8-6-2008

The Genome of Cyanothece 51142, a Unicellular Diazotrophic Cyanobacterium Important in the Marine Nitrogen Cycle.

Eric A. Welsh
Washington University in St Louis

Michelle Liberton
Washington University in St Louis

Jana Stöckel
Washington University in St Louis

Thomas Loh
Washington University in St Louis

Chunyan Wang
Washington University School of Medicine in St. Louis

See next page for additional authors

Recommended Citation

Welsh, Eric A.; Liberton, Michelle; Stöckel, Jana; Loh, Thomas; Wang, Chunyan; Wollam, Aye; Fulton, Robert S.; Clifton, Sandra W.; Jacobs, Jon M.; Aurora, Rajeev; Sherman, Louis A.; Smith, Richard D.; Wilson, Richard K.; and Pakrasi, Himadri B., "The Genome of Cyanothece 51142, a Unicellular Diazotrophic Cyanobacterium Important in the Marine Nitrogen Cycle." (2008). *Department of Biological Sciences Faculty Publications.* Paper 38.
http://dx.doi.org/10.1073/pnas.0805418105

Authors
Eric A. Welsh, Michelle Liberton, Jana Stöckel, Thomas Loh, Chunyan Wang, Aye Wollam, Robert S. Fulton, Sandra W. Clifton, Jon M. Jacobs, Rajeev Aurora, Louis A. Sherman, Richard D. Smith, Richard K. Wilson, and Himadri B. Pakrasi

Complete genome sequence of *Cyanothece* 51142, a cyanobacterium that uses metabolic compartmentation for bioenergy production

Eric A. Welsh[1,6], Michelle Liberton[1,6], Jana Stöckel[1], Thomas Loh[1], Chunyan Wang[2], Aye Wollam[2], Robert S. Fulton[2], Sandra W. Clifton[2], Jon M. Jacobs[3], Rajeev Aurora[4], Louis A. Sherman[5], Richard D. Smith[3], Richard K. Wilson[2] & Himadri B. Pakrasi[1]

[1]Department of Biology, Washington University, St. Louis, MO 63130, USA. [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108, USA. [3]Pacific Northwest National Laboratory, Richland, WA 99352, USA. [4]Department of Molecular Microbiology & Immunology, Saint Louis University School Of Medicine, St. Louis, MO 63104, USA. [5]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to H.B.P. (pakrasi@wustl.edu).

ABSTRACT

Cyanobacteria are photosynthetic prokaryotes that directly convert solar energy to energy-rich biochemical products. The unicellular diazotrophic cyanobacterium *Cyanothece* 51142 has the special ability to store large quantities of carbohydrates as glycogen during the day and subsequently utilize this stored energy during the night to meet the metabolic needs of nitrogen fixation. Analysis of the 5,460,377 base pair *Cyanothece* genome revealed 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. Annotation of the *Cyanothece* genome was significantly aided by the use of high-throughput global proteomics data. This genome sequence provided important insights into the ability of *Cyanothece* to use metabolic compartmentalization and energy storage that makes it an ideal system for bioenergy research, as well as studies of how a unicellular organism balances multiple, often incompatible, processes in the same cell.

Bioenergy production has recently become a topic of intense interest due to increased concern regarding limited petroleum-based fuel supplies and the contribution of the use of these fuels to atmospheric CO_2 levels. A number of organisms have received attention as biofuel producers, including corn, sugar cane, soybean, oil palm, and algae[1]. Cyanobacteria present an alternative system which offers certain advantages in bioenergy production. Cyanobacteria are oxygenic photosynthetic bacteria that have significant roles in global biological carbon sequestration, oxygen production, and the nitrogen cycle [2, 3]. They occupy a diverse range of habitats, from open ocean to hot springs, deserts, and arctic waters. Many strains of cyanobacteria have well-developed methods for targeted gene replacement and modification, and can be readily grown to high density in photobioreactors (Nedbal L, Trtílek M, Červený J, Komárek O, Pakrasi HB. (2008) Biotechnology & Bioengineering. *in press).* Cyanobacterial strains capable of photoautotrophic, mixotrophic, and heterotrophic growth have been isolated. Of particular importance is the diversity of metabolic pathways in cyanobacteria that allow these organisms to succeed in a wide variety of environments and which provide a wealth of targets for metabolic engineering of energy-rich biomolecules.

Cyanothece are unicellular cyanobacteria with the ability to fix atmospheric nitrogen[4]. *Cyanothece* strains have been isolated worldwide from a variety of habitats, including fresh and salt waters, where they demonstrate significant diversity in their size, cellular shape, amount of extracellular polysacharride, and growth rates[5, 6]. A member of this genus, *Cyanothece* sp. PCC 7822, has been shown to evolve hydrogen and produce lactate, acetate, and ethanol[7]. Another *Cyanothece* strain, *Cyanothece* sp. ATCC 51142 (hereafter referred to as *Cyanothece* 51142), has been extensively characterized at the molecular and physiological levels[8]. Nitrogen fixation is biochemically incompatible with oxygenic photosynthesis, as the nitrogenase enzyme is inactivated by oxygen[9]. To resolve this problem, *Cyanothece* 51142 performs photosynthesis during the day and nitrogen fixation at night, thus temporally separating these processes in the same cell[10]. During a diurnal period, *Cyanothece* 51142 cells actively accumulate and subsequently degrade different storage inclusion bodies for the products of photosynthesis and N_2-fixation[10, 11]. We therefore identified *Cyanothece* 51142 as a natural biofactory for the conversion of carbon dioxide to carbohydrates, and

their subsequent utilization, and sought to understand the interplay between the metabolic processes in *Cyanothece* 51142 by analysis of its complete genome sequence, aided by the use of high-throughput proteomics data.

RESULTS

Architecture of the *Cyanothece* 51142 genome

The *Cyanothece* 51142 genome was sequenced using an initial 9.3x coverage shotgun library, followed by fosmid libraries and several rounds of 454 sequencing. The finished assembly was independently confirmed using an optical restriction map generated by OpGen Technologies, Inc. (Madison, Wisconsin). Optical mapping is a technique that has been successfully used in bacterial genome sequence finishing [12], and was here used to generate a restriction endonuclease-based map of the entire *Cyanothece* 51142 genome. The combination of these two sequencing approaches confirmed that the 5.46 Mb *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 notably, a 430 kb linear chromosome (Fig. 1 and Supplementary Table 1). Both the genome assembly and the optical map (Supplementary Fig. 1) indicate that the ~430 kb chromosome is linear. The finding of a linear element in the *Cyanothece* 51142 genome was unanticipated, and is the first report of such an element in the genome of a photosynthetic bacterium. Linear genomic elements have previously been found in other bacterial genera such as *Borrelia*[13], *Streptomyces*[14], and *Agrobacterium*[15]. Among photosynthetic bacteria, only circular chromosomal elements have been described, although these can range in size and organization: *Rhodobacter sphaeroides*, an anoxygenic photosynthetic bacterium, has two circular chromosomes[16], and the filamentous cyanobacterial strains *Anabaena* sp. PCC 7120 and *Anabaena variabilis* sp. ATCC 29413 contain large (~640 kb) circular plasmids[17].

The gene content of the linear chromosome was examined to investigate its possible origin and importance to the organism. The linear chromosome is 429,701 bp long, and contains 449 predicted protein-coding sequences, 127 (28.3%) with assigned function (Table 1). Of the remaining protein-coding sequences, 24 (5.3%) are predicted to be unknown genes, while the remaining 298 (66.4%) are hypothetical. Overall, the linear chromosome contains a much higher percentage of genes with no assigned function (71.7% vs. 45.7%) compared to the large circular chromosome. The GC content for the

linear and circular chromosome is comparable (Table 1). Codon usage within the linear chromosome was similar to that of the circular chromosome of *Cyanothece* 51142, and to the genomes of closely related cyanobacteria (data not shown).

Of the 108 genes on the linear chromosome assigned to functional categories, 38 have a corresponding copy on the circular chromosome or plasmids (Supplementary Table 2). These multi-copy genes include, on the linear chromosome, the genes in a *coxBAC* operon encoding the terminal cytochrome *c* oxidase of the respiratory electron transport chain, and several genes (*gpm, pyk, pgi, eno, ackA, xpk, glgP, ald*) within a cluster related to glucose metabolism. Analysis of these multi-copy genes did not uncover any significant conserved synteny between the linear and circular chromosomes of *Cyanothece* 51142. However, further examination revealed synteny between the linear chromosome and the genome of only one other bacterial strain, *Cyanothece* sp. CCY 0110. In addition, 43 genes are unique to the linear chromosome (Supplementary Table 2). These include the operon encoding the large and small subunits of the exodeoxyribonuclease *xseA/xseB* and the cobalamin-independent methionine synthase *metE.* Importantly, also within the glucose metabolism cluster on the linear chromosome is the only gene encoding L-lactate dehydrogenase in the *Cyanothece* 51142 genome. This enzyme catalyzes the terminal step in the production of lactate, a fermentation product previously identified in other cyanobacterial strains[18].

An important question concerning the linear chromosome of *Cyanothece* 51142 is whether the genes located on it are transcribed and translated. From the data generated as part of a global proteomics analysis (described below) that was used to aid the genome annotation, we identified the products of 122 of the 449 protein-coding genes on the linear chromosome (Supplementary Table 2).

We also examined the linear chromosome to determine how it is replicated and maintained. Sequence analysis, however, failed to reveal near the ends of the linear element the presence of any feature, such as inverted repeats or stem-loop structures, known to be related to such functions in previously characterized bacterial genera[19].

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 (Table 1). 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 (For details, see Methods section). Proteomic data were used, in conjunction with early draft genomic sequence, to build a mass and time (MT) library[20] for use in quantitative proteomics experiments. The observed peptides in the MT library covered approximately 50% of the predicted proteome. The MT library was compared to the predicted proteome based on the initial draft genome sequence, and of the 1,989 proteins for which there was no significant sequence homology to proteins of known function, 506 (25.4% of this group) were reclassified as 'unknowns' due to the observation of corresponding tryptic peptides. Additionally, the observed tryptic peptides were matched against a set of ~12,000 low confidence open reading frame predictions, resulting in the inclusion of 38 additional open reading frames that would not have otherwise been included 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 over 500 genes, and lent an additional level of validation to the genome annotation.

Phylogenetic analysis

To understand the relationship between *Cyanothece* 51142 and other nitrogen- and non-nitrogen fixing cyanobacteria, a phylogenetic tree was constructed from 435 orthologous proteins present in 28 cyanobacterial strains with sequenced genomes (Fig. 2). This group of cyanobacteria included nitrogen-fixing as well as non-nitrogen-fixing strains, chosen as a wide range of representative strains. The large set of proteins was used to average out any differences in mutation rates between proteins under different selective pressures, and to remove any bias from proteins involved in processes, such as nitrogen-fixation, that are not universally present in all of the cyanobacteria analyzed. *Gloeobacter* was designated as the outgroup, as it is thought to be the most anciently branching cyanobacterium[21]. The deduced tree branched into two distinct groups, the $\alpha-$ and $\beta-$cyanobacteria, as previously described based on carbon uptake/fixation

related proteins[22]. The overabundance of the HIP1 (GCGATCGC)[23] and HIP1-like (GGGATCCC) octameric palindromic repeats appears to be endemic to the β-cyanobacteria, with low abundance in *Gloeobacter* and all α-cyanobacteria (Supplementary Table 3), and thus can be used as independent verification of the tree. The presence of a β-cyanobacterial carbon uptake system, an overabundance of the HIP1 repeat, and its position in the phylogenetic tree place *Cyanothece* 51142 within the β-cyanobacterial group.

As shown in Figure 2, the diazotrophic fresh water thermophilic strains *Synechococcus* JA-3-3Ab and JA-2-3B'a(2-13) appear to be the earliest branching of the β-cyanobacteria after *Gloeobacter*. The non-thermophilic unicellular N_2-fixing strains (*Cyanothece* 51142 and *Crocosphaera*) branch separately from the N_2-fixing filamentous strains (*Anabaena, Nostoc* and *Trichodesmium*), with the non-heterocyst-forming *Trichodesmium* branching the earliest among the filamentous strains. Importantly, this phylogeny of the diazotrophs agrees with the physiological differences between the strains, even though proteins related to nitrogen fixation and heterocyst formation were excluded from our analysis. Surprisingly, *Synechocystis* sp. PCC 6803, a non-diazotrophic unicellular fresh water strain, branches together with *Cyanothece* 51142 and *Crocosphaera*, rather than with the non-diazotrophic *Thermosynechococcus elongatus.* This implies that *Synechocystis* is a distant member of the *Cyanothece* group that might have lost the *nif* cluster and other genes associated with diurnal nitrogen fixation capability, and suggests that the ability to fix nitrogen was present in the common ancestor to the *Cyanothece* and *Anabaena* branches and was subsequently lost in *Synechocystis*.

Nitrogen fixation

Nitrogen fixation is an energy intensive process requiring the use of 16 ATP molecules per N_2 molecule converted to ammonia. The energy to perform nitrogen fixation comes from carbohydrates produced by photosynthesis and stored in the form of glycogen,

which is consumed early in the dark period[10, 24]. Fixed nitrogen in turn is stored in cyanophycin granules, which are fully depleted during the next light period to provide the cell with nitrogen for amino acid and nucleic acid synthesis[8]. In the genome of *Cyanothece* 51142*,* most of the genes involved in nitrogen fixation are located in a contiguous cluster of 34 genes separated by no more than 3 kb. 28 of these genes have conserved synteny with those found in other sequenced nitrogen-fixing cyanobacteria. We sought to understand in detail the organization of the genes involved in nitrogen fixation in the genomes of *Cyanothece* 51142 and other cyanobacteria. As shown in Fig. 3, *Cyanothece* 51142 contains the largest contiguous cluster of nitrogen fixation-related genes. *Crocosphaera watsonii* WH 8501 contains a nearly identical cluster to that of *Cyanothece* 51142, with the exception of a missing putative transcriptional regulator (*cce_0556*) between *nifB* and *nifS*. *Trichodesmium erythraeum* IMS 101 also contains a single cluster, but with several genes absent from it. The *Nostoc punctiforme* PCC 73102 gene cluster has the same general organization as in *Cyanothece* 51142, but with a 5.2 kb insert between the *cysE2* and *nifB* operons, and a 20 kb insert between *nifD* and *nifK*. In the heterocyst-forming *Nostoc* and *Anabaena* strains, one or more inserts ranging in size from 9 kb - 24 kb break the cluster into several smaller clusters, with a number of genes duplicated or missing between clusters. In the most anciently branching of the nitrogen fixing cyanobacteria, *Synechococcus* sp. JA-2-3B'a(2-13) and *Synechococcus* sp. JA-3-3Ab, the middle section of the single cluster is inverted relative to that of *Cyanothece* (denoted in brackets), and several genes are missing. Such patterns of conserved synteny across these cyanobacteria together with the proteome-wide phylogenetic tree (Fig. 2) support a single acquisition event of the *nif* cluster in a common ancestor.

Photosynthesis, respiration, and circadian control

Cyanobacteria are unique in that the internal thylakoid membrane system houses the components of both the photosynthetic and respiratory electron transport chains, and that components of these processes overlap[25]. The *Cyanothece* 51142 genome contains all the expected photosynthesis and light harvesting genes, including the genes required for a functional cytochrome b_6f complex, which is shared between

photosynthesis and respiration. *Cyanothece* 51142 has 5 copies of the *psbA* gene encoding the D1 protein of Photosystem II, two more than are found in *Synechocystis* 6803, in which photosynthesis and the D1 protein have been extensively studied[26]. The genome of *Cyanothece* 51142 includes the complete set of genes required for respiration. Similar to other cyanobacteria, different types of terminal oxidases are present to catalyse the final step in oxidative phosphorylation: a cytochrome *bd*-quinol oxidase (*cydA* and *cydB*), and three copies of the *coxBAC* operon encoding aa_3-type cytochrome *c* oxidases.

Photosynthetic capacity and respiratory activity have both previously been shown to exhibit strong diurnal oscillations in *Cyanothece* 51142[27]. Our data show that many of the circadian clock genes are present in the *Cyanothece* 51142 genome. These include orthologs for central elements of the circadian clock (*kaiA*, *kaiB*, and *kaiC)*, components of input pathways (*cikA* and *ldpA*), as well as elements identified to participate in output pathways (*sasA*, *rpaA*, *cmpA*, and *sigB*). However, a *pex* ortholog is missing in the *Cyanothece* 51142 genome and also in the draft genome of the closely related strains *Cyanothece* sp. CCY 0110 and *Crocospaera watsonii*. The *pex* gene from *Synechococcus* sp. PCC 7942 is thought to modulate the function of the central clock oscillator[28].

Energy metabolism and fermentation

In diverse organisms including archaea, eubacteria, fungi, and animals, glucose is stored as glycogen granules. Pathways involved in glucose metabolism in *Cyanothece* 51142 are shown in Fig. 4a. *Cyanothece* 51142 cells are programmed to undergo diurnal cycles of glycogen synthesis in the light, followed by degradation and utilization in the following dark period[10]. Accordingly, the *Cyanothece* 51142 genome includes the necessary genes for glycogen synthesis, such as the glycogen synthase *glgA*, the 1,4-alpha-glucan branching enzyme *glgB* and the glucose-1-P adenylyltransferase *glgC*, each with two copies. Furthermore, *Cyanothece* 51142 contains three different genes encoding glycogen phosphorylases (*glgP1-P3)* for glycogen degradation, with *glgP2* found on the linear chromosome (Fig. 4b and Supplementary Fig. 2). Interestingly, two

distinctly different glycogen debranching enzymes are present, one that is found only in the β-cyanobacteria (*cce_3194*), and another (*glgX*; *cce_3465*) that is found mainly in the α-cyanobacteria.

The pathways of central carbon metabolism (glycolysis, the pentose phosphate pathway, and the tricarboxylic acid cycle) are highly conserved among a wide range of different organisms. In *Cyanothece* 51142*,* nearly all of the genes associated with glycolysis are present as multi-copy genes, with copies on both the circular and linear chromosomes. In contrast, all the genes of the tricarboxylic acid cycle are only single copy genes (data not shown). Similar to other well-studied cyanobacteria, *Cyanothece* 51142 contains an incomplete tricarboxylic acid cycle, with genes encoding for components of the 2-oxoglutarate dehydrogenase complex being absent. *Cyanothece* 51142 contains a gene encoding the phosphoenolpyruvate carboxykinase (PEPCK; *cce_0508*), the enzyme which performs the first step in gluconeogenesis, circumventing the irreversible reaction of phosphoenolpyruvate to pyruvate in glycolysis. This enzyme therefore links carbon, organic acid, and amino acid metabolism. PEPCK is also present in *Rhodopseudomonas palustris*[29]*,* a purple photosynthetic bacterium that belongs to the α-proteobacteria, but is missing in nearly all other cyanobacterial strains sequenced to date.

Analysis of the genes involved in pyruvate metabolism and fermentation revealed significant differences between *Cyanothece* 51142 and other cyanobacteria. *Cyanothece* 51142 contains all the genes necessary for ethanol, lactate, and acetate fermentation (Fig. 4a), which requires an anoxic environment. *Cyanothece* 51142 creates such an intracellular environment during the early dark period in order to fix nitrogen. The production of ethanol and energy rich organic acids has been shown experimentally to occur in another related strain, *Cyanothece* PCC 7822[7, 18]. Interestingly, while genes related to carbohydrate and energy metabolism are localized at multiple loci on the circular chromosome, a 20.2 kb gene cluster found on the linear chromosome contains several genes involved in glucose metabolism (Fig. 4b). Within this cluster is the only gene encoding L-lactate dehydrogenase *(ldh; cce_5187)* in the *Cyanothece* 51142 genome. This enzyme is required for the terminal step of lactate

fermentation, suggesting that the linear chromosome is important for fermentation. This gene cluster on the linear chromosome does not show any conserved synteny to other cyanobacterial strains, or to any KEGG genomes, and is therefore unique to *Cyanothece* 51142.

Carbon sequestration

Much of the carbon required for carbon fixation and glycogen storage is imported into the cell in the form of CO_2 and bicarbonate. *Cyanothece* 51142 has Form 1B Rubisco and carboxysomes, and all of the CO_2 and bicarbonate transporters typical of β-cyanobacteria[22]. These include the high affinity BCT1 (*cmpABCD* operon) and inducible SbtA bicarbonate transporters, the low affinity BicA bicarbonate transporter, and the NDH-1_3 and NDH-1_4 CO_2 transporters. Many β-cyanobacteria contain both high affinity *cmpABCD* bicarbonate and *nrtABCD* nitrate ABC transporters, which are highly homologous to one another. *Cyanothece* 51142 has only one such operon, as does *Nostoc punctiforme* PCC 73102. Based on sequence homology to other strains, as well as the active site residues from crystal structures[30, 31], this operon in *Cyanothece* 51142 likely encodes a bicarbonate transporter. Thus, instead of an NrtABCD nitrate transporter, *Cyanothece* 51142 has a high-affinity nitrate/nitrite bispecific transporter[32], NrtP, which is found mainly in α-cyanobacterial *Synechococcus* strains.

DISCUSSION

Among the well-studied oxygenic photosynthetic microbes, *Cyanothece* sp. ATCC 51142 has a unique and significant ability to synthesize, store, degrade, and utilize large quantities of storage products as part of a robust diurnal cycle, an ability that suggests important future applications in bioenergy production. *Cyanothece* 51142 is the first of the *Cyanothece* family to be fully sequenced, and thus provides a basis upon which the regulation of numerous metabolic processes and storage capabilities within a single N_2-fixing oxygenic photosynthetic cell may be further investigated.

In total, more than 30 cyanobacterial genomes have been sequenced to date. Compared to them, of considerable interest is the organization of the 5.46 Mb *Cyanothece* 51142 genome, with the presence of a linear chromosome in addition to circular elements, an arrangement previously not reported in a photosynthetic bacterium. High throughput proteomic analysis determined that more than 25% of the genes assigned to functional categories on the linear chromosome are translated, and that among those are genes involved in important metabolic pathways in *Cyanothece* 51142, specifically for glucose metabolism (Fig. 4). Furthermore, transcriptional analyses have revealed that the expression of genes within this gene cluster is highly coregulated, with maximal transcript abundance during the early dark period (Stöckel et al., under review). We therefore conclude from these analyses that the linear chromosome is a *bona fide* genomic constituent with important functional roles. Interestingly, our analysis of the unfinished genome sequence of *Cyanothece* sp. CCY 0110 (research.venterinstitute.org/moore) revealed that this cyanobacterium contains four contigs covering ~171,130 bp of sequence that is highly homologous to that of the linear chromosome of *Cyanothece* 51142, implying that the genome of *Cyanothece* sp. CCY 0110 may also contain a linear chromosome.

Ethanol and hydrogen are examples of two biofuels that largely require anoxic environments for their production. As cyanobacteria produce oxygen as a byproduct of photosynthesis, it is necessary for a biofuel-producing cyanobacterium to have the ability to create an anoxic intracellular environment. In *Cyanothece* 51142, a respiratory

burst at the beginning of the dark period contributes energy for nitrogen fixation, and serves to deplete O_2 in the cell that would inhibit nitrogen fixation. *Synechocystis* 6803, a close relative of *Cyanothece* 51142, lacks the diurnal cycling of photosynthesis and nitrogen fixation that produces such an anoxic environment, as well as the enzymes necessary for fermentation. In our examination of the fermentation pathways in *Cyanothece* 51142, we found that the gene encoding lactate dehydrogenase, which catalyzes the conversion of pyruvate to lactate, is found as a single copy gene on the linear chromosome, and that corresponding peptides for this gene are detectable in the proteomics data (Supplementary Table 2). Thus, it is likely that *Cyanothece* 51142 has the ability to ferment ethanol, and may do so in the early dark period to generate additional ATP molecules for nitrogen fixation.

Hydrogen is produced by nitrogen-fixing cyanobacteria by two different enzymes, a bidirectional hydrogenase (*hoxHYUFE),* and the nitrogenase. Our studies have shown that the expression of the genes encoding both the nitrogenase and the bidirectional hydrogenase are upregulated during the early dark period (Stöckel et al., under review). It has been proposed that the bidirectional hydrogenase may also function in the conversion of pyruvate to acetyl-CoA in the fermentative pathway[18].

It is notable that genes involved in the biosynthesis and assembly of the nitrogenase enzyme are found together in *Cyanothece* 51142 in the most highly conserved gene cluster of all cyanobacteria examined. The large size of the *Cyanothece* 51142 cluster and the organization of *nif* genes in other cyanobacteria (Fig. 3) suggests that the *Cyanothece* 51142 *nif* cluster may be most representative of an ancestral *nif* cluster organization, consisting of two operons on opposite strands. The differences between the other strains can be most parsimoniously explained by starting with a single *Cyanothece* 51142 -like cluster and introducing various deletion, mutation, and translocation events. The patterns of conserved synteny across all of the cyanobacteria analyzed (Fig. 3) strongly supports a *Cyanothece* 51142-like organization of the *nif* cluster in a shared common ancestor.

The classification of cyanobacteria into α and β groups was based on the occurrence of Form 1A or Form 1B Rubisco[22]. While phylogenetic analysis placed

Cyanothece within the β-cyanobacteria, the lack of an NrtABCD nitrate transporter and the subsequent presence of the NrtP nitrate transporter, combined with the presence of an α-cyanobacterial glycogen debranching enzyme in *Cyanothece* and other β-cyanobacteria was unexpected, and may imply a larger degree of lateral gene transfer between different groups than was previously supposed.

The *Cyanothece* 51142 genome details the pathways of storage granule accumulation as glycogen, cyanophycin and polyphosphate, and reveals further insights into interconnections between different pathways (Fig. 5). In the *Cyanothece* 51142 genome, we have identified over 140 genes related to nitrogen fixation, photosynthesis, carbon uptake and sequestration, fermentation, as well as granule formation and degradation, critical metabolic processes that exhibit diurnal rhythms. The complete genome sequence of *Cyanothece* 51142 provides an important road map for the use of systems and synthetic biology approaches for bioenergy production.

METHODS

Cell growth. Cyanothece cells were routinely grown under 12 hr light/12 hr dark conditions (50 µmol photons $m^{-2}s^{-1}$ at 30°C) in ASP2[4] media without added nitrate ($NaNO_3$).

Optical mapping. Whole cell samples were sent to OpGen, Inc (Madison, WI), where the genome was assembled using optical mapping techniques[12]. DNA was isolated and sheered into large molecular weight (> 200 kb) fragments, which were then loaded onto microfluidic slides on which individual strands of DNA were stretched into linear extended conformations. The endonuclease *Afl*II was washed over the slides, the DNA fluorolabeled, and the digests imaged using fluorescence microscopy. These images were then converted into "barcodes" representing the digest patterns. OpGen's assembly software was used to assemble all overlapping barcode patterns into the finished assembly (Supplementary Fig. 1). Barcode patterns computationally generated from the final genome sequence (Supplementary Fig. 1) agree with the optical mapping to within the error of the optical mapping technique (+/- 5 kb). Plasmids were not optically mapped, due to their small size.

Genome annotation. Open reading frames (ORFs) were predicted using a combination of CRITICA v1.05[33] and GLIMMER v2.13[34], and a database of all available cyanobacterial genomes[17, 35] for comparison. ScalaBLAST[36] hits were generated versus UNIPROT release 7.7[37], as well as the proteomes of all then-available sequenced cyanobacteria. InterProScan v4.0 was used to scan InterPro v12.1 [38] to identify domains within each ORF, and the ORFs were mapped onto custom local KEGG pathways using highest scoring homologs in KEGG release 38[39]. Manual annotation was performed based on all information provided by the ScalaBLAST hits, InterProScan domains, and KEGG pathways. High throughput proteomics data was used to reclassify genes with no informative sequence homology as unknowns, based on the observation of at least two different unique peptides per protein in the MT library, or only a single unique peptide if the protein was conserved in 6 or more strains. The annotated genome was rendered into the circular and linear diagrams composing Fig 1.

using the GenomePlot software[40]. All tRNAs were assigned using tRNAscan-SE v1.23[41] using the general model. All other ncRNAs were assigned using INFERNAL v0.72[42] and the Rfam database[43], searching only for those ncRNAs identified in other cyanobacterial genomes on the Rfam website.

Analysis of HlP1 and HlP1-like repeats. The genome of each cyanobacterium was scanned for all possible 8-mers. The GC content was determined for each genome, and used to calculate the expected frequency of each 8-mer at random in the genome. The overabundance of each 8-mer was calculated as the ratio of the observed frequency over the expected frequency. The results are summarized in Supplementary Table 3.

Phylogenetic analysis. Orthologous proteins were identified through an all versus all blastp analysis of 28 cyanobacterial proteomes[17, 35]. Orthology was defined as reciprocal best match hits between proteomes, matching over 66% the length of the longer of the two proteins, with scores $\geq$ 1/10 the higher of the self-self scores. Any highest-scoring protein with multiple identical-scoring hits was discarded. Sets of orthologs were considered to be conserved if all proteins within the set were orthologous to one another, resulting in 435 sets of 28 proteins each. Each set was aligned using MAFFT[44] and concatenated into a single alignment, removing all columns containing gaps. The PHYLIP v3.64[45] package was used to generate the final consensus tree, using the Fitch-Margoliash method with 100 bootstraps, global rearrangement, and 784 jumbles. Bootstrap values for all branches were 100.

Cell growth and sample preparation for proteomic analysis. After 7 days of growth, 150 mL cell samples were taken every 2h for 2 days, starting one hour into the dark period (time point D1). Cells were washed with 100 mM ammonium bicarbonate buffer (pH 8.0) and broken by two passages through a French Press at 20,000 psi. Soluble and membrane fractions were separated by centrifugation at 150,000*g* at 4°C for 20 min and then stored at –80°C. The soluble portion of the proteomic samples was denatured and reduced using 8 M Urea and 5 mM tributylphosphine (Sigma-Aldrich, St. Louis, MO) at 37˚C for 60 min. The membrane portion of the samples was treated identically,

except for an additional 5 min of sonication immediately prior to incubation. All samples were then diluted 5 times in 25 mM ammonium bicarbonate prior to tryptic digestion. Samples subjected to strong cation exchange chromatography (SCX) used a PolySulfoethyl A, 200 mm x 2.1 mm, 5 µM, 300-Å column with 10 mm x 2.1 mm guard column (PolyLC, Inc., Columbia, MD) with a flow rate of 0.2 mL/min. The SCX peptide fractionation step has been previously described[46, 47]. The peptides were resuspended in 900 µL of mobile phase A, and separated on an Agilent 1100 system (Agilent, Palo Alto, CA) equipped with a quaternary pump, degasser, diode array detector, peltier-cooled autosampler and fraction collector (both set at 4 ºC). A total of 25 fractions were collected for each sample.

Reversed phase LC separation and MS/MS analysis of peptides. This method has been extensively reported[48] with the coupling of a constant pressure (5,000 psi) reversed phase capillary liquid chromatography system (150 µm i.d. × 360 µm o.d. × 65 cm capillary; Polymicro Technologies Inc., Phoenix, AZ) with a Finnigan LTQ ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) using an electrospray ionization source manufactured in-house. Each unfractionated and SCX fraction was analyzed via capillary RPLC-MS/MS.

LC-MS/MS data analysis. SEQUEST analysis software was used to match the MS/MS fragmentation spectra to sequences from the initial draft *Cyanothece* 51142 proteome. The criteria selected for filtering followed methods based upon a reverse database false positive model which has been shown to give ~95% confidence for an entire protein dataset[49]. A mass and time (MT) tag database containing the calculated mass and normalized elution time (NET) for each identified peptide was generated to assist with subsequent high sensitivity, high-throughput analysis of *Cyanothece* 51142 samples using the accurate mass and time (MT) tag approach.

Homology analysis of the linear chromosome. Protein coding sequences from the linear chromosome were searched against UNIPROT release 7.7[37] and all currently available cyanobacterial genomes[17, 35], including the circular chromosome and plasmids of *Cyanothece* 51142, using BLAST v2.2.15[50] with an E-value cutoff of 0.01 and low

complexity filtering disabled. Top scoring hits were used to generate Supplementary Table 2. Identification of proteins unique to the linear chromosome required no homolog to exist within the circular chromosome or plasmids, and the presence of a homolog in at least one other organism. The criteria used for homolog assignment were BLAST scores $\geq$1/3 the self-self score, and hit lengths spanning $\geq$ 2/3 the lengths of the proteins.

Analysis of *nif* cluster. Proteins encoded by the *Cyanothece* 51142 *nif* cluster were searched against all available cyanobacterial genomes using BLASTP and TBLASTN, with an E-value cutoff of 0.01 and low complexity filtering disabled. Hits within 10 kb of each other within each genome were identified as part of a cluster of Nif proteins. For the data presented in Fig. 3, synteny to the *Cyanothece* 51142 cluster was identified manually and homologous genes were assigned.

Accession codes. Genome sequence data have been deposited in GenBank under accession numbers CP000806 - CP000811.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGEMENTS

We thank all members of the Pakrasi Lab for collegial discussions. This work was supported by funding from the Danforth Foundation at Washington University. This work is also part of a Membrane Biology EMSL 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 (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle.

AUTHOR CONTRIBUTIONS

E.A.W and T.L. analyzed sequence data and genome annotation. C.W., A.W., R.S.F., S.W.C. and R.K.W. contributed to genome sequencing. M.L. and J.S. designed

research, performed experiments, and analyzed data. J.M.J and R.D.S. contributed proteomic data. E.A.W, M.L. and J.S. wrote the paper. H.B.P., L.A.S. and R.A. oversaw the project. All authors discussed the results and commented on the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

FIGURE LEGENDS

Figure 1. General genome features and distribution of gene functions within the *Cyanothece* 51142 genome. Labels are given in megabase pairs for the circular chromosome (**a**), and kilobase pairs for the linear chromosome (**b**) and plasmids (**c**). Genes are colored by functional category as follows: energy, fatty acid, phospholipid metabolism (red); cell envelope (orange); cellular processes (steel blue); central intermediary metabolism (light green); photosynthesis and respiration (dark green); regulation (cyan); DNA, transcription, translation (dark blue); small molecule biosynthesis (magenta); transport and binding (purple); unknown/hypothetical (grey); other (black); non-coding RNA (yellow). The two rRNA operons on the circular chromosome are indicated in yellow at 3.95 Mbp and 4.10 Mbp in (**a**). The glucose metabolism cluster on the top strand of the linear chromosome is shown in light green and red between 375 kbp and 400 kbp in (**b**).

Figure 2. Phylogenetic tree of cyanobacteria. The tree was generated from the analysis of 435 sets of proteins, co-orthologous in all 28 of the analyzed strains. All branches have bootstrap values of 100.

Figure 3. Clusters of N_2-fixation-related genes. Shown are genes with conserved synteny between *Cyanothece* 51142 and other nitrogen-fixing cyanobacteria. Two genes in *Anabaena* sp. PCC 7120, *all1516* and *all1454*, which are spliced during heterocyst formation, are denoted by asterisks (*).

Figure 4. Genes encoding enzymes in glucose metabolism pathways in *Cyanothece* 51142. (**a**) Each arrow shows the direction in which the reaction takes place; broken arrow indicates 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*; (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**) A cluster of genes on the linear chromosome that encode various enzymes involved in glucose metabolism.

Figure 5. 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[8]. Fixed nitrogen is stored in cyanophycin granules, which are completely depleted during the following day.

Table 1. Comparison of the general features of the linear and circular chromosomes of *Cyanothece* 51142.

	Total Genome	(%)	Circular Chromosome	(%)	Linear Chromosome	(%)
Size (bp)	5,460,377		4,934,271		429,701	
G+C content (%)	37.9		37.9		38.6	
Protein-coding genes	5,304	100.0	4,762	100.0	449	100.0
Assigned function	2,735	51.6	2,584	54.3	127	28.3
Unknown	506	9.5	468	9.8	24	5.3
Hypothetical	2,063	38.9	1,710	35.9	298	66.4

REFERENCES

1. Hankamer, B. *et al.* Photosynthetic biomass and H_2 production by green algae: from bioengineering to bioreactor scale-up. *Physiologia Plantarum* **131**, 10-21 (2007).
2. Bryant, D.A. & Frigaard, N.U. Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* **14**, 488-496 (2006).
3. Zehr, J.P. *et al.* Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* **412**, 635-638 (2001).
4. Reddy, K.J., Haskell, J.B., Sherman, D.M. & Sherman, L.A. Unicellular, aerobic nitrogen-fixing cyanobacteria of the genus *Cyanothece*. *J Bacteriol* **175**, 1284-1292 (1993).
5. Rippka, R. Isolation and purification of cyanobacteria. *Methods Enzymol* **167**, 3-27 (1988).
6. Porta, D., Rippka, R. & Hernandez-Marine, M. Unusual ultrastructural features in three strains of *Cyanothece* (cyanobacteria). *Arch Microbiol* **173**, 154-163 (2000).
7. van der Oost, J., Bulthuis, B.A., Feitz, S., Krab, K. & Kraayenhof, R. Fermentation metabolism of the unicellular cyanobacterium *Cyanothece* PCC 7822. *Arch Microbiol* **152**, 415-419 (1989).
8. Sherman, L.A., Meunier, P. & Colon-Lopez, M.S. Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth Res* **58**, 25-42 (1998).
9. Zehr, J.P., Methe, B. & Foster, R. New nitrogen-fixing microorganisms from the oceans: biological aspects and global implications. In *Biological Nitrogen Fixation, Sustainable Agriculture and the Environment*. (eds. Wang, Y.-P., Lin, M., Tian, Z.-X., Elmerich, C. & Newton, W. E.) 361-365, (Springer Netherlands, Dordrecht, 2005).
10. Schneegurt, M.A., Sherman, D.M., Nayar, S. & Sherman, L.A. Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the

cyanobacterium *Cyanothece* sp. strain ATCC 51142. *J Bacteriol* **176**, 1586-1597 (1994).

11. Li, H., Sherman, D.M., Bao, S. & Sherman, L.A. Pattern of cyanophycin accumulation in nitrogen-fixing and non-nitrogen-fixing cyanobacteria. *Arch Microbiol* **176**, 9-18 (2001).
12. Jo, K. *et al.* A single-molecule barcoding system using nanoslits for DNA analysis. *Proc Natl Acad Sci U S A* **104**, 2673-2678 (2007).
13. Ferdows, M.S. & Barbour, A.G. Megabase-sized linear DNA in the bacterium *Borrelia burgdorferi*, the Lyme disease agent. *Proc Natl Acad Sci U S A* **86**, 5969-5973 (1989).
14. Kinashi, H., Shimaji-Murayama, M. & Hanafusa, T. Integration of SCP1, a giant linear plasmid, into the *Streptomyces coelicolor* chromosome. *Gene* **115**, 35-41 (1992).
15. Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L. & Ramuz, M. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J Bacteriol* **175**, 7869-7874 (1993).
16. Suwanto, A. & Kaplan, S. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *J Bacteriol* **171**, 5850-5859 (1989).
17. Kulikova, T. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **32**, D27-30 (2004).
18. Stal, L.J. & Moezelaar, R. Fermentation in cyanobacteria. *FEMS Microbiol Rev* **21**, 179-211 (1997).
19. Volff, J.N. & Altenbuchner, J. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* **186**, 143-150 (2000).
20. Lipton, M.S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci U S A* **99**, 11049-11054 (2002).
21. Nakamura, Y. *et al.* Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137-145 (2003).

22. Badger, M.R., Price, G.D., Long, B.M. & Woodger, F.J. The environmental plasticity and ecological genomics of the cyanobacterial CO_2 concentrating mechanism. *J Exp Bot* **57**, 249-265 (2006).
23. Robinson, N.J. *et al.* Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. *Nucleic Acids Res* **23**, 729-735 (1995).
24. Bergman, B., Gallon, J.R., Rai, A.N. & Stal, L.J. N_2 fixation by non-heterocystous cyanobacteria. *FEMS Microbiol Rev* **19**, 139-185 (1997).
25. Vermaas, W.F.J. Photosynthesis and Respiration in Cyanobacteria. in Encyclopedia of Life Sciences 245-251, (Nature Publishing Group, London; 2001).
26. Roose, J.L. & Pakrasi, H.B. Evidence that D1 processing is required for manganese binding and extrinsic protein assembly into photosystem II. *J Biol Chem* **279**, 45417-45422 (2004).
27. Meunier, P.C., Colon-Lopez, M.S. & Sherman, L.A. Temporal Changes in State Transitions and Photosystem Organization in the Unicellular, Diazotrophic Cyanobacterium *Cyanothece* sp. ATCC 51142. *Plant Physiol* **115**, 991-1000 (1997).
28. Kutsuna, S., Kondo, T., Aoki, S. & Ishiura, M. A period-extender gene, *pex*, that extends the period of the circadian clock in the cyanobacterium *Synechococcus* sp. strain PCC 7942. *J Bacteriol* **180**, 2167-2174 (1998).
29. Larimer, F.W. *et al.* Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nat Biotechnol* **22**, 55-61 (2004).
30. Koropatkin, N.M., Pakrasi, H.B. & Smith, T.J. Atomic structure of a nitrate-binding protein crucial for photosynthetic productivity. *Proc Natl Acad Sci U S A* **103**, 9820-9825 (2006).
31. Koropatkin, N.M., Koppenaal, D.W., Pakrasi, H.B. & Smith, T.J. The structure of a cyanobacterial bicarbonate transport protein, CmpA. *J Biol Chem* **282**, 2606-2614 (2007).

32. Sakamoto, T., Inoue-Sakamoto, K. & Bryant, D.A. A novel nitrate/nitrite permease in the marine Cyanobacterium *Synechococcus* sp. strain PCC 7002. *J Bacteriol* **181**, 7363-7372 (1999).
33. Badger, J.H. & Olsen, G.J. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**, 512-524 (1999).
34. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641 (1999).
35. Markowitz, V.M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**, D344-348 (2006).
36. Oehmen, C.S. & Nieplocha, J. ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis. *IEEE Trans. Parallel Distributed Systems* **17**, 740-749 (2006).
37. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119 (2004).
38. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116-120 (2005).
39. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**, D354-357 (2006).
40. Gibson, R. & Smith, D.R. Genome visualization made fast and simple. *Bioinformatics* **19**, 1449-1450 (2003).
41. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
42. Nawrocki, E.P. & Eddy, S.R. Query-Dependent Banding (QDB) for Faster RNA Similarity Searches. *PLoS Comput Biol* **3**, e56 (2007).
43. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S.R. Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439-441 (2003).
44. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511-518 (2005).

45. Felsenstein, J. Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).
46. Jacobs, J.M. *et al.* Proteome analysis of liver cells expressing a full-length hepatitis C virus (HCV) replicon and biopsy specimens of posttransplantation liver from HCV-infected patients. *J Virol* **79**, 7558-7569 (2005).
47. Jacobs, J.M. *et al.* Multidimensional proteome analysis of human mammary epithelial cells. *J Proteome Res* **3**, 68-75 (2004).
48. Shen, Y. *et al.* Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal Chem* **73**, 1766-1775 (2001).
49. Qian, W.J. *et al.* Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J Proteome Res* **4**, 53-62 (2005).
50. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).