5 mismatches per read (≥90% nucleotide sequence identity) (see below for validation of these alignment criteria). Sequences were first aligned to the above-mentioned rRNA database (LSUrdb+SSUrdb) to remove rRNA sequences. rRNA sequences were not analyzed because the much shorter SOLiD sequences have much less phylogenetic resolution than those provided from the pyrosequencing results due to their shorter length (~50 bp compared to ~225 bp). The remaining sequences were aligned to the metagenome database. A sequence was assigned to a specific gene if at least half of the sequence was aligned to the coding region of the gene. For each gene, the cDNA sequences uniquely mapped to it, as well as those with multiple best hits that had been randomly assigned to it according to the bwa algorithm parameters, were then counted for each sample separately. Only the counts for uniquely mapped cDNA sequences were used to infer gene regulation patterns for most genes. For genes that were expressed at moderate to high levels, manual inspection of randomly selected genes revealed that the inclusion of the uncertain fraction of mRNA sequences (those that did not uniquely map to one gene) did not affect inferences concerning regulation patterns significantly because there were already large numbers of uniquely mapped cDNA sequences for these genes.

Simulation of SOLiD™ sequence alignments. In order to find the best alignment criteria for SOLiD™ datasets and in particular to determine the optimal number of mismatches allowed per sequence, a simulation was performed. Using the above-mentioned metagenome scaffolds as the data source, one million 50-bp sequences were generated using MetaSim (Richter et al., 2008). The composition of the test database was adjusted to reflect the relative proportions of sequences generated from *Chloroflexi*, *Cyanobacteria*, *Chlorobi*, *Acidobacteria* based upon the average rRNA sequence abundance (See Figure 1). A 50-bp empirical sequencing error model was adapted from a 62-bp error model downloaded from MetaSim webpage. The generated sequences were sequentially aligned to the metagenome scaffolds six times using bwa allowing 0, 1, 2, 3, 4 and 5 mismatches, respectively (more than five mismatches was too costly computationally). The alignment results were compared to determine both the organismal origin and actual sequence position of the generated test sequences. The number of sequences aligned incorrectly as unique best hits did not increase significantly as the number of mismatches allowed increased and were negligible (0.007%) compared to the number of sequences aligned correctly (Suppl. Table S3). The percentage of incorrectly mapped sequences is an upper limit, because the rRNA and repeat sequences were not removed prior to generating the test sequence dataset. These results clearly suggested that increasing the number of mismatches allowed to as many as 5 would not undermine the accuracy of the alignments. The simulations also showed that the number of sequences aligned began to saturate at 3 mismatches. However, the sequences in the actual datasets are derived from populations whose sequences can differ slightly from the consensus, which would cause a greater number of mismatches in their alignments to the consensus. Taking this small difference into account, 5 mismatches per sequences seemed to represent a reasonable choice for our analyses, ensuring alignment accuracy, a high level of correctly mapped sequences, a low error rate, and a reasonable computation time.

Statistical analyses. For each gene in both datasets, and for every possible pair-wise comparison of the four samples, a Fisher's exact test was conducted to determine the probability that the gene was differentially expressed in a statistically significant manner. In the 2×2 Fisher's exact test, the two columns were two different RNA samples (different timepoints); one row was the number of mRNA sequences assigned to the gene, and the other was the proportion of the total number of sequences (mapped sequences in SOLiD™ datasets). Ideally, the proportionality factor should be the percentage of cells of the species to which the gene belongs in the sample. However, because we did not have this information, the data were instead compared to the percentage of rRNA sequences assigned to the kingdom, to which the genome or scaffold that harbors any given gene belonged or was assigned. Percentages of rRNA sequences at the kingdom level should be proportional to the number of ribosomes in a given group of organisms and thus to the number of metabolically active cells; these values were the best data available to estimate with confidence the percentage of cells of different kingdoms in the community. No corrections for cell number, cell volume, and cellular physiological status were introduced. Use of this factor helped to offset the differences in cell abundances that could have originated during the sampling, and which might be significant in some cases. For each gene, the smallest p-value out of the six possible pairs was retained.

Normalization of expression level. For each gene that was differentially expressed in a statistically significant manner between at least a pair of samples ($p < 0.001$), a set of normalized expression levels were calculated for straightforward comparison purposes. Expression values E were first calculated for each sample using the following formula.

$$E_i = n_i / (N_i * p_i)$$

In the formula above, n denotes the number of mRNA sequences assigned to a gene in sample i; N denotes the number of total sequences (454 datasets) or total mapped sequences (SOLiD™ datasets), and p denotes the percentage of rRNA sequences of the kingdom to which the genome or scaffold that harbors the gene belong or assigned. We did not consider gene length, which certainly contributes to the total number of mRNA sequences mapped for a given gene, because this is not a factor when only comparing the expression of a given gene under different conditions. These four expression values were then normalized by the mean of the four values for the purpose of comparing expression patterns of genes with different transcription levels.

Figure S1. Composition of ribosomal RNA sequences (sunset sample) by kingdom. Only RNA sequences generated by pyrosequencing were analyzed. The numbers and the sizes of the circles next to the taxon names are proportional to the relative abundance of the rRNA sequences of that taxon. These figures were generated by MEGAN. (absolute score cutoff = 71, relative cutoff = 10% of best hit)

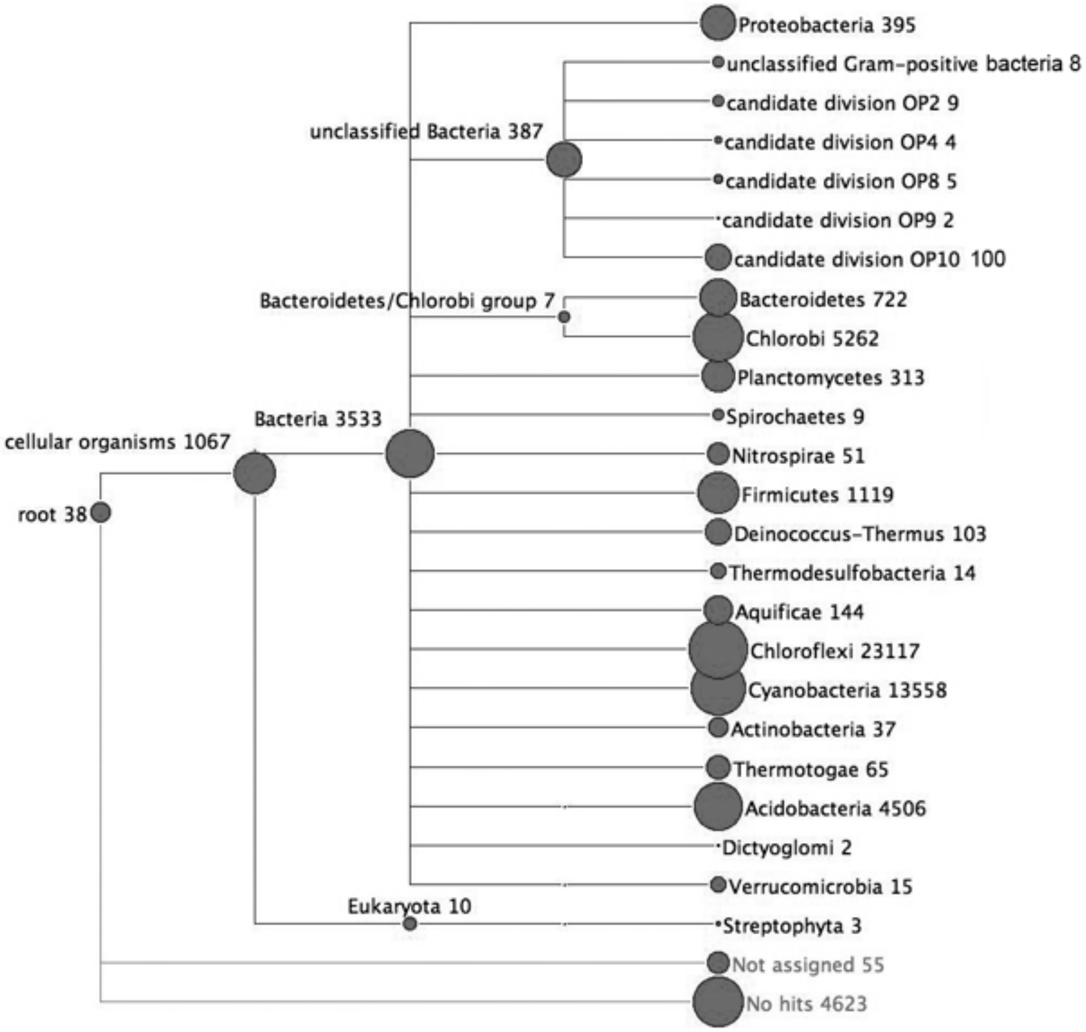

Figure S2. Composition of ribosomal RNA sequences of *Cyanobacteria* (A), *Chlorobi* (B & C) and *Acidobacteria* (D) in one sample (sunset). The analysis in Panel B was generated using the *Chlorobi* "zoom-in" 16S rRNA database while the analysis in Panel C was generated based on the same database with the addition of the 16S rRNA sequence recovered from the metagenome of an unclassified *Chlorobiales*-like organism from Mushroom Spring. Only rRNA sequences generated by pyrosequencing were analyzed. The sizes of the circles at the side of the taxon names is proportional to the relative abundance of the rRNA sequences assigned to that taxon. These figures were generated by MEGAN (absolute score cutoff = 71, relative cutoff = 0% of best hit).

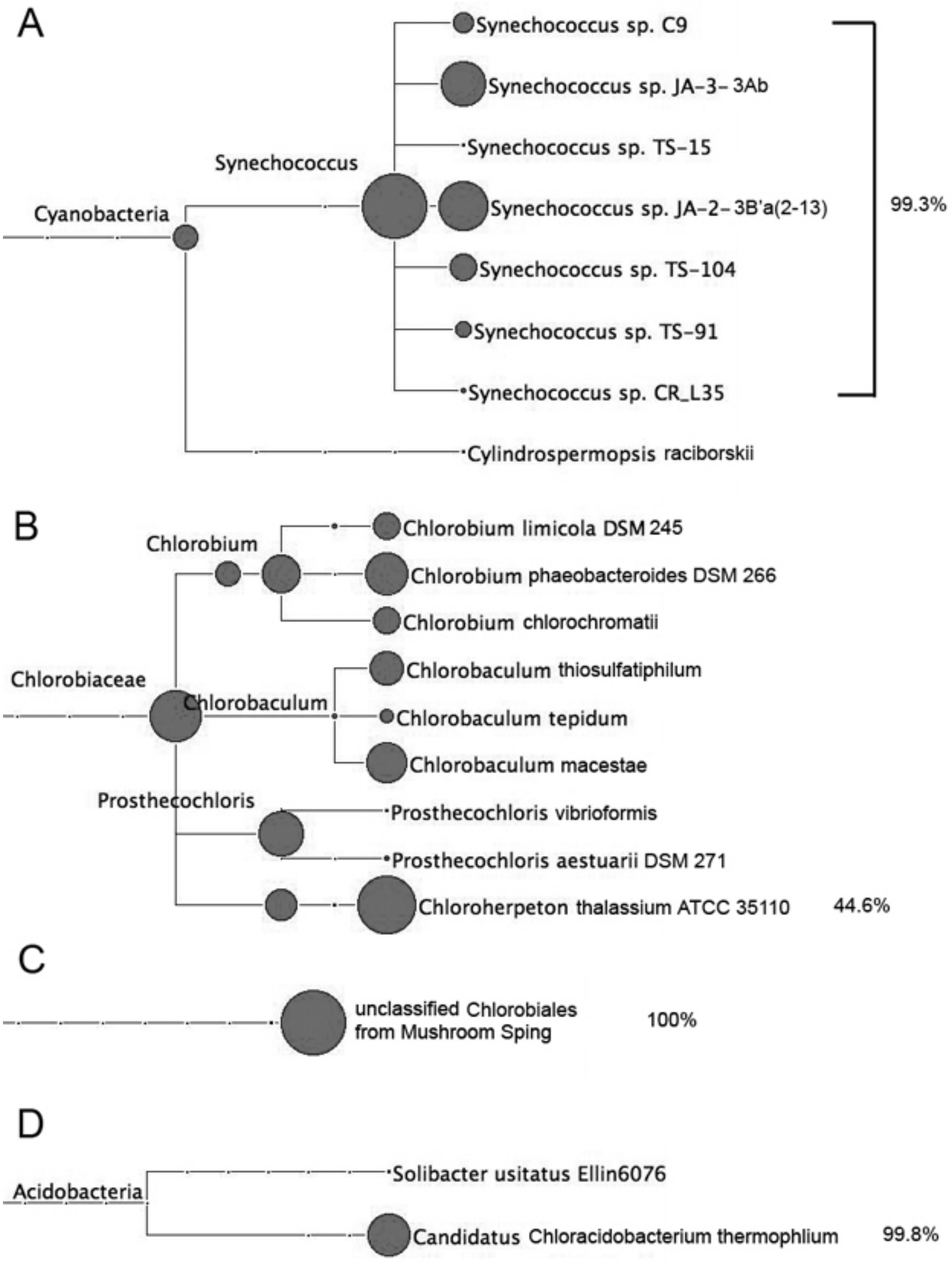
A
Cyanobacteria
Synechococcus
Synechococcus sp. C9
Synechococcus sp. JA-3-3Ab
Synechococcus sp. TS-15
Synechococcus sp. JA-2-3B'a(2-13)
Synechococcus sp. TS-104
Synechococcus sp. TS-91
Synechococcus sp. CR_L35
Cylindrospermopsis raciborskii
99.3%
B
Chlorobiaceae
Chlorobium
Chlorobium limicola DSM 245
Chlorobium phaeobacteroides DSM 266
Chlorobium chlorochromatii
Chlorobaculum
Chlorobaculum thiosulfatiphilum
Chlorobaculum tepidum
Chlorobaculum macestae
Prosthecochloris
Prosthecochloris vibrioformis
Prosthecochloris aestuarii DSM 271
Chloroherpeton thalassium ATCC 35110
44.6%
C
unclassified Chlorobiales
from Mushroom Sping
100%
D
Acidobacteria
Solibacter usitatus Ellin6076
Candidatus Chloracidobacterium thermophlium
99.8%

Table S1. Rank of selected photosynthesis-related genes among the top 20 genes with most mRNA sequences in both data sets.

Gene	Pyrosequencing		SOLiD™	
	No. of sequences[a]	Rank[b]	No. of sequences[a]	Rank[b]
Roseiflexus spp. *pufLM*	102	2	8,258	3
[c]Uncultivated *Chloroflexi* spp. *pufC*	14	5	4,091	2
[c]Uncultivated *Chloroflexi* spp. *pufL*	16	4	1,315	6
[c]Uncultivated *Chloroflexi* spp. *pucC*	3	>20	609	19
Chloroflexus spp. *bchH*	2	18	930	6
Synechococcus spp. *psaA*	183	1	3,232	3
Synechococcus spp. *psaB*	78	2	5,576	1
Chlorobiales spp. *pscA*	150	1	22,152	1
Chlorobiales spp. *fmoA*	141	2	16,993	2
Chlorobiales spp. *csmA*	21	7	3,781	6
Cand. Chloracidobacterium thermophilum *pscA*	84	1	14,216	1
Cand. Chloracidobacterium thermophilum *pscB*	19	6	1,925	12

[a] Sum of mRNA sequences for all four samples.

[b] Rank among genes with most mRNA sequences in corresponding phylogenetic groups.

[c] See Klatt et al. (2011) and Supplemental Materials and Methods for additional information.

Table S2. Selected examples of differentially transcribed orthologous genes of *Synechococcus* spp. A-like and B′-like[a]

	No. of sequences aligned to gene in A-like (Normalized relative expression level)				No. of sequences aligned to gene in B′-like (Normalized relative expression level)			
Annotation	Sunset	Sunrise	Low light morning	High light morning	Sunset	Sunrise	Low light morning	High light morning
universal stress protein family	14 (0.79)	5 (0.32)	13 (0.84)	36 (2.1)	461 (1.5)	201 (0.75)	281 (1.1)	201 (0.67)
universal stress protein family	5 (Not siginificant)	8	18	19	70 (1.6)	22 (0.56)	42 (1.1)	35 (0.80)
chaperonin, 10 kDa	7 (0.25)	18 (0.73)	40 (1.6)	38 (1.4)	58 (0.11)	289 (0.62)	910 (2.0)	666 (1.3)
Response regulator	43 (1.2)	5 (0.16)	31 (0.98)	59 (1.7)	2 (Not significant)	2	5	15
Response regulator	23 (1.0)	11 (0.53)	15 (0.74)	39 (1.7)	1 (Not significant)	3	5	4
Sensor histidine kinase	0 (0)	0 (0)	1 (0.3)	14 (3.7)	158 (0.61)	281 (1.2)	314 (1.4)	195 (0.77)
ubiquinone/menaquinone biosynthesis methyltransferase (MenG)	1 (Not significant)	1	0	4	49 (0.81)	30 (0.56)	78 (1.5)	69 (1.2)
4-hydroxybenzoate polyprenyl transferase (UbiA)	6 (0.37)	7 (0.48)	24 (1.7)	24 (1.5)	70 (0.40)	117 (0.74)	284 (1.8)	181 (1.0)
O-succinylbenzoate--CoA ligase (MenE)	0 (Not significant)	0	5	6	22 (0.43)	22 (0.48)	62 (1.4)	87 (1.7)
putative glycolate oxidase, iron-sulfur subunit	7 (0.46)	6 (0.44)	14 (1.0)	31 (2.1)	76 (0.47)	265 (1.9)	188 (1.3)	53 (0.33)
glycerate kinase	0 (0)	2 (0.27)	8 (1.1)	22 (2.6)	14 (0.11)	121 (1.1)	247 (2.3)	55 (0.46)

[a] Based on total mRNA sequences for the two strains, the ratio of *Synechococcus* spp. A-like to B′-like is approximately 1:2

Table S3. Simulation of 1 million 50-bp sequence alignments allowing different numbers of mismatches per sequence[a]

	Correctly aligned with unique best hit	Incorrectly aligned with unique best hit[b]	Aligned with multiple best hits[c]	Not aligned
0 Mismatches	383,274	28 (0.007%)	11,818 (3.08%)	604,880
1 Mismatches	683,232	44 (0.006%)	21,292 (3.12%)	295,432
2 Mismatches	798,149	62 (0.008%)	24,862 (2.11%)	176,927
3 Mismatches	827,133	62 (0.007%)	25,775 (3.12%)	147,030
4 Mismatches	832,652	63 (0.007%)	25,970 (3.12%)	141,315
5 Mismatches	833,675	64 (0.007%)	26,013 (3.12%)	140,248[d]

[a] The simulations were performed using MetaSim software (Richter et al., 2008) and using the metagenomic sequence database as the source of the sequences. The proportion of sequences for each kingdom (*Chloroflexi, Chlorobi, Cyanobacteria*, and *Acidobacteria*) was adjusted to reflect the average composition of the mats as inferred from the percentage of rRNA (see Figure 1).

[b] Percentage values in parentheses represent the alignment error rate among sequences aligned with unique best hits.

[c] Sequences with multiple best hits were not counted and were likely to result from repeated sequences and rRNAs, which were removed prior to alignment for the actual calculations.

[d] Most of the remaining unaligned sequences (99.8%) were deemed to be "unalignable" because they contained 6 or more Ns. These Ns were included in the metagenome scaffolds as spacers between contigs and scaffolds.

References

Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, et al. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucl Acids Res 35:D 169–172.

Huson DH, Auch AF, Qi J, Schuster SC (2007). MEGAN analysis of metagenomic data. Genome Res 17:377–386.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Heidelberg JF, Bhaya D, et al. (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. Accompanying paper.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl Acids Res 35:7188–7196.

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008): MetaSim—A sequencing simulator for genomics and metagenomics. PLoS ONE 3(10): e3373.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Molecular Biology 7:3.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3:e2527.

Ward DM, Cohan FM, Bhaya D, Heidelberg JF, Kühl M and Grossman AR (2008) Genomics, environmental genomics and the issue of microbial species. Nature Heredity 100:207--219