The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle

Eric A. Welsh*†, Michelle Liberton*†, Jana Stöckel*†, Thomas Loh*, Thanura Elvitigala*, Chunyan Wang‡, Aye Wollam‡, Robert S. Fulton‡, Sandra W. Clifton‡, Jon M. Jacobs§, Rajeev Aurora¶, Bijoy K. Ghosh*, Louis A. Sherman‖, Richard D. Smith§, Richard K. Wilson‡, and Himadri B. Pakrasi*,**

*Department of Biology, Washington University, St. Louis, MO 63130; ‡Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108; §Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352; ¶Department of Molecular Microbiology and Immunology, Saint Louis University School of Medicine, St. Louis, MO 63104; and ‖Department of Biological Sciences, Purdue University, West Lafayette, IN 47907

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved August 6, 2008 (received for review June 3, 2008)

Unicellular cyanobacteria have recently been recognized for their contributions to nitrogen fixation in marine environments, a function previously thought to be filled mainly by filamentous cyanobacteria such as *Trichodesmium*. To begin a systems level analysis of the physiology of the unicellular N_2-fixing microbes, we have sequenced to completion the genome of *Cyanothece* sp. ATCC 51142, the first such organism. *Cyanothece* 51142 performs oxygenic photosynthesis and nitrogen fixation, separating these two incompatible processes temporally within the same cell, while concomitantly accumulating metabolic products in inclusion bodies that are later mobilized as part of a robust diurnal cycle. The 5,460,377-bp *Cyanothece* 51142 genome has a unique arrangement of one large circular chromosome, four small plasmids, and one linear chromosome, the first report of a linear element in the genome of a photosynthetic bacterium. On the 429,701-bp linear chromosome is a cluster of genes for enzymes involved in pyruvate metabolism, suggesting an important role for the linear chromosome in fermentative processes. The annotation of the genome was significantly aided by simultaneous global proteomic studies of this organism. Compared with other nitrogen-fixing cyanobacteria, *Cyanothece* 51142 contains the largest intact contiguous cluster of nitrogen fixation-related genes. We discuss the implications of such an organization on the regulation of nitrogen fixation. The genome sequence provides important information regarding the ability of *Cyanothece* 51142 to accomplish metabolic compartmentalization and energy storage, as well as how a unicellular bacterium balances multiple, often incompatible, processes in a single cell.

diurnal rhythm | linear chromosome | nitrogen fixation | optical mapping | proteomics

Although dinitrogen (N_2) is the most plentiful gas in the atmosphere, it is inert and therefore not readily available for use in biological systems. The introduction of new nitrogen via nitrogen fixation has a central role in marine environments where the bioavailability of nitrogen determines the level of primary productivity (1). Many strains of cyanobacteria have the ability to fix atmospheric nitrogen to a biologically accessible form and therefore significantly contribute to the oceanic biological nitrogen cycle (2, 3). Nitrogen fixation is catalyzed by nitrogenase, a multiprotein enzyme exquisitely sensitive to oxygen. Cyanobacteria are the only diazotrophic organisms that produce oxygen gas as a by-product of photosynthesis and, consequently, must reconcile the presence of oxygen with the activity of an oxygen-sensitive enzyme. A variety of mechanisms have evolved to accommodate these incompatible processes, including the formation of heterocysts, specialized cells for nitrogen fixation in filamentous strains such as *Anabaena* (4–6), temporal separation of photosynthesis and nitrogen fixation in unicellular organisms such as *Crocosphaera* and *Cyanothece*, and complex combinations of spatial and temporal separation in nonheterocystous filamentous cyanobacteria such as *Trichodesmium* (7).

In marine environments, earlier work identified *Trichodesmium* as the key oceanic nitrogen-fixing cyanobacterium (8). More recently, Zehr and coworkers determined that multiple strains of unicellular cyanobacteria play a significant role in oceanic nitrogen fixation (3), leading to a new interpretation of N_2 fixation in this ecosystem. In fact, the abundance of unicellular diazotrophs may make their contribution to marine N_2 fixation more significant than that of *Trichodesmium* (1). Despite their importance in the nitrogen dynamics of oceanic environments, a genome level description of a unicellular diazotrophic marine cyanobacterium has not previously been reported.

Cyanothece strains have been isolated worldwide from various habitats, including salt waters, where they demonstrate significant diversity in their cellular size, shape, and growth rates (9). One marine diazotrophic strain, *Cyanothece* sp. ATCC 51142, has a robust diurnal cycle in which photosynthesis is performed during the day and nitrogen fixation at night (10). As part of this diurnal cycle, *Cyanothece* 51142 actively accumulates and subsequently degrades and utilizes large quantities of different storage inclusion bodies, including those for the products of photosynthesis and nitrogen fixation (10) (see Fig. 1). Details of the diurnal cycle in *Cyanothece* 51142 have been explored by recent global transcriptional analyses, which revealed that 30% of genes, essentially those involved in central metabolic pathways, showed strong cyclic expression patterns (11). Furthermore, functionally related genes were maximally expressed at distinct phases during the diurnal period. Taken together, these physiological traits describe an organism in which tight control of cellular processes combined with storage of metabolic products for later usage is paramount for its ecological success and survival.

Author contributions: M.L., J.S., R.A., B.K.G., L.A.S., R.D.S., R.K.W., and H.B.P. designed research; M.L., J.S., C.W., A.W., R.S.F., S.W.C., and R.K.W. performed research; C.W., J.M.J., and R.D.S. contributed new reagents/analytic tools; E.A.W., M.L., J.S., T.L., T.E., R.S.F., S.W.C., and J.M.J. analyzed data; and E.A.W., M.L., and J.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The genome sequence reported in this paper has been deposited in GenBank (accession nos. CP000806–CP000811).

†E.A.W., M.L., and J.S. contributed equally to this work.

**To whom correspondence should be addressed at: Department of Biology, Washington University, One Brookings Drive CB1137, St. Louis, MO 63130. E-mail: pakrasi@wustl.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0805418105/DCSupplemental.

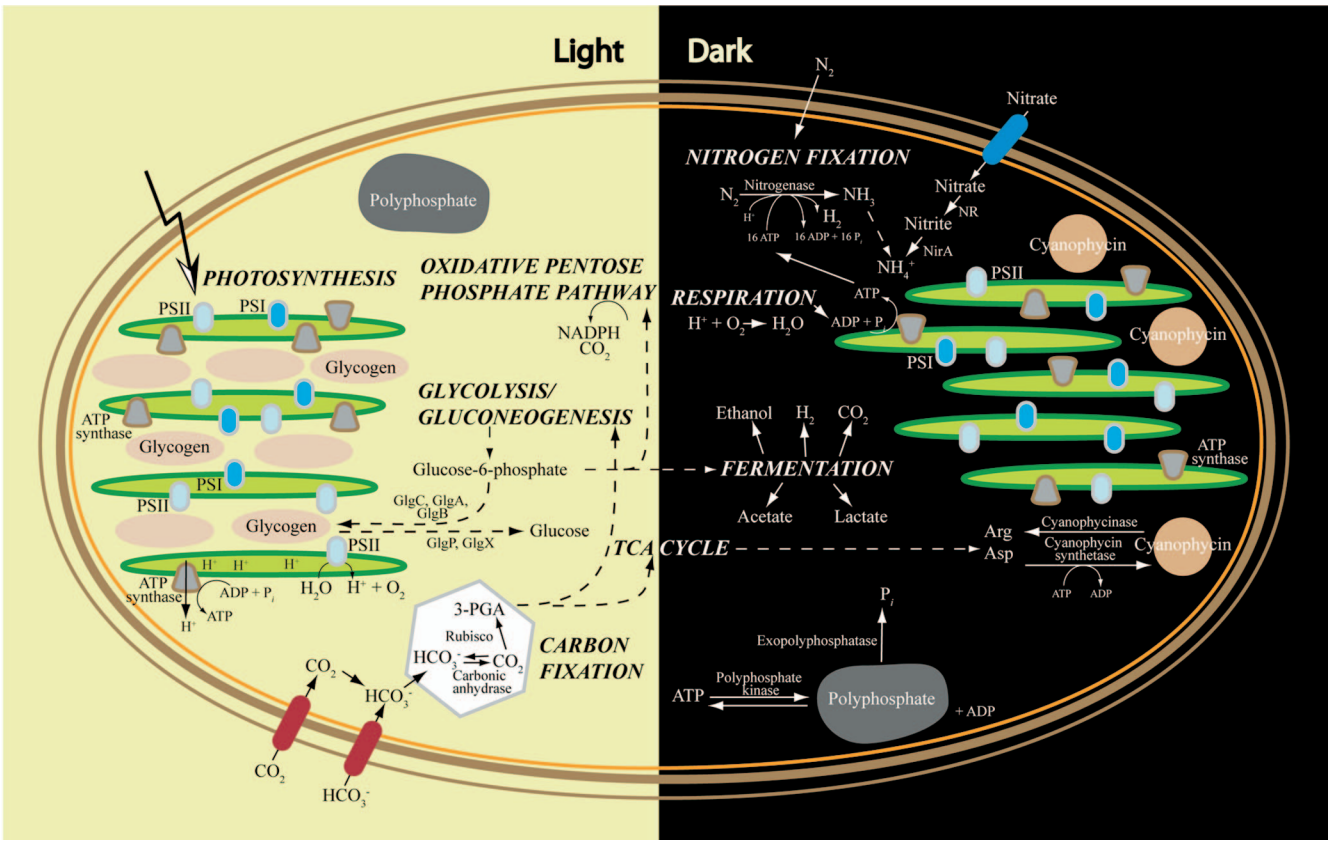

Fig. 1. Overview of processes involved in daily metabolic cycling in *Cyanothece* 51142. Photosynthesis fixes carbon during the day, which is stored in glycogen granules. Glycogen is rapidly consumed during a burst of respiration in the early dark period, which coincides with peak nitrogenase activity, fermentation, and a minimum of photosynthetic capacity (10). Fixed nitrogen is stored in cyanophycin granules, which are completely depleted during the following day. Phosphate is stored in polyphosphate bodies.

To further understand the genomic basis for the biochemical and physiological adaptations in unicellular diazotrophic cyanobacteria, we report the complete genome sequence of *Cyanothece* sp. ATCC 51142. Our analysis of this genome revealed a unique combination of highly conserved gene clusters involved in nitrogen fixation and glucose metabolism, as well as the entire set of genes required for heterolactic acid and acetate fermentation. Furthermore, the pathways relevant for storage of metabolic products, including nitrogen (cyanophycin), carbon (glycogen), and phosphorus (polyphosphate), in intracellular inclusion bodies are represented by multicopy genes that may provide redundancy or additional regulation. These genome-based insights highlight the overall strategy of *Cyanothece* 51142 to maximize efficiency by tightly regulating major metabolic processes throughout a diurnal cycle.

Results and Discussion

Organization of the *Cyanothece* 51142 Genome. Cyanobacterial genomes reported to date vary greatly in size, a diversity likely related to the capability to adapt to different environmental niches. The 1.6-Mb genome of unicellular *Prochlorococcus marinus* str. CCMP1986 (MED4) may represent one of the smallest sets of genes enabling survival in the open ocean (12), whereas genome expansion, as in freshwater heterocystous *Anabaena* sp. PCC 7120 (7.2 Mb) (13) and marine unicellular *Acaryochloris marina* (6.5 Mb) (14), may illustrate the ability to acquire novel metabolic traits to succeed in specific environments. Interestingly, the genome of *Cyanothece* 51142, at 5.46 Mb in size, is ≈35% larger than that of the closely related *Synechocystis* sp. PCC 6803, an increase in size possibly accounting for temporal regulation and nitrogen fixation. This 5.46-Mb genome size is smaller than that of *Anabaena* and *Nostoc*, consistent with a correlation between heterocystous filamentous lifestyles and genome expansion.

Despite differences in genome sizes, the genome organizations of all cyanobacteria reported to date have been similar, with one large circular chromosome and several smaller circular plasmids (15). However, shotgun sequencing and finishing of the *Cyanothece* 51142 genome led to the possibility of an independent 430-kb linear element. We then used optical restriction mapping, a technique that has been successfully used in bacterial genome finishing (16), as an independent approach to confirm that the 430-kb element is indeed linear [supporting information (SI) Fig. S1]. Verified by these combined approaches, the completed *Cyanothece* 51142 genome consists of six separate elements: a 4.93-Mb circular chromosome, four plasmids ranging in size from 10 kb to 40 kb, and the 430-kb linear chromosome (Fig. 2 and Table 1).

The most unique feature of the organization of the *Cyanothece* 51142 genome is the linear chromosome, which is 429,701 bp long, and contains 449 predicted protein-coding sequences, 127 (28.3%) with assigned function (Table 1). Overall, this chromosome contains a much higher percentage of genes with no assigned function (71.7% vs. 45.7%) compared with the large circular chromosome. Of the 95 genes on the linear chromosome assigned to functional categories, 42 are unique to the linear chromosome and have homologs in other organisms, whereas 38 genes have a corresponding copy on the circular chromosome or plasmids (Table S1).

Although linear genomic elements have independently

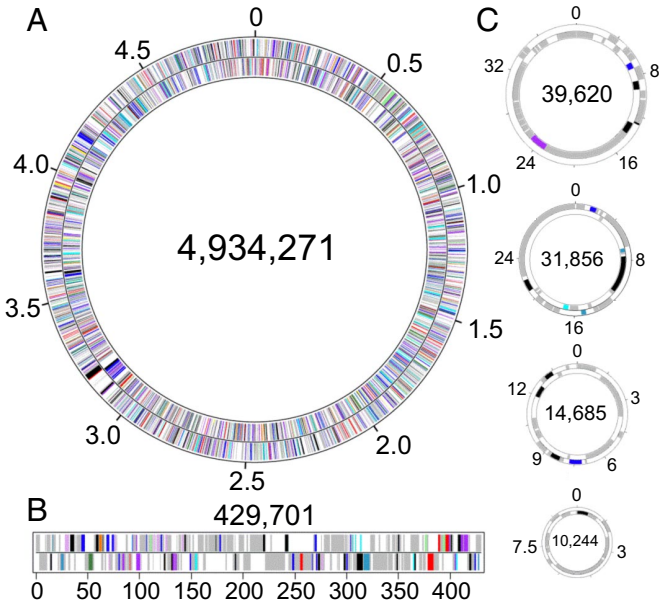

Fig. 2. General genome features and distribution of gene functions within the *Cyanothece* 51142 genome. Labels are given in base pairs for the circular chromosome (*A*), linear chromosome (*B*), and plasmids (*C*). Genes are colored by functional category as follows: energy, fatty acid, and phospholipid metabolism (red), cell envelope (orange), cellular processes (steel blue), central intermediary metabolism (light green), photosynthesis and respiration (dark green), regulation (cyan), DNA, transcription, and translation (dark blue), small molecule biosynthesis (magenta), transport and binding (purple), unknown/hypothetical (gray), other (black), and noncoding RNA (yellow). The two rRNA operons on the circular chromosome are indicated in yellow at 3.95 Mb and 4.10 Mb in *A*. The glucose and pyruvate metabolism cluster on the linear chromosome is shown in red and light green between 375 kb and 400 kb in *B*.

evolved in other bacterial genera such as *Borrelia*, *Streptomyces*, and *Agrobacterium*, mechanistic details concerning the origin, maintenance, and selective advantage of linear chromosomes are poorly understood (17, 18). Analysis of the *Cyanothece* 51142 linear chromosome failed to reveal the presence of any distinguishing feature near its ends, such as inverted repeats and stem-loop structures (17), nor the presence of any telomere-processing proteins known to be associated with linear chromosomes in characterized bacterial genera (19, 20). The origin of replication could not be determined for either the circular or linear chromosome by using standard GC skew and DnaA box analysis (21, 22), consistent with the behavior of these algorithms on related cyanobacteria (22, 23). Thus, the mechanism for replication and maintenance of the linear chromosome in *Cyanothece* 51142 remains unknown.

Proteomics-Assisted Genome Annotation. The total number of predicted protein-coding genes in the finished *Cyanothece* 51142 genome is 5,304, and for 2,735 (51.6%) of these, a likely function could be assigned. Of the remainder, 506 (9.5%) are of unknown function and 2,063 (38.9%) are hypothetical (Table 1). The annotation of genes of unknown function was greatly aided by data from a high-throughput proteomics analysis (see *Materials and Methods* for details). Proteomic data were used, in conjunction with early draft sequence, to build an accurate mass and time (AMT) tag library (24) that covered ≈50% of the predicted proteome. Because of the observation of corresponding tryptic peptides, 506 (25%) of 1,989 predicted "hypothetical" proteins with no significant homology to those of known function were reclassified as "unknowns". Additionally, the observed tryptic peptides were matched against a set of ≈12,000 low-confidence ORF predictions, resulting in the inclusion of 38 additional ORFs in the final genome annotation. The combined analysis of proteome and genome data is an important new approach that resulted in the inclusion or reclassification of nearly 550 genes (10%) and lent an additional and valuable level of validation to the genome annotation.

Genome-Based Model of *Cyanothece 51142*. The complete genome information has led to a detailed picture of the metabolic capacity of *Cyanothece* 51142 (Fig. 1). In the following sections, we will highlight key processes that are important in the life cycle of this organism.

Nitrogen Fixation. Nitrogen fixation has great influence on cellular metabolism in *Cyanothece* 51142 because the nitrogenase enzyme is synthesized, functions, and is then degraded during each diurnal period (10), a process that consumes considerable cellular resources. Phylogenetic analysis showed that *Cyanothece* 51142 and *Crocosphaera watsonii*, both capable of N_2 fixation, are most closely related to *Synechocystis* sp. PCC 6803, a unicellular cyanobacterium that cannot fix nitrogen (Fig. S2). Although the origin of nitrogen fixation in cyanobacteria is currently a topic of much debate, this phylogeny suggests that nitrogen fixation may have been present in the ancestor of *Crocosphaera watsonii*, *Cyanothece* 51142, and *Synechocystis* 6803, and was later lost in *Synechocystis* 6803, a view consistent with Swingley *et al.* (25).

In the *Cyanothece* 51142 genome, the majority of genes involved in nitrogen fixation are located in a contiguous 28-kb cluster of 34 genes separated by no more than 3 kb (11, 26). Within the gene cluster are the structural genes encoding the Fe-protein (*nifH*) and Mo-Fe protein (*nifDK*) of the molybdenum-iron nitrogenase enzyme, genes involved in Mo-Fe cofactor biosynthesis (*nifB*, *fdxN*, *nifS*, *nifU*, and *nifV*), Mo-Fe cofactor assembly (*nifE* and *nifN*), iron uptake (*feoA* and *feoB*), and genes of unknown function. The proximity of these genes to each other, combined with their highly synchronous expression (11, 26), strongly suggests that all genes in this cluster are involved in nitrogen fixation. The organization of genes in operons or

Table 1. General features of the *Cyanothece* 51142 genome

	Total	Percent	Circular	Percent	Linear	Percent	pCT42a	Percent	pCT42b	Percent	pCT42c	Percent	pCT42d	Percent
Accession No.			CP000806		CP000807		CP000808		CP000809		CP000810		CP000811	
Size, bp	5,460,377		4,934,271		429,701		39,620		31,856		14,685		10,244	
G + C content, %	37.9		37.9		38.6		36.8		41.5		38.1		37.0	
Open reading frames	5,304	100.0	4,762	100.0	449	100.0	38	100.0	25	100.0	19	100.0	11	100.0
Assigned function	2,735	51.6	2,584	54.3	127	28.3	10	26.3	7	28.0	5	26.3	2	18.2
Unknown	506	9.5	468	9.8	24	5.3	5	13.2	5	20.0	4	21.1	0	0.0
Hypothetical	2,063	38.9	1,710	35.9	298	66.4	23	60.5	13	52.0	10	52.6	9	81.8
Ribosomal RNAs	2		2		—		—		—		—		—	
Transfer RNAs	43		43		—		—		—		—		—	

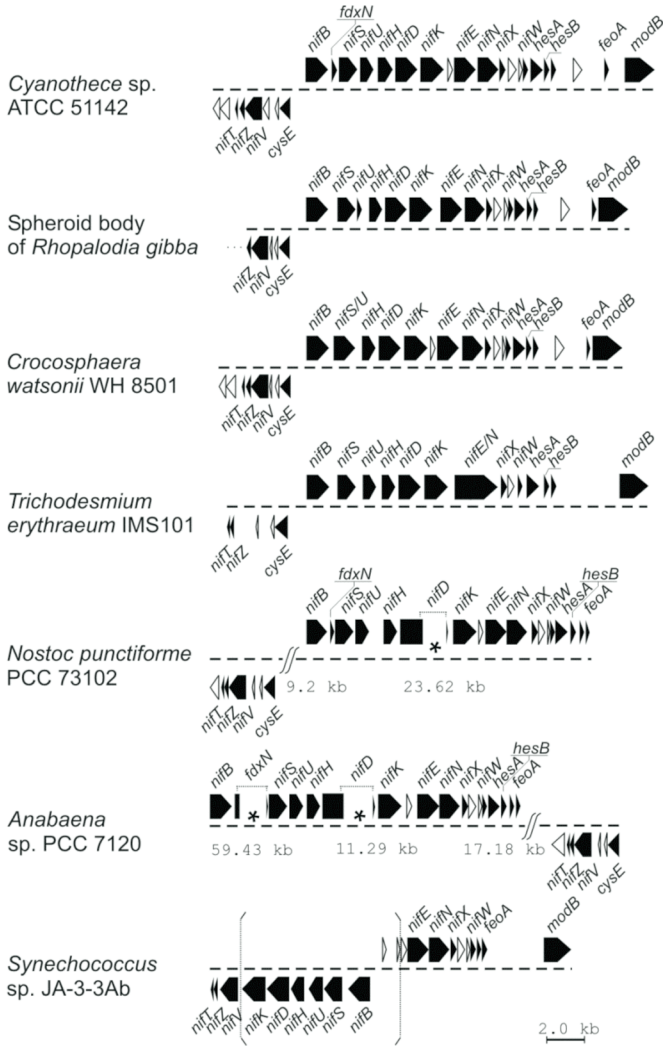

Fig. 3. Clusters of N_2-fixation-related genes. Shown are genes with conserved synteny between *Cyanothece* 51142 and other nitrogen-fixing cyanobacteria. Black arrows represent genes assigned to functional categories and white arrows correspond to hypothetical genes and genes of unknown function. Missing sequence information from the spheroid body of *Rhopalodia gibba* is indicated by dots. (∫∫) denotes gaps in the sequence, with the length of the omitted sequence in kb. (*) indicates the location of DNA insertion elements, with the size of the element in kb. A possible inversion event in *Synechococcus* sp. JA-3–3Ab is highlighted in brackets. GenBank accession numbers for the sequences used are as follows: *Cyanothece* sp. ATCC 51142, CP000806; spheroid body of *Rhopalodia gibba*, AY728387; *Crocosphaera watsonii* WH 8501, AADV02000024; *Trichodesmium erythraeum* IMS101, CP000393; *Nostoc punctiforme* PCC 73102, CP001037; *Anabaena* sp. PCC 7120, BA000019; and *Synechococcus* sp. JA-3–3Ab, CP000239.

transcriptional regulons is an attribute of prokaryotes that allows efficient control of genes participating in the same process. It may also provide the potential for lateral multigene transfer in a single step. Both of these factors may be at play regarding the *nif* gene clusters in cyanobacteria.

We compared the *nif* gene cluster in *Cyanothece* 51142 to those found in a group of representative nitrogen-fixing cyanobacteria and to the cyanobacterial endosymbiont of the fresh-water diatom *Rhopalodia gibba*, which has a *nif* locus closely related to that of *Cyanothece* 51142 (27). From this comparison, we found that *Cyanothece* 51142 contains the largest contiguous cluster of nitrogen fixation-related genes in this group (Fig. 3 and Table S2). Importantly, the organization of the cluster is highly conserved among the nonheterocyst forming cyanobacteria: a single cluster consisting of two adjacent regulons on opposite strands, with the gene for a molybdate ATP-binding cassette transporter permease, *modB*, at the end. This general organization was found in all of the nonheterocystous strains examined, including *Cyanothece* 51142, *Crocosphaera watsonii* WH8501, *Trichodesmium erythraeum* IMS101, *Synechococcus sp. JA-3-3Ab*, and also in the spheroid body of *Rhopalodia gibba* (Fig. 3). This is notable because transcriptional regulation is likely triggered by different mechanisms due to the diverse strategies for nitrogen fixation used by these organisms. In contrast, the corresponding clusters in the heterocystous cyanobacteria *Nostoc punctiforme* PCC 73102 and *Anabaena* sp. PCC 7120 are disrupted by several insertion elements and are missing *modB*, which is located elsewhere in the genomes. However, these clusters still share the same gene order within the separate subclusters. The cluster in *Synechococcus* sp. JA-3–3Ab, the most anciently branching of the group, is missing several genes and contains a unique inversion between *nifV* and *nifE*. Together, the organization of these clusters suggests a divergence from a common ancestor. The pattern of these differences between clusters can be most parsimoniously explained by starting with a single *Cyanothece* 51142-like ancestral *nif* cluster and introducing various inversion, deletion, mutation, duplication, and translocation events to yield the organizations found in the other strains. Interestingly, the similarities between the gene clusters resemble the relationships shown in the phylogenetic tree (Fig. S2), even though nitrogen fixation-related sequences were excluded from the tree-building process (see *SI Text*).

Energy Metabolism and Fermentation. *Cyanothece* 51142 cells are programmed to undergo diurnal cycles of glycogen synthesis in the light, followed by degradation and utilization in the subsequent dark period (Fig. 1) (28). Accordingly, the genome includes the genes for glycogen synthesis and three different genes for glycogen degradation, with *glgP2* located on the linear chromosome (Fig. 4*A*, Reaction 2). *Cyanothece* 51142 contains all of the fermentation-related genes necessary to produce ethanol, lactate, acetate, and hydrogen (Fig. 4*A*), processes that require an anoxic intracellular environment similar to that created during the dark period when nitrogen fixation occurs. Interestingly, while genes in the *Cyanothece* 51142 genome related to carbohydrate and energy metabolism are localized at multiple loci on the circular chromosome, a 20.2-kb gene cluster on the linear chromosome contains several genes involved in glucose and pyruvate metabolism (Fig. 4*B*). Within this cluster is the only gene encoding L-lactate dehydrogenase (*ldh*) in the genome. This enzyme is required for the terminal step of lactate fermentation, suggesting that the linear chromosome is important for this process. Clustering of these genes may provide an advantage in transcriptional regulation under conditions where fermentation becomes the major energy-deriving process. This gene cluster on the linear chromosome does not show any conserved synteny to other cyanobacterial strains or to any genomes in the Kyoto Encyclopedia of Genes and Genomes (29) pathway database and is therefore unique to *Cyanothece* 51142.

Notably, *Cyanothece* 51142 contains a gene encoding the phosphoenolpyruvate carboxykinase (PEPCK), an enzyme that performs the first step in gluconeogenesis, circumventing the irreversible reaction of phosphoenolpyruvate to pyruvate in glycolysis. This enzyme links carbon, organic acid, and amino acid metabolism. This gene is also present in the purple photosynthetic bacterium *Rhodopseudomonas palustris* (30) and in the partially sequenced *Cyanothece* CCY0110 but is missing in nearly all other cyanobacterial strains sequenced to date.

Linear Chromosome. Linear chromosomes have been identified in a number of bacterial species, now including a photosynthetic

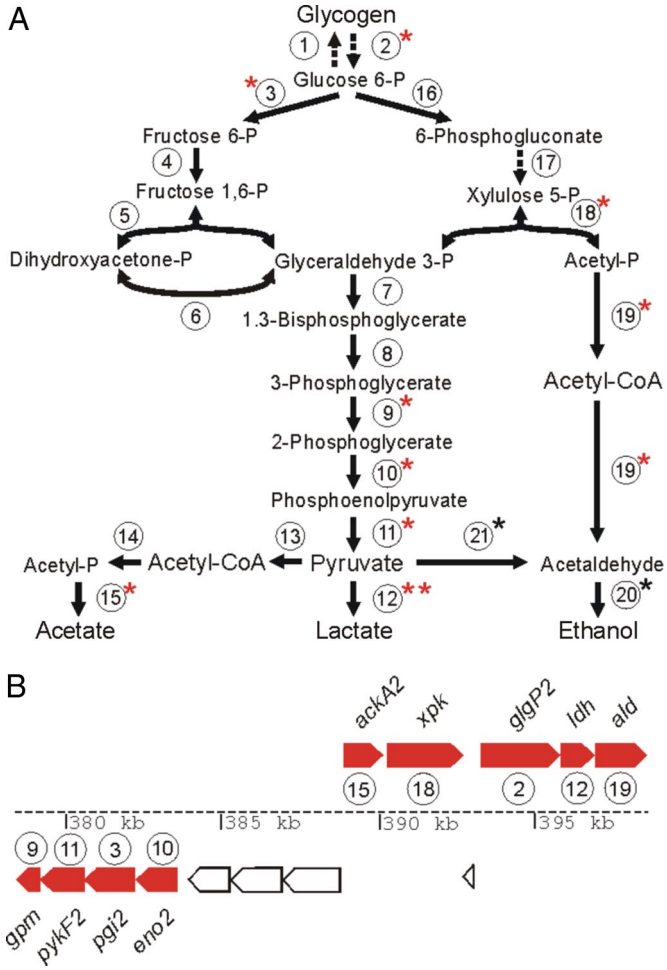

Fig. 4. Genes encoding enzymes in glucose metabolism pathways in *Cyanothece* 51142. Pathways were generated by mapping *Cyanothece* 51142 genes onto known fermentative pathways (29, 40). (*A*) Each arrow shows the direction of the reaction. Broken arrows indicate that more than one catalytic step is involved. The numbers correspond to the enzymes involved: (1) enzymes for glycogen synthesis, (2) enzymes for glycogen degradation, (3) glucose 6-P isomerase *pgi*, (4) 6-phosphofructokinase *pfkA*, (5) fructose-bisphosphate aldolase *fba*, (6) triosephosphate isomerase *tpi*, (7) glyceraldehyde-3-P dehydrogenase *gap*, (8) phosphoglycerate kinase *pgk*, (9) phosphoglycerate mutase *gpm*, (10) enolase *eno*, (11) pyruvate kinase *pykF*, (12) lactate dehydrogenase *ldh*, (13) pyruvate dehydrogenase *pdhA*, (14) phosphotransacetylase *pta*, (15) acetate kinase *ackA*, (16) glucose-6-P dehydrogenase *zwf*, (17) 6-phosphogluconate dehydrogenase *gnd*, (18) xylulose-5-P phosphoketolase *xpk*, (19) aldehyde dehydrogenase *ald*, (20) alcohol dehydrogenase *adh*, and (21) pyruvate decarboxylase *pdc*. One star illustrates enzymes with a gene copy present on the linear chromosome. Two stars signify uniqueness to the linear chromosome. Red stars correspond to genes organized in the gene cluster shown in *B*. (*B*) Cluster of genes on the linear chromosome that encodes various enzymes involved in glucose metabolism.

cyanobacterium. However, the implications of the presence of such a chromosome in any of these bacteria remain unresolved (31). An important question concerning the linear chromosome of *Cyanothece* 51142 is whether the genes located on it are expressed. Global transcriptional analyses have revealed that only a small number of genes on the linear chromosome are within the group of significantly cycling genes (11). In fact, of a total of 1,445 cycling genes, only 17 are found on the linear chromosome (Fig. S3). However, from the data generated as part of a global proteomics analysis (described above), we identified the products of 50 of the protein-coding genes on the linear chromosome (Table S1).

Analysis of the genes on the linear chromosome that have corresponding copies elsewhere in the *Cyanothece* 51142 genome did not uncover any significant conserved synteny, with the exception of the *coxBAC* operon that is found twice on the circular chromosome. Although linear chromosomes have not been reported in photosynthetic bacteria, it is possible that they are specifically confined to the *Cyanothece* genus. We compared the linear chromosome and the genome sequences of other bacteria and found synteny only between the *Cyanothece* 51142 linear chromosome and the partially sequenced genome of *Cyanothece* sp. CCY0110. The draft genome sequence of *Cyanothece* sp. CCY0110 is currently in 163 contigs, and we identified 7 contigs covering a region of 182,835 bp that have synteny to the *Cyanothece* 51142 linear chromosome (Fig. S4). Given the unfinished status of the *Cyanothece* sp. CCY0110 genome, the existence of a linear element cannot be unequivocally ascertained, but it remains possible that this genome also contains a linear chromosome. Additional genome sequences of other *Cyanothece* strains will be necessary to determine the extent of this occurrence within the *Cyanothece*.

A higher evolution rate for genes located near the chromosome ends has been reported for the linear chromosome in *Streptomyces*. This is particularly interesting, because most of the multicopy genes on the linear chromosome in *Cyanothece* 51142 are indeed localized closer to the ends. Gene redundancy could lead to alterations and selection for modifications in one gene copy that are beneficial to the organism.

Conclusion

To date, more than 50 cyanobacterial genomes, including freshwater, benthic, and open ocean strains, have been at least partially sequenced (14). As the first completely assembled genome of a unicellular nitrogen-fixing cyanobacterium, the genome of *Cyanothece* 51142 represents a new class because, in addition to a conventional circular chromosome and plasmids, it contains one linear chromosome, the first description of a linear element in the genome of a photosynthetic bacterium. Additionally, marine diazotrophs such as *Cyanothece* 51142 need to accommodate a variety of environmental changes, particularly in nutrient availability. These organisms can adjust their cellular machinery according to nitrogen availability, which involves the capability to synthesize, store, degrade, and use large quantities of storage products during a diurnal period. This adaptation of *Cyanothece* 51142 to a nitrogen-deficient marine environment is reflected at the genomic level. Notably, the genes associated with nitrogen fixation and glycogen metabolism are organized in regulons and/or found as multicopy genes. This presumably presents a mechanism for efficient control of gene expression and, together with the capability to store metabolic products in inclusion bodies, provides an advantage for unicellular marine diazotrophs such as *Cyanothece* 51142 in their natural nutrient-poor habitats.

Materials and Methods

Cell Growth. Cells were routinely grown under 12-h light/dark conditions (50 μmol photons $m^{-2}s^{-1}$ at 30°C) in ASP2 (11) medium without added nitrate ($NaNO_3$).

Genome Sequencing. At the Washington University Genome Sequencing Center, an Applied Biosystem 3730 instrument was used to sequence an initial 9.3× coverage shotgun library, followed by sequencing of a fosmid library (EpiFos, Epicentre) and several rounds of pyrosequencing on a Roche 454 sequencer.

Genome Annotation. ORFs were predicted by using a combination of CRITICA (32) and GLIMMER (33), followed by detailed manual annotation. ScalaBLAST (34) hits were generated vs. UNIPROT release 7.7 (35) and the proteomes of all then-available sequenced cyanobacteria. GenomePlot software (36) was used to render the genomic maps. All tRNAs were assigned by using tRNAscan-SE (37). Additional ncRNAs were assigned by using INFERNAL (38) to search for

RFAM families (39) found in other cyanobacteria. Ribosomal rRNAs were identified through sequence homology to other cyanobacterial strains.

Proteomic Analysis. Soluble and membrane fractions were subjected to strong cation exchange chromatography (SCX), followed by reversed phase liquid chromatography (RPLC) separation and MS/MS analysis of peptides on a Finnigan LTQ ion trap mass spectrometer (ThermoFinnigan). Each unfractionated and SCX fraction was analyzed via capillary RPLC-MS/MS. SEQUEST software was used to match the MS/MS fragmentation spectra to sequences from the initial draft *Cyanothece* 51142 proteome. An AMT tag database containing the calculated mass and normalized elution time for each identified peptide was generated to assist with subsequent high sensitivity, high-throughput analysis of *Cyanothece* 51142 samples by using the AMT tag approach (24).

Analysis of the *nif* Cluster. Proteins encoded by the *Cyanothece* 51142 *nif* cluster were searched against other cyanobacterial genomes by using BLASTP and TBLASTN, with an E-value cutoff of 0.01. Hits within 10 kb of each other within each genome were identified as part of a cluster of Nif proteins. Synteny to the *Cyanothece* 51142 cluster was identified manually and homologous genes assigned. Of the 34 genes in the *Cyanothece* 51142 cluster, 6 are not well conserved across the examined strains and were omitted from the analysis (Fig. 3 and Table S2).

ACKNOWLEDGMENTS. We thank all members of the Pakrasi Lab for collegial discussions and active participation in the manual annotation of the *Cyanothece* 51142 genome. This work was supported by the Danforth Foundation at Washington University, the Department of Energy–Basic Energy Science program (H.B.P.), and the National Science Foundation-Frontiers in Integrative Biological Research program (H.B.P., R.A., and B.G). This work is also part of a Membrane Biology Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Science Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research program (Pacific Northwest National Laboratory).

1. Montoya JP, *et al.* (2004) High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* 430:1027–1032.
2. Bryant DA, Frigaard NU (2006) Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* 14:488–496.
3. Zehr JP, *et al.* (2001) Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* 412:635–638.
4. Haselkorn R (1978) Heterocysts. *Annu Rev Plant Physiol* 29:319–344.
5. Wolk CP (2000) Heterocyst formation. *Anabaena in Prokaryotic Development*, eds Brun YV, Shimkets L (Am Soc Microbiol, Washington, DC), pp 83–104.
6. Haselkorn R (2007) Heterocyst differentiation and nitrogen fixation in cyanobacteria. *Associative and Endophytic Nitrogen-Fixing Bacteria and Cyanobacterial Associations*, eds Elmerich C, Newton WE (Springer, Dordrecht), pp 233–255.
7. Berman-Frank I, Lundgren P, Falkowski P (2003) Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res Microbiol* 154:157–164.
8. Lipschultz F, Owens NJP (1996) An assessment of nitrogen fixation as a source of nitrogen to the North Atlantic Ocean. *Biogeochemistry* 35:261–274.
9. Rippka R (1988) Isolation and purification of cyanobacteria. *Methods Enzymol* 167:3–27.
10. Sherman LA, Meunier P, Colon-Lopez MS (1998) Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth Res* 58:25–42.
11. Stöckel J, *et al.* (2008) Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Natl Acad Sci USA* 105:6156–6161.
12. Dufresne A, *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* 100:10020–10025.
13. Kaneko T, *et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* 8:205–213; 227–253.
14. Swingley WD, *et al.* (2008) Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci USA* 105:2005–2010.
15. Kulikova T, *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 32:D27–D30.
16. Jo K, *et al.* (2007) A single-molecule barcoding system using nanoslits for DNA analysis. *Proc Natl Acad Sci USA* 104:2673–2678.
17. Volff JN, Altenbuchner J (2000) A new beginning with new ends: Linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* 186:143–150.
18. Cui T, *et al.* (2007) *Escherichia coli* with a linear genome. *EMBO Rep* 8:181–187.
19. Tourand Y, *et al.* (2006) Differential telomere processing by *Borrelia* telomere resolvases *in vitro* but not *in vivo*. *J Bacteriol* 188:7378–7386.
20. Bao K, Cohen SN (2003) Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev* 17:774–785.
21. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S (2004) Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res* 32:3781–3791.
22. Gao F, Zhang CT (2008) Ori-Finder: A web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics* 9:79.
23. Nakamura Y, *et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* 9:123–130.
24. Lipton MS, *et al.* (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci USA* 99:11049–11054.
25. 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.
26. Toepel J, Welsh E, Summerfield TC, Pakrasi H, Sherman LA (2008) Differential transcriptional analysis of the cyanobacterium *Cyanothece* sp. ATCC 51142 during light-dark and continuous-light growth. *J Bacteriol* 190:3904–3913.
27. Kneip C, Voss C, Lockhart PJ, Maier UG (2008) The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC Evol Biol* 8:30.
28. Schneegurt MA, Sherman DM, Nayar S, Sherman LA (1994) Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the cyanobacterium *Cyanothece* sp. strain ATCC 51142. *J Bacteriol* 176:1586–1597.
29. Kanehisa M, *et al.* (2006) From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res* 34:D354–357.
30. Larimer FW, *et al.* (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium. *Rhodopseudomonas palustris*. *Nat Biotechnol* 22:55–61.
31. Hinnebusch J, Tilly K (1993) Linear plasmids and chromosomes in bacteria. *Mol Microbiol* 10:917–922.
32. Badger JH, Olsen GJ (1999) CRITICA: Coding region identification tool invoking comparative analysis. *Mol Biol Evol* 16:512–524.
33. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27:4636–4641.
34. Oehmen CS, Nieplocha J (2006) ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis. *IEEE Trans Parallel Distributed Systems* 17:740–749.
35. Apweiler R, *et al.* (2004) UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res* 32:D115–119.
36. Gibson R, Smith DR (2003) Genome visualization made fast and simple. *Bioinformatics* 19:1449–1450.
37. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.
38. Nawrocki EP, Eddy SR (2007) Query-Dependent Banding (QDB) for Faster RNA Similarity Searches. *PLoS Comput Biol* 3:e56.
39. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: An RNA family database. *Nucleic Acids Res* 31:439–441.
40. Stal LJ, Moezelaar R (1997) Fermentation in cyanobacteria. *FEMS Microbiol Rev* 21:179–211.