UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BIOINFORMATIC INVESTIGATIONS IN MARINE MICROBIAL ECOLOGY

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS

by

Philip Heller

December 2014

The Dissertation of Philip Heller is approved:

_____
Professor Jonathan Zehr, chair

_____
Professor Josh Stuart

_____
Rex Malmstrom, Ph.D.

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables and Figures

Chapter 1: ARBitrator: A software pipeline for on-demand retrieval of auto-curated *nifH* sequences from GenBank

Chapter 2: Metagenomics of Uncultivated UCYN-A Cyanobacteria

Chapter 3: Dexter: A tool for exploring diel expression data sets

Abstract

Philip Heller

Bioinformatic investigations in marine microbial ecology

Chapter 1: The number of known sequences for the *nifH* gene, commonly used to assess community potential for N_2 fixation, has been rapidly growing over the past few decades. Obtaining these sequences from the National Center for Biotechnology Information's GenBank database is problematic because of annotation errors, nomenclature variation, and paralogues; moreover, GenBank's tools are not conducive to searching solely by function. A software retrieval and curation pipeline called ARBitrator was developed that uses a BLAST search followed by a screening phase based on conserved domain similarity to retrieve *nifH* sequences from Genbank. A total of 34,420 *nifH* sequences were identified in GenBank. The false-positive rate is 0.033%. The pipeline can be adapted for other genes.

Chapter 2: Marine cyanobacteria capable of fixing molecular nitrogen ("diazotrophs") are key in biogeochemical cycling; the nitrogen fixed is a major

external source of nitrogen to the open ocean. *Candidatus* Atelocyanobacterium thalassa (UCYN-A) is a diazotrophic cyanobacterium known for its widespread geographic distribution, unusually reduced genome, and symbiosis with a single-celled prymnesiophyte alga. Recently a novel strain of this organism, called UCYN-A2, was detected in coastal waters off southern California. We assembled and analyzed the metagenome of this UCYN-A2 population. UCYN-A2 and the open-ocean UCYN-A1 strain share most protein-coding genes with high synteny, yet average amino-acid sequence identity between orthologous genes is only 86%. Our results suggest that UCYN-A1 and UCYN-A2 had a common ancestor and diverged after genome reduction.

Chapter 3: Gene expression in cells fluctuates over time in response to internal and external stimuli. Cyanobacteria, whose metabolism is tightly coupled to the sunlight cycle, have evolved complex patterns of gene expression that may derive from optimizing growth by coordinating photosynthesis and protein synthesis. These patterns can provide much information on how microorganisms grow and respond to the environment. However, analyzing complex information from whole-genome expression studies is difficult, requiring computational approaches that support visual exploration and analysis of data in order to detect related expression signatures.

A Java application called Dexter was developed to facilitate analysis of single or multiple gene expression time-series data sets. The value of the program was

demonstrated by using it to improve operon predictions.

Introduction

Background

The world's oceans harbor abundant single-celled life. Recent estimates compute the global marine prokaryote population (bacteria plus archaea) 1.2 x 10^29 cells, with biomass in excess of 300 Pg of carbon (Whitman et al., 1998). Cyanobacteria (photosynthetic bacteria) in the ocean, principally *Prochlorococcus* and *Synechococcus* (Partensky et al., 1999; Johnson et al., 2006; Scanlan et al., 2009), are important in global biogeochemical processes (Karl et al., 2002; Gruber & Galloway, 2008). For example, marine cyanobacteria annually convert 45-50 Gt of carbon in the form of atmospheric CO_2, to marine biomass (Longhurst 1995, Field 1998, Carr 2006). A significant fraction of biomass generated by this process, known as primary production, enters the food web and eventually sinks to the deep ocean, thus sequestering CO_2 (Falkowski 1997).

Primary production in the open ocean is often limited by scarcity of bioavailable reduced nitrogen in the forms of nitrate, nitrite, ammonium, urea and amino acids (Stal, 2009). A small number of prokaryotic organisms, known as diazotrophs, are able to provide new bioavailable nitrogen to microbial communities by reducing atmospheric N_2; this reaction hydrolyzes 16 units of ATP per molecule of molecular nitrogen reduced, and is one of the most metabolically expensive processes in nature (Raymond et al., 2004). Reduction is catalyzed by the nitrogenase enzyme, a tetramer composed of the products of the *nifH, nifD*, and *nifK* genes (Peters et al., 1995; Igarashi, 2003). Nitrogenase is inactivated by

even low concentrations of O_2 (Allison 1940; Parker & Scutt, 1960), and cyanobacteria, which evolve O_2 through photosynthesis, have consequently evolved several strategies for separating nitrogen fixation from the oxygen evolved by photosynthesis (Fay 1992; Bergman et al. 2012; . Several filamentous cyanobacteria, for example *Nostoc* (Fleming & Haselkorn, 1973), are known to form heterocysts, specialized cells that contain no chlorophyll and fix nitrogen that is transferred to the vegetative cells of the filaments (Fay 1992). Some unicellular cyanobacteria, notably *Crocosphaera watsonii*, only fix nitrogen at night (Sherman et al., 1998). The symbotic cyanobacterium *Candidatus* Atelocyanobacterium thalassa ("UCYN-A") has lost its photosystem II genes and obtains the products of photosynthesis from its eukaryotic single-celled hosts in exchange for reduced nitrogen (Tripp et al., 2010; Bombar et al., 2014). Intriguingly, *Trichodesmium erythraeum* is filamentous but does not form heterocysts, and fixes nitrogen during the day despite the lack of any known segregation strategy (Capone et al. 1997).

Biological nitrogen fixation is an essential component of the marine nitrogen biogeochemical cycle. In order to fully understand this and other important marine biogeochemical cycles, such as the phosphorus and iron cycles, it is necessary to understand the role of bacteria in transferring abiotic material into communities. However, studying marine microbes by traditional microbiological approaches is hampered by at least two major difficulties. First, most marine bacteria have not been cultivated (although some progress is being made in this direction; see for example Markou 2011), and therefore cannot be studied *in vitro*. Second, it is difficult to identify microorganisms visually and microbial morphology is too simple to be a sound basis for phylogenetic classification

(Olsen et al., 1986; Amann et al., 1990). The shortcomings of traditional investigative tools prompted the application of techniques from molecular biology to problems of microbial oceanography. In 1987, Carl Woese used sequences of the 16S ribosomal RNA (rRNA) subunit to construct a phylogenetic tree of life that revealed 3 top-level "ur-kingdoms" that he termed Archaebacteria, Eubacteria, and Eukaryota (Woese 1987; Woese, Kandler, & Wheelis, 1990). 16S rRNA continues to be a useful basis for general phylogeny and identification, while sequences of other genes are used to investigate specific metabolic processes, including *nasA* for nitrate assimilation (Allen et al., 2001), *nifH* for nitrogen fixation (Zehr & McReynolds, 1989), *nirS* for denitrification (Ward et al., 2007), and *nosZ* for nitrous oxide reduction (Jones et al., 2013). Today molecular techniques are applied to at least three kinds of studies: studies that focus on a single gene, for example to assess community potential to carry out metabolic processes associated with the gene(refs); metagenomic studies, which analyze all DNA sequences from an environmental sample; and transcriptome/metatranscriptome studies, which analyze gene expression of cultures or environmental populations and communities.

With the increased application of molecular methods and decreasing nucleic acid sequencing costs, data collection has outpaced analysis in all three kinds of studies mentioned above. Analysis is challenging because of the size and complexity of experimental results, and computational methods are necessary to determine relationships within large molecular biology data sets. This thesis presents three computation-based studies that resulted in important discoveries in single-gene genomics, metagenomics, and transcriptomics data, by applying a combination of existing and novel bioinformatics

techniques. The first project, described below in the section titled "*nifH* sequence retrieval and curation", involved *nifH,* the gene that encodes the two identical subunits of the iron protein of nitrogenase. Accelerating accumulation of sequences of this gene in the GenBank database has made manual retrieval and curation of such sequences intractable and impeded advancement of *nifH* diversity and phylogeny studies. To address this problem, a novel algorithm was developed and implemented in a software tool called ARBitrator. The second project, described below in the section titled "Metagenomics of UCYN-A2", involved assembly from metagenomic data of the genome of a new strain of a symbiotic diazotrophic cyanobacterium called UCYN-A, which may be a globally important nitrogen-fixer. This work provides a basis for comparing the genomes of the two strains; future study of new strains may benefit from the bioinformatic techniques developed for the assembly and analysis. The third project, described below in the section titled "Transcriptomics and diel expression of cyanobacteria", involved time series analysis of cyanobacterial transcriptome data. Cyanobacteria, whose metabolism is tightly coupled to the daily cycle of sunlight, have evolved complex daily patterns of gene expression. Although several time-series expression studies have been performed on cyanobacteria in axenic cultures, single-organism and comparative analysis are challenging because of a lack of computational tools. A Java application called Dexter was developed and applied to compare and analyze cyanobacterial gene expression time series datasets.

nifH sequence retrieval and curation

The identification in 1989 by Zehr and McReynolds of the role of *Trichodesmium* in fixing

nitrogen in *Trichodesmium* aggregates was carried out using degenerate PCR primers that targeted a broad variety of *nifH* sequences (Zehr & McReynolds, 1989). *nifH en*codes the iron protein , a component of nitrogenase, the enzyme that catalyzes reduction of molecular nitrogen to ammonia, and is an effective molecular proxy for environmental nitrogen fixation (Zehr et al., 2007). The success of this work led to reuse of the primers and discovery of *nifH* genes in diverse marine and non-marine habitats, including woody dicotyledenous plants (Simonet 1991), rice roots (Ueda, et al., 1995), termite guts (Ohkuma 1996), stromatolites (Steppe et al., 1996), central ocean gyres (Zehr et al., 1998), and salt marshes (Lovell et al., 2000). Many newly discovered genes were sequenced, and the sequences were deposited in the GenBank database (Benson et al., 2004; Benson et al., 2011), allowing cross-study phylogenetic analysis. *nifH* phylogeny is important because diazotrophs are distributed broadly but sparsely throughout the bacterial and archaeal domains (Young 2005; Gruber & Galloway, 2008); therefore organism phylogeny is not a useful predictor of the capacity to fix nitrogen.

nifH sequences in the GenBank database have been accumulating at an accelerating pace for two decades, causing a curation challenge that is no longer tractable except by computational approaches. In 1994, the database contained 19 *nifH* sequences (Chien & Zinder, 1994); this rose to 100 in 1998. A census of *nifH* diversity initiated in 2007 identified over 17,000 *nifH* sequences in GenBank (Gaby & Buckley, 2011); in 2011 there were over 30,000, with approximately 250 additional sequences deposited monthly. Diversity studies and sequence collections had become hampered by the inability to efficiently collect all *nifH* sequences from GenBank. Chapter 1 of this thesis describes a

software pipeline that rapidly and automatically retrieves and curates *nifH* sequences from GenBank with high sensitivity and specificity, thus enabling ongoing phylogeny and diversity studies.

Metagenomics of UCYN-A2

Trichodesmium has long been recognized as an important marine nitrogen fixer (Capone et al., 1997). However, estimates of nitrogen fixation by *Trichodesmium* amount to only 25% - 50% of geochemically derived rates (Mahaffey et al., 2005; Capone et al., 2005). This discrepancy has motivated the search for additional significant marine diazotrophs.

In 1998 (Zehr et al., 1998) and 2000 (Zehr et al., 2000), diverse diazotroph populations were discovered by *nifH gene* amplification at station ALOHA, 100 km north of Oahu. Phylogenetic analysis of *nifH* genes revealed two sub-clusters, referred to as Group A and Group B. Group B was identified to be *Crocosphaera watsonii*, a previously cultivated marine cyanobacterium. Group A represented a previously unknown clade, commonly called "UCYN-A" (for "unicellular cyanobacteria nitrogen-fixing, Group A") after nomenclature introduced by Mazard (Mazard et al., 2004) to describe a clade of unicellular nitrogen fixers detected by 16S ribosomal DNA phylogeny; UCYN-A was recently given the temporary name *Candidatus* Atelocyanobacterium thalassa. Cells from water samples were sorted by flow cytometry to enrich for UCYN-A cells and then the DNA was sequenced (Zehr et al., 2008) and later assembled (Tripp et al., 2010) to reveal a genome that lacked major metabolic components, including photosystem II. This unusual genomic

streamlining implied that UCYN-A is a symbiont receiving photosystem II products in exchange for fixed nitrogen. The symbiont's host has been identified as *Braarudosphaera bigelowii* (Thompson et al., 2012), a nanoplankton.

Research in the Pacific Ocean (Montoya 2004; Moisander et al., 2010) suggested that UCYN-A may be as least as important in nitrogen fixation as *Trichodesmium* and prompted the search for additional UCYN-A genomes from different habitats. Chapter 2 of this thesis describes the assembly and analysis of the metagenome of a second strain of UCYN-A from coastal waters off Scripps Oceanic Institute in La Jolla, California, an environment substantially different from that at Station ALOHA (Chavez 2002). As with the original strain, water was partially purified by flow cytometry and paired-end sequencing was performed prior to metagenome assembly. Assembly revealed that the two strains have 97% of their coding genes in common and a high degree of synteny, but have surprisingly low nucleotide and amino acid similarity. The availability of the genome of a second strain provides a basis for investigating how genomic differences between strains may reflect adaptation to different habitats, while the bioinformatic techniques developed for the assembly and analysis are likely to prove of value in the study of additional UCYN-A strains.

Transcriptomics and diel expression of cyanobacteria

Whereas genomics and metagenomics measure the metabolic potential of organisms and communities, transcriptomics and metatranscriptomics measure the expression level of

genes. Time series transcriptomic and metatranscriptomic data can reveal the temporal dynamics of metabolic processes in organisms and communities.

Representatives of all 3 domains of life have been observed to express genes in cycles that repeat every 24 hours, controlled by external cues such as light or temperature or by endogenous circadian clocks. Cyanobacteria can be expected to exhibit especially prominent and complex daily periodic expression patterns, as their lifestyles are coordinated to the availability of sunlight; this may be especially true in the case of diazotrophic cyanobacteria like *Crocosphaera* that temporally separate photosynthesis from nitrogen fixation. Time-series expression studies of cyanobacteria in culture are available for a few organisms, including *Crocosphaera watsonii* WH8501 (Shi 2010), *Cyanothece* sp. Strain ATCC 51142 (Stoeckel 2008; Toepel 2008), *Prochlorococcus marinus* MED4 (Zinser 2009), and *Trichodesmium erythraeum* IMS101 (I. Shilova and J. Zehr, unpub. data).Analysis of diel expression patterns appears to support the hypothesis that genes with similar expression patterns have similar function. Few community-wide environmental time-series expression studies have been performed; one recent metatranscriptome time-series study (Otteson 2014) revealed intriguing waves of coordinated expression patterns that begin with cyanobacterial primary producers and propagate to heterotrophic bacteria.

Systematic mining of time series expression data sets, and comparison of data sets for different organisms or ecotypes, is challenging because of a lack of software tools for visualization, exploration, and analysis; the problem is exacerbated by variations in

experimental design, which make comparison of data sets difficult. Chapter 3 of this thesis presents a novel software tool called Dexter that facilitates visualization, exploration, and analysis of multiple time-series expression data sets. The value of the program was demonstrated by using it to explore gene expression similarity within operons and to improve operon predictions in three cyanobacterial strains.

References

Allen, A. E., Booth, M. G., Frischer, M. E., Verity, P. G., Zehr, J. P., & Zani, S. (2001). Diversity and detection of nitrate assimilation genes in marine bacteria. *Applied and Environmental Microbiology, 67*(11), 5343–5348. doi:10.1128/AEM.67.11.5343-5348.2001

Allison, F., Lidwig, C.A., Hoover, S.R., & Minor, F.W. Biochemical nitrogen fixation studies. I. Evidence for limited oxygen supply within the nodule. (1940). *The Botanical Gazette*, March 1940, 513-533.

Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., & Stahl, D. A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology,*, *56*(6), 1919–1925.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2004). GenBank: update. *Nucleic Acids Research*, *32*(Database issue), D23–6. doi:10.1093/nar/gkh045

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2011). GenBank. *Nucleic Acids Research*, *39*(Database issue), D32–7. doi:10.1093/nar/gkq1079

Bergman, B., & Carpenter, E. J. (1991). Nitrogenase confined to randomly distributed trichomes in the marine cyanobacterium *Trichodesmium Thiebautii*. *Journal of Phycology*, *27*(2), 158–165. doi:10.1111/j.0022-3646.1991.00158.x

Bombar, D., Heller, P., Sánchez-Baracaldo, P., Carter, B. J., & Zehr, J. P. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *The ISME Journal*, 1–13. doi:10.1038/ismej.2014.167

Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B., & Carpenter, E. J. (1997).

Trichodesmium, a Globally Significant Marine Cyanobacterium. *Science*, *276*(5316), 1221–1229. doi:10.1126/science.276.5316.1221

Capone, D. G., Burns, J. A., Montoya, J. P., Subramaniam, A., Mahaffey, C., Gunderson, T., et al. (2005). Nitrogen fixation by Trichodesmium spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochemical Cycles*, *19*(2), n/a–n/a. doi:10.1029/2004GB002331

Carr, M-E, Friedrichs, M.A.M., Schmeltz, M., Aita, M.N., Antoine, D., Arrigo, K.R., et al. (2006). A comparison of global estimates of marine primary production from ocean color. *Deep-Sea Research II* 53 (2006) 741-770.

Chavez FP, Pennington JT, Castro CG, Ryan JP, Michisaki RM, Schlining B et al. (2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. Progr Oceanogr 54: 205–232.

Chien, Y., & Zinder, S. (1994). Cloning, DNA sequencing, and characterization of a nifD-homologous gene from the archaeon Methanosarcina barkeri 227 which resembles nifD1 from the eubacterium Clostridium pasteurianum. *Journal of Bacteriology*, *176*(21), 6590.

Delong, E. F. (2009). The microbial ocean from genomes to biomes. *Nature*, *459*(7244), 200–206. doi:10.1038/nature08059

Falkowski, P. (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. *Nature, 387,* 272-275.

Fay, P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiological Reviews*, *56*(2), 340–373.

Field, C.B., Behrenfeld, M.J., Renderson, J.T., & Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science (281)* 10 July 1988, 237-240.

Fleming, H., & Haselkorn, R. (1973). Differentiation in Nostoc muscorum: nitrogenase is synthesized in heterocysts. *Proceedings of the National Academy of Sciences*, *70*(10), 2727–2731.

Gaby, J. C., & Buckley, D. H. (2011). A global census of nitrogenase diversity. *Environmental Microbiology*, *13*(7), 1790–1799. doi:10.1111/j.1462- 2920.2011.02488.x

Gruber, N., & Galloway, J. N. (2008). An Earth-system perspective of the global nitrogen cycle. *Nature*, *451*(7176), 293–296. doi:10.1038/nature06592

Igarashi, R. Y. (2003). Nitrogen Fixation: The Mechanism of the Mo-Dependent

Nitrogenase. *Critical Reviews in Biochemistry and Molecular Biology*, *38*(4), 351– 384. doi:10.1080/10409230390242380

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., & Chisholm, S. W. (2006). Niche partitioning among Prochlorococcus ecotypes along ocean- scale environmental gradients. *Science*, *311*(5768), 1737–1740. doi:10.1126/science.1118052

Jones, C.M., Graf, D.R.H., Bru, D., Philippot, L., and Hallin, S. (2013) The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *ISME Journal* 7: 417-426.

Karl, D., Michaels, A., Bergman, B., Capone, D., Carpenter, E., Letelier, R., et al. (2002). Dinitrogen fixation in the world's oceans. In E. W. Boyer & R. W. Howarth, *The Nitrogen Cycle at Regional to Global Scales* (pp. 47–98). Dordrecht: Springer Netherlands. doi:10.1007/978-94-017-3405-9_2

Longhurst, A., Sathyendranath, S., Platt, T., & Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research (17)* 6:1245-1271.

Lovell, C. R., Piceno, Y. M., Quattro, J. M., & Bagwell, C. E. (2000). Molecular analysis of diazotroph diversity in the rhizosphere of the smooth cordgrass, Spartina alterniflora. *Applied and Environmental Microbiology*,, *66*(9), 3814–3822. doi:10.1128/AEM.66.9.3814-3822.2000

Mahaffey, C., Michaels, A. F., & Capone, D. G. (2005). The conundrum of marine N2 fixation. *American Journal of Science*, *305*(6-8), 546–595. doi:10.2475/ajs.305.6- 8.546

Markou, G. & Georgakakis, D. (2011). Cultivation of filamentous cyanobacteria (blue-green algae) in agro-industrial wastes and wastewaters: A review. *Applied Energy 88* (2011) 3389–3401.

Mazard, S.L., Fuller, N.J., Orcutt, K.M., Bridle, O., & Scanlan, D.J. (2004). PCR Analysis of the Distribution of Unicellular Cyanobacterial Diazotrophs in the Arabian Sea. *Applied and Environmental Microbiology* 2004, 70(12): 7355.

Mohr, Wiebke, Tobias Großkopf, Douglas W R Wallace, & Julie LaRoche. (2010). Methodological Underestimation of Oceanic Nitrogen Fixation Rates. Edited by Zoe Finkel. *PLoS ONE* 5 (9). Public Library of Science: e12583. doi:10.1371/journal.pone.0012583.

Moisander, P. H., Beinart, R. A., Hewson, I., White, A. E., Johnson, K. S., Carlson, C. A., et al. (2010). Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. *Science*, *327*(5972), 1512–1514. doi:10.1126/science.1185468

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986). Microbial

ecology and evolution: a ribosomal RNA approach. *Annual Review of Microbiology*, *40*(1), 337–365. doi:10.1146/annurev.mi.40.100186.002005

Otteson, E.A., Young, C.R., Gifford, S.M., Eppley, J.M., Marin, R., Schuster, S., Scholin, C., & DeLong, E. F. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345(6193) 207-212.

Parker, C.A., & Scutt, P.B. The effect of oxygen on nitrogen fixation by Azotobacter (1960). *Biochimica et Biophysica Acta 38* (1960) 230-238.

Partensky, F., Hess, W. R., & Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews : MMBR*, *63*(1), 106–127.

Peters, J. W., Fisher, K., & Dean, D. R. (1995). Nitrogenase structure and function: a biochemical-genetic perspective. *Annual Reviews in Microbiology*, *49*(1), 335– 366. doi:10.1146/annurev.mi.49.100195.002003

Raymond, J., Siefert, J. L., Staples, C. R., & Blankenship, R. E. (2004). The natural history of nitrogen fixation. *Molecular Biology and Evolution*, *21*(3), 541–554. doi:10.1093/molbev/msh047

Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., et al. (2009). Ecological genomics of marine picocyanobacteria. *Microbiology and Molecular Biology Reviews : MMBR*, *73*(2), 249–299. doi:10.1128/MMBR.00035- 08

Sherman, L.A., Meunier, P., Colon-Lopez, M.S. (1998). Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynthesis Research (58)*: 25-42.

Shi, T., Ilikchyan, I., Rabouille, S., & Zehr, J. P. (2010). Genome-wide analysis of diel gene expression in the unicellular N2-fixing cyanobacterium Crocosphaera watsonii WH 8501. *The ISME Journal*, 1–12. doi:10.1038/ismej.2009.148

Stal, L. J. (2009). Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? *Environmental Microbiology*, *11*(7), 1632–1645. doi:10.1111/j.1758-2229.2009.00016.x

Steppe, T. F., Olson, J. B., Paerl, H. W., Litaker, R. W., & Belnap, J. (1996). Consortial N2 fixation: a strategy for meeting nitrogen requirements of marine and terrestrial cyanobacterial mats. *FEMS Microbiology Ecology*, *21*(3), 149–156. doi:10.1111/j.1574-6941.1996.tb00342.x

Stocker, R. (2012). Marine microbes see a sea of gradients. *Science*, *338*(6107), 628– 633. doi:10.1126/science.1208929

Thompson, A. W., Foster, R. A., Krupke, A., Carter, B. J., Musat, N., Vaulot, D., et al. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science*, *337*(6101), 1546–1550. doi:10.1126/science.1222700

Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., et al. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, *464*(7285), 90–94. doi:10.1038/nature08786

Ueda, T., Suga, Y., Yahiro, N., & Matsuguchi, T. (1995). Remarkable N2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of nifH gene sequences. *Journal of Bacteriology*, *177*(5), 1414–1417.

Ward, B. B., Capone, D. G., & Zehr, J. P. (2007). What's new in the nitrogen cycle? *Oceanography-Washington Dc- ....*

Whitman, W., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578–6583.

Woese, C. (1987). Bacterial Evolution. *Microbiological Reviews, 0146-0729/87/020221-51*, 221-271.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, *87*(12), 4576–4579.

Zehr, J. P., & McReynolds, L. A. (1989). Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii. *Applied and Environmental Microbiology,*, *55*(10), 2522–2526.

Zehr, J.P., Montoya, J.P., Jenkins, B.D., Hewson, I., Mondragon, E., Short, C., et al. (2007). Experiments linking nitrogenase gene expression to nitrogen fixation in the North Pacific subtropical gyre. *Limnology and Oceanography (52)*1, 2007, 169-183.

Zehr, J. P., Bench, S. R., Carter, B. J., Hewson, I., Niazi, F., Shi, T., et al. (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science*, *322*(5904), 1110–1112. doi:10.1126/science.1165340

Zehr, J. P., Mellon, M. T., & Zani, S. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of Nitrogenase (nifH) genes. *Applied and Environmental Microbiology*, *64*(9), 3444–3450.

Zehr, J.P., Carpenter, E.J., & Villareal, T.A. (2000). New perspectives on nitrogen-fixing microorganisms in tropical and subtropical oceans. *Trends in Microbiology 8(2)*, 68-73.

Zehr, J., Jenkins, B., Short, S., & Steward, G. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environmental Microbiology*, *5*(7), 539–554.

Chapter 1: ARBitrator: A software pipeline for on-demand retrieval of auto-curated *nifH* sequences from GenBank[1]

[1]This chapter has been published as: Heller, P., Tripp, H.J., Turk-Kubo, K., and Zehr, J.P. (2014). ARBitrator: a software pipeline for on-demand retrieval of auto-curated *nifH* sequences from GenBank. Bioinformatics 2014 30 (20): 2883-2890 doi:10.1093/bioinformatics/btu417

Abstract

Motivation: Studies of the biochemical functions and activities of uncultivated microorganisms in the environment require analysis of DNA sequences for phylogenetic characterization and for the development of sequence-based assays for the detection of microorganisms. The numbers of sequences for genes that are indicators of environmentally important functions such as nitrogen (N_2) fixation have been rapidly growing over the past few decades. Obtaining these sequences from the National Center for Biotechnology Information's GenBank database is problematic because of annotation errors, nomenclature variation, and paralogues; moreover, GenBank's structure and tools are not conducive to searching solely by function. For some genes, such as the *nifH* gene commonly used to assess community potential for N_2 fixation, manual collection and curation are becoming intractable due to the large number of sequences in GenBank and the large number of highly similar paralogues. If analysis is to keep pace with sequence discovery, an automated retrieval and curation system is necessary.

Results: ARBitrator uses a two-step process comprised of a broad collection of potential homologues followed by screening with a best-hit strategy to conserved domains. 34,420 *nifH* sequences were identified in GenBank as of Nov 20, 2012. The false-positive rate is approximately 0.033%. ARBitrator rapidly updates a public *nifH* sequence database, and we show that it can be adapted for other genes.

Availability and Implementation: Java source and executable code are freely

available to non-commercial users at

http://pmc.ucsc.edu/~wwwzehr/research/database/.

Introduction

Microorganisms catalyze a variety of biogeochemical transformations, such as nitrogen (N_2) fixation, that are critical for ecosytem function. Since many microorganisms in the environment have not been cultivated, studies of such functions depend on amplification and sequencing of genes that encode proteins involved in the functions of interest (Zehr *et al*., 1995; Ueda *et al*., 1995; Zehr *et al*., 2003). This approach facilitates detection and phylogenetic characterization of uncultivated microorganisms.

Dinitrogen (N_2) is the most abundant gas in the atmosphere, but is not bioavailable unless it is reduced to ammonia by N_2 fixation. Biological N_2 fixation can require the protein products of up to 20 *nif* genes. The *nifH*, *nifD*, and *nifK* gene products are the structural components of the conventional molybdenum (Mo)-containing nitrogenase enzyme (Igarashi, 2003). The products of other *nif* genes are involved in roles such as regulation and biosynthesis (Rubio and Ludden, 2008). In environmental studies *nifH*, which encodes the Fe protein of Mo, V and Fe nitrogenases, is the most commonly utilized nitrogenase gene for the investigation of microorganisms with the potential to fix N_2, since it is the most highly conserved in sequence (Young, 1992). This gene can be used to examine the diversity of N_2-fixing microorganisms in the environment,

provides insight into the evolution and ecology of N_2 fixation, and can indicate the potential for N_2 fixation in microbial communities (Zehr and Capone, 1996; Lovell *et al*., 2001).

The application of degenerate *nifH* PCR primers (Zehr and McReynolds, 1989) enabled the discovery of novel *nifH* sequences in the environment. This approach fueled studies of N_2 fixation across a broad range of habitats, including woody dicotyledenous plants (Simonet *et al*., 1991), rice roots (Ueda et al., 1995), termite guts (Ohkuma *et al*., 1996), stromatolites (Steppe *et al*., 1996), central ocean gyres (Zehr *et al*., 1998), and salt marshes (Lovell *et al*., 2001). The size of phylogenetic trees based on all available nucleotide and amino acid sequences expanded from 19 sequences in 1994 (Chien and Zinder, 1994) to approximately 100 sequences in 1997 (Zehr *et al*., 1997) and to approximately 1500 sequences in 2003 (Zehr *et al*., 2003). Four (or five, depending upon author) major phylogenetic clusters have been described (Chien and Zinder, 1994; Zehr *et al*., 2003; Raymond *et al*., 2004; Gaby and Buckley, 2011). Only 3 of these phylogenetically-related clusters contain true nitrogenase-encoding *nif* genes. Cluster I *nifH* primarily comprises "conventional" *nifH*, which encodes the Fe protein of Mo nitrogenase (Igarashi, 2003), as well as *vnfH* genes that encode the Fe protein of vanadium (V) nitrogenase (note that vanadium nitrogenase genes cluster differently based on *nifD* or *nifK* genes)(Raymond *et al*., 2004). Organisms that contain Cluster 1 *nifH* genes include cyanobacteria and alpha-, beta-, and gamma-proteobacteria. Cluster II *nifH* genes

encode the Fe protein of "alternative" nitrogenases that contain iron but do not contain Mo or V (Lehman and Roberts, 1991; Joerger *et al*., 1989). Cluster III is dominated by genes encoding Fe proteins of nitrogenases primarily of anaerobes, including methanogens and sulfate reducers; these nitrogenases likely contain Mo. Clusters IV and V (sometimes grouped as Cluster IV) contains *nifH* paralogues whose functions include photopigment biosynthesis (Young, 2005) and non-N_2- fixation electron transport (Raymond *et al*., 2004). It has also been suggested that the function of Cluster IV *nifH* paralogues found in non-N_2-fixing Archaea is the biosynthesis of cofactor F430, essential to the production of methane (Staples *et al*., 2007; Boyd *et al*., 2011). The gene name *nflH*, for "*nifH*-like", has been proposed for these *nifH* paralogues (Staples *et al*., 2007).

Since sequences are accumulating rapidly in gene sequence databases, ongoing efforts to retrieve and analyze new *nifH* records are necessary in order to elucidate relationships between phylogenetic categories and identify phylotypes in different ecosystems. These efforts can be frustrated by the sheer number and growth rate of *nifH* gene sequences deposited into the National Center for Biotechnology Information (NCBI) GenBank database (Benson, 2004). For example, in February 2009, Gaby and Buckley identified *nifH* sequences from the nonredundant nucleotide collection database (nr/nt) at GenBank for a global census of nitrogenase diversity (Gaby and Buckley, 2011) and records were manually curated to form a database of approximately 17,000 sequences. We estimate that in the intervening time between the download and publication of the related article in 2011, at least 10,000 *nifH*

sequences were added to the database. We estimate that as of January 2012 there were over 32,000 *nifH* sequences in that database, with a growth rate of over 300 sequences per month. Retrieval involving human intervention now requires a significant investment of manpower that will increase over time; an automated retrieval pipeline is needed to allow analysis to keep pace with data collection. However, automating retrieval of all sequences of any specific gene from GenBank is difficult. Moreover, the most common query idioms for searching GenBank are variants of BLAST (Altschul et al., 1990), which searches for sequence similarity rather than function. Approaches based on BLAST alone are likely to be overly sensitive, as hits to *nifH* homologues with functions other than N_2 fixation are not easily distinguished from genuine *nifH* hits. Text searches that analyze sequence annotation can be misled by misannotation (Tripp *et al*., 2011) or misspellings in the annotation fields; a text search for "nifh" in the nr protein database found only 6,173 sequences, of which we believe 527 are not actually *nifH*. The Fungene database (http://fungene.cme.msu.edu) provides a repository of collections of functional genes from GenBank, classified by hidden Markov models (Krogh *et al*., 1994; Eddy, 1998); however, updates from GenBank are infrequent and the hidden Markov model approach is prone to false calls. An alternative nifH database available to the public, maintained at Cornell (Gaby and Buckley, 2001; Gaby and Buckley, 2014), requires the manual retrieval and curation of Genbank *nifH* sequences.

To resolve these issues, a software pipeline was developed, called ARBitrator, that requires little human intervention and retrieves up-to-date *nifH* sequence collections

within a few hours. The software is adaptable to collecting sequences for genes other than *nifH*, and is especially helpful for discriminating genes of interest from their paralogues, since it incorporates an auto-curation feature based upon best reversePSI BLAST hits to GenBank's Conserved Domain Database. ARBitrator's output is formatted for input into other programs, such as the ARB phylogenetic software environment (Ludwig et al., 2004). Supplementary Appendix A provides a procedure for incrementally updating an existing *nifH* ARB database using the output of ARBitrator, thus facilitating maintenance of a comprehensive updated gene database.

System and Methods

Design Criteria

In order to retrieve sequences for a single gene (in this case the *nifH* gene) from a large public database, and to facilitate maintenance of a sequence database for that gene, a pipeline needs to meet specific design criteria. Based on experience with maintaining a public database of *nifH* sequences (*http://pmc.ucsc.edu/~wwwzehr/research/database/*), the following requirements were identified:

(1) The pipeline should be easy to invoke, and should require no manual setup or runtime intervention. In particular, the pipeline should use the public on-line GenBank database and services provided by NCBI, rather than using a local copy of the database which would need to be downloaded and updated prior to pipeline

execution.

(2) The data should not require manual curation. Given the current size of the GenBank database and the rate at which new *nifH* sequences are submitted, any step involving even trivial manual inspection of individual records would introduce excessive delays and possible errors.

(3) The pipeline should have acceptable error rates. Automated classification systems have inherent error rates which must be controlled to within acceptable tolerances. In the case of *nifH* classification, sequences with strong similarity to *nifH* but different function might be accepted on the basis of homology (false positives). Similarly, sequences with *nifH* function that have strong similarity to non-*nifH* genes might be rejected (false negatives). Both kinds of error must be minimized.

(4) Sequence identification should not rely on annotations in the database. Although annotations are useful as quality control checks (e.g. for determining false positive/negative rates), classification should be based only on sequence content. A solution based on annotations could be no better than the false annotation rate, which is unknown but may be too high for reliable classification. Moreover, any such solution would complicate the software by requiring natural language processing.

(5) The output of the pipeline should contain all metadata associated with the identified sequences, presented in a standard format. The pipeline's results should be easy to analyze. We have traditionally used ARB (Ludwig et al., 2004) to build

phylogenetic trees of *nifH* sequences; therefore the pipeline must produce output in EMBL format (Kanz, 2004), which ARB is able to read. All annotation information and metadata should be included in the EMBL records.

(6) The pipeline should be adaptable to the retrieval of genes other than *nifH*. To this end, code should be open-source under standard licenses, and adaptation should not require a high level of programming expertise or technical support.

Algorithm

ARBitrator evaluates candidate sequences on two criteria, which we designate "quality" and "superiority". Quality measures a candidate sequence's similarity to the most similar member of a representative set of 15 *nifH* sequences; superiority measures the degree to which a candidate is more similar to *nifH* than to genes coding for other known proteins. ARBitrator first executes a sensitivity phase in which it collects candidates based on quality (Figure 1). Initial investigations showed that this phase is effective at finding *nifH* sequences; however, many non-*nifH* (false positive) sequences are also collected. In the subsequent specificity phase, false positive candidates are eliminated based on superiority. In a final formatting phase, nucleotide sequences and annotations are retrieved for each accepted sequence and an output file is generated in EMBL format.

Implementation

In the sensitivity phase, a set of *nifH* protein sequences is BLASTed against the nr

database at GenBank using the blastp (protein query, protein subject) program. Given the large number of known *nifH* sequences (over 15,000 at the inception of this project), a "representative set" was selected consisting of 15 sequences that are evenly distributed throughout the *nifH* phylogenetic tree (Figure 2). By only BLASTing these representative sequences, rather than all known sequences, time spent in this phase of the algorithm is reduced by 3 orders of magnitude. For all candidate sequences (hits) retrieved, we define "quality" as the negative $\log_{10}$ of the E-value of the hit. If a subject is hit by multiple queries from the representative set, then quality is defined as the negative $\log_{10}$ of the smallest E-value across all hits. The sensitivity phase accepts sequences with quality >= 2 (i.e. all E-values are <= .01). The output of this phase is a collection of GIs (GenInfo Identifiers).

To support the specificity phase, a database of conserved domains was constructed based on the total set of GenBank Conserved Domains in the Sub-family Hierarchy for cd01983, Fer4_Nifh. Candidate GIs that meet the sensitivity criteria are reverse-PSI BLASTed against this database. The reverse-PSI BLAST algorithm uses position-specific scoring (Marchler-Bauer et al., 2001; 2011), and is thus appropriate for conserved domain analysis since differences within a conserved domain are penalized more heavily than differences in non-conserved regions. The 3 best-scoring hits for each candidate are analyzed. Any hits to the cd02117 (NifH_like) conserved domain are discarded as uninformative; candidate sequences are accepted if the best remaining hit is to the cd02040 (NifH) conserved domain, and

the E-value of this hit is at least 10x smaller than the E-value of the next-best hit. We define a sequence's *superiority* as $\log_{10}$(E-value of best non-cd02040 hit) - $\log_{10}$(E-value of cd02040 hit). Thus, candidates are accepted if their superiority is >= 1. The output of the sensitivity phase is a subset of the protein GIs generated by the sensitivity phase, representing sequences that ARBitrator classifies as *nifH*. For analysis requiring only protein GIs, the pipeline may optionally be terminated at this point.

The formatting phase retrieves the nucleotide coding sequence and annotations associated with each protein GI, and builds a record in EMBL format (Kulikova et al., 2007). The GenBank protein record for each GI is retrieved from NCBI via the eUtils utility (http://www.ncbi.nlm.nih.gov/books/NBK25497/pdf/chapter2.pdf). This record is scanned for the "coded_by" subfield of the "CDS" field to extract the identifier of a nucleotide sequence – typically a complete genome – along with coordinates for the start and end of the protein sequence. A second eUtils query retrieves the nucleotide record, from which the *nifH* nucleotide sequence is extracted. The public-domain ReadSeq utility (http://iubio.bio.indiana.edu/soft/molbio/readseq/java/) converts this record to EMBL format. Lastly, for compatibility with ARB's input filtering, multi-line literal quote field values in the EMBL record are concatenated into single monolithic lines.

After execution of the pipeline, new records can be incorporated into an existing ARB database using a protocol described in Supplementary Appendix A. Briefly, new

records are imported into ARB using a custom import filter. All protein sequences are then exported and aligned to the NifH/FrxC protein family using a hidden Markov model. The aligned sequences are then imported into ARB. Nucleotide sequences are back-aligned to the protein alignment using the ARB backalign tool. The nucleotide and amino acid sequences can then be used for phylogenetic analysis and probe design.

Tuning

The software's classification algorithm is fine-tuned by three parameters: the representative set of verified *nifH* sequences used as the BLAST queries of the sensitivity phase, and the threshold values for quality and superiority. The representative set was chosen to broadly represent the known clades in the *nifH* phylogenetic tree. Members of the set are shown in Figure 2.

To determine threshold values for quality and superiority, positive and negative training sets were created. The positive training set contains the 15,513 *nifH* sequences in a manually curated *nifH* database (*http://pmc.ucsc.edu/~wwwzehr/research/database/)* as of September 1, 2010. The negative training set contains 766 sequences that are believed not to be *nifH* but that are moderately similar to *nifH*. Optimal quality and superiority thresholds were computed by exhaustive search of quality-superiority space. When quality threshold = 2 and superiority threshold = 1, miscalls are minimized in both training sets. These thresholds were verified by N-fold cross validation.

Error rates

To determine false positive and false negative rates, ARBitrator was executed immediately after the November 2012 run reported herein, with the quality threshold relaxed to zero and the superiority threshold relaxed to -10. This generated a sample set consisting of sequences that were accepted by ARBitrator as *nifH*, as well as sequences that were rejected by a small margin. Sequences were aligned to a Pfam-curated multiple alignment of the NifH/frxC family (Fer4_NifH; PF00142) using the HMMalign module of the HMMer software package (Finn *et. al.*,2011*)*. Sequences that were too short for reliable alignment were omitted from the error rate computation. To classify "accept" sequences, a neighbor-joining phylogenetic tree (Saitou and Nei, 1987) was constructed using sequences that had amino acid residues in the region most widely PCR amplified by *nifH* primers (Zehr *et.al.* 2003), as the majority of the sequences submitted to GenBank are from PCR amplification. Sequences with short branches within Clusters I through III were classified as *nifH*; sequences with short branches within Cluster IV (which contains *nifH* homologues and non-functional genes) could not be classified with confidence, and were omitted from the error-rate computation; sequences with long branches within Cluster IV were classified as not *nifH*. ARBitrator's false positive rate was computed as the number of "accept" sequences classified phylogenetically as not *nifH*, divided by the total number of "accept" sequences that could be aligned. A false positive rate for sequences reported as *nifH* by Fungene was computed by the same method. To classify "reject" sequences, any sequence that aligned poorly with the NifH/frxC

family model was classified as not *nifH*. Those that aligned well but did not have amino acid residues in the region most widely amplified by *nifH* primers were omitted from the error rate computation, as their phylogenetic analysis would not be reliable. A neighbor-joining tree was constructed using the remaining "rejects" (i.e. those whose alignments permitted phylogenetic analysis), and the *nifH* representative set sequences. Classification was performed as described above, except that sequences with short branches to Cluster IV were conservatively classified as *nifH* to determine an upper bound for the false positive rate.

To assess the reliability of annotations among *nifH* and similar sequences, annotated gene function was retrieved for all "accept" and "reject" sequences. Records whose annotated function was *nifH* or synonymous to *nifH* ("dinitrogenase reductase", "nitrogenase Fe protein", and 138 others) were classified as annotated as *nifH*. Phylogeny-based misannotation rates were computed using the approach described to determine error rates. Sequences whose annotation contradicted phylogenetic classification were blasted against the GenBank nr database to assess the likelihood of misannotation.

Extension beyond *nifH*

To test ARBitrator's effectiveness on genes other than *nifH*, the pipeline was configured to retrieve sequences of the *nifD* gene (that encodes the alpha subunit of the Mo, V and Fe nitrogenase proteins). The representative sequence set consisted of the *nifD* sequences of the organisms that contributed to the *nifH* representative set, as

well as 3 sequences of *vnfD* (the vanadium-using form of *nifD*). A positive training set was built by running the pipeline with the *nifD* representative and the quality/superiority settings used for *nifH*, and hand-selecting 73 sequences from the results; 43 additional sequences were added from six published studies (Dedysh, 2004; Fani et al., 2000; Henson et al., 2004; Holmes, 2004; Parker et al., 2002; Rodríguez-Echeverría, 2010). A negative training set was built by selecting 82 sequences of five genes that are known to have high similarity to *nifD*: *nifE*, *nifH*, *nifK*, *nifN*, and protochlorophyllide reductase. Optimal quality and superiority thresholds were computed by exhaustive search of quality- superiority space, and the pipeline was executed with quality=8.1 and superiority=0.1. As with the *nifH* configuration, the thresholds were then relaxed to generate a sample set for computation of true- and false-negative rates.

Results

Sequences Retrieved on Nov 20, 2012

On November 20, 2012, ARBitrator returned 34,420 *nifH* sequences from GenBank, of which 1,757 were new to GenBank since the previous ARBitrator run on July 11, 2012. The list of protein sequence GIs may be downloaded from *http://pmc.ucsc.edu/~wwwzehr/research/database/*. Figure 3 shows a phylogenetic tree of representatives of the 34,420 sequences.

Figure 4 shows the distribution by quality and superiority of all sequences in the

sample set. Linear regression analysis of the accepted sequences (upper-right quadrant) shows a linear relationship with superiority = 1.3 + .45 x quality, and a coefficient of determination of .91. Most sequences accepted by ARBitrator cluster around the point (quality=63, superiority=28). Intriguingly, there is a cluster of rejected sequences around the point (quality=2, superiority=-5); these are predominantly annotated as septum-site determining proteins or cobyrinic acid a,c-diamide synthase (*cobB*).

Error Rates

Of 34,420 sequences in the sample set that are called *nifH* by ARBitrator (upper-right quadrant in Figure 4), 2,208 are too short for phylogenetic analysis and are omitted from the error-rate computation. 30,051 cluster with short branches with Clusters I, II, and III, and are confirmed as *nifH*. 2,151 sequences that cluster with short branches within Cluster IV cannot be confidently classified as *nifH* or not-*nifH*, and are omitted from the computation. 10 sequences associate via long branches with Cluster IV and are classified as not *nifH*. Thus the phylogenetically-derived false positive rate is 10 / (10 + 30,051) = 0.033%. Of the 32,227 sequences in the sample that are rejected by ARBitrator (upper-left, lower-left, and lower- right quadrants in Figure 4), 2,067 are too short for phylogenetic analysis and are omitted from the error-rate computation. 28,846 align poorly with the NifH/frxC family and are confirmed as not *nifH*. 1,134 cluster with long branches within Cluster IV, and are also confirmed as not *nifH*. 8 associate via short branches with Cluster IV and are omitted from the error-rate

computation. Thus the phylogenetic analysis detected no false negative errors. See Supplementary Figure S1 for phylogenetic trees of accepted and rejected sequences.

104 sequences that ARBitrator accepts are annotated as not *nifH*; 87 of these are confirmed by phylogeny as *nifH* (i.e. phylogeny confirms ARBitrator's call), 10 are classified by phylogeny as not *nifH*, and 10 cannot be phylogenetically classified. 88 sequences in the sample set that ARBitrator rejects are annotated as *nifH*; 82 of these are classified by phylogeny as not *nifH* (i.e. phylogeny confirms ARBitrator's call), 8 are classified by phylogeny as *nifH*, and 8 cannot be phylogenetically classified.

Comparison to other *nifH* databases

The sequences were compared to the 31,970 sequences in the Fungene database as of November 20, 2012. ARBitrator and Fungene had 29,836 sequences in common. ARBitrator collected 4,584 sequences that Fungene rejected, and Fungene collected 2,134 sequences that ARBitrator rejected; of these, ARBitrator rejected 1,744 due to quality and 390 due to superiority. For the sequences reported by both ARBitrator and Fungene, the phylogenetically derived false positive rate is .033%, the same as the overall ARBitrator false positive rate. For the 2,134 sequences that only Fungene reports as *nifH*, 1,964 did not align with the NifH/frxC family and are classified as not *nifH*. Thus the false positive rate for these sequences is >= 1,964/2,134 = 92%. The overall Fungene false positive rate is 6.2%.

The second update of the Cornell nifH database reported in Gaby and Buckley

(2014), based on a May 2012 snapshot of GenBank, contained 32,854 nucleotide records, 28,742 of which have corresponding protein records specified by a "db_xref" field thus permitting comparison to ARBitrator. ARBitrator rejected 479 of these records (1.7%): 440 due to low quality, and 39 due to low superiority.

nifD results

The *nifD* configuration of the pipeline returned 2,747 sequences, of which 2,726 are annotated as *nifD*, for a false positive rate of 0.76% conditioned on annotation accuracy. The sample set contained 8715 rejected sequences, of which 76 are annotated as *nifD,* for an annotation-conditioned false negative rate of 0.87%. For the 2,972 *nifD* sequences reported by Fungene for which unambiguous annotations could be retrieved, the annotation-conditioned false positive rate is 48%.

Discussion

nifH gene diversity studies, and indeed all single-gene diversity studies, are hampered by the difficulty of collecting all sequences associated with the gene of interest. NCBI's GenBank, the database where newly discovered sequences are deposited, provides no service for selecting all records that are annotated as representing a specified gene. A direct text search of the database would be of dubious value due to misannotations. GenBank's main query idiom is the BLAST search, which selects based on sequence similarity regardless of function. Consequently a sequence collection pipeline that simply BLASTs a representative set of query genes will

accept paralogues with the wrong function. The challenge of supporting diversity studies is exacerbated by the rapid growth of GenBank: the collection of *nifH*, and presumably other genes, appears to have been growing exponentially for the past several years. Thus there exists an opportunity to facilitate diversity studies by increasing the efficiency of sequence retrieval.

Necessity for Both Quality and Superiority Criteria

As Figure 4 shows, sequences obtained and accepted as *nifH* by ARBitrator (above and to the right of the purple crosshairs) cluster around the point (quality=63, superiority=28). When the quality and superiority criteria are relaxed (to the left of and/or below the crosshairs) a second cluster appears around quality=0.1, superiority=-5; sequences in this second cluster are predominantly annotated as *MinD* (membrane ATPase of the MinC-MinD-MinE system) or *CobB* (cobyrinic acid a-c diamide synthase, involved in the biosynthesis of vitamin B12). There is an approximately linear relationship between quality and superiority (superiority = 1.3 + .45 quality), with $r^2 = .91$. This relationship is not strong enough to allow either quality or superiority alone to be used as a selection criterion. For example, without the quality criterion, all sequences to the right of the vertical crosshair would be accepted, including many sequences from the *MinD*/cobyrinic acid peak. Similarly, without the superiority criterion, all sequences below the horizontal crosshair would be accepted.

Error Rates

ARBitrator's low error rates (.033% false positives, no detectable false negatives) can be understood in light of the underlying similarities between the ARBitrator algorithm and the error-rate analysis. In both approaches, candidate sequences are aligned against known *nifH* sequences. With ARBitrator, the candidate sequences are the contents of the GenBank protein database and the known *nifH* sequences are the 15 representative sequences; during the sensitivity phase, the BLAST step aligns each query (known *nifH* representative) against each database member (candidate). Candidates are provisionally accepted if they align well enough with the representatives, with the specificity phase providing necessary additional accuracy to the measurement of alignment quality. In the error-rate analysis, the candidate sequences are the members of the ARBitrator sample set, which are aligned against the known *nifH* members of the Fer4_NifH Pfam family. The alignment algorithms in the two approaches are not identical in all implementation details, but in both cases sequences are accepted if and only if they are similar to known *nifH* sequences, with similarity measured by alignment score.

When a sequence's annotation contradicts its ARBitrator classification (i.e. ARBitrator accepts a sequence that is not annotated as *nifH,* or rejects a sequence that is annotated as *nifH*), phylogenetic analysis supports the ARBitrator call in 90% of cases. 104 sequences are classified by ARBitrator as *nifH* but annotated as not *nifH*. When these were BLASTed against the nr database and the top 20 non-self hits for

each query were inspected, 48 queries hit exclusively to *nifH* subjects or to non-*nifH* subjects from the same study as the query. We propose that these subject sequences, which come from two studies that submitted multiple sequences to GenBank, were systematically misannotated by the researchers and should have been annotated as *nifH*. Similarly, ARBitrator rejects 88 sequences that are annotated as *nifH*. When these were blasted against the nr database and the top 20 hits of each query were inspected, it was found that 54 of the queries hit exclusively to non-*nifH* subjects, or to *nifH* subjects from the same study as the query. These subject sequences come from five studies which submitted multiple sequences to GenBank, and we propose that these sequences were systematically misannotated as *nifH*.

Comparison to other *nifH* databases

The ARBitrator and Fungene databases as of Nov 20, 2012 had 29,836 *nifH* sequences in common. ARBitrator accepted 4,584 sequences that Fungene rejected. The false positive rate for these sequences is approximately the same as for the overall sample set (.03%). Fungene accepted 2,134 sequences that ARBitrator rejected; the false positive rate for these sequences is 90%. This discrepancy can perhaps be attributed to the fundamental difference between ARBitrator's BLAST-based algorithm and the hidden Markov model that underlies the Fungene pipeline. ARBitrator classifies according to similarity to representative *nifH* sequences and to the NifH conserved domain; Fungene classifies according to similarity to a composite profile model.

ARBitrator's results are substantially in agreement with the latest update of the Buckley Lab database, which is not generated by auto-curating software and requires manual processing (Gaby and Buckley, 2014).

Extension beyond nifH

The *nifD* pipeline configuration, despite being based on cruder positive and negative training sets than the *nifH* configuration, nevertheless produced annotation-based error rates that were less than 1%. This result supports the applicability of the ARBitrator algorithm to other genes besides *nifH*.

The discrepancy between the superiority thresholds for *nifH* (1.0) and *nifD* (0.1) can be explained by the evolutionary history of *nifD*, which apparently originated as an ancestral gene that underwent a duplication event, giving rise to an ancestral bicistronic operon that later duplicated again to produce the present-day *nifD, nifE, nifK, and nifN* genes (Fani et al., 2000). Thus many *nifD* genes have low superiority because their similarity to the *nifE, nifK, and nifN* conserved domains is only slightly worse than their similarity to the *nifD* conserved domain. A higher superiority threshold would reject such sequences and increase the false negative rate. The similarity of *nifD* to other genes makes it a particu- larly rigorous test of the extension of the ARBitrator approach.

In conclusion, by combining a quality-based sensitivity phase with a superiority-based specificity phase, we have been able to implement a pipeline that meets all

design criteria. Records provide nucleotide and amino acid sequences as well as complete annotations. Computed false-positive and false-negative rates are acceptably low, and actual rates may be even lower. Results from the Nov 20, 2012 run are generally in good agreement with the *nifH* sequence collection at Fungene; however, the ARBitrator results are more extensive, have a lower error rate, and can be updated whenever a user wishes to rerun the pipeline. ARBitrator is designed to support ongoing *nifH* phylogeny research into the future as the GenBank collection continues to grow exponentially. If error rates increase, adjustments can be made to the quality and superiority thresholds. If NifH/FrxC family and CobB sequences continue to be major contributors to the false positive rate, the software may need to be adapted to detect and reject these special cases.

In addition to its original application to *nifH* phylogeny, ARBitrator can be adapted to other genes, as evidenced by the *nifD* results. An immediate use for adapted versions of the pipeline would be the collection *nif* genes other than *nifH* and *nifD*. Comparison of phylogenies of multiple *nif* genes could provide new insight into N_2 fixation diversity and ecology.

Acknowledgements

The authors are grateful to Deniz Bombar for his help in organizing this article.

Funding: This work was supported by NSF grant EF0424599 for the Center for Microbial Oceanography: Research and Education (CMORE), a first phase Gordon

and Betty Moore Marine Investigator grant (J.Z.) and the Microbial Environmental

Genomics Applications: Modeling, Experimentation, and Remote Sensing

(MEGAMER) facility of the University of California, Santa Cruz.

Conflict of Interest: None declared.

References

Altschul S, *et al*. (1990). Basic Local Alignment Search Tool. J Mol Biol **215**, 403-410.

Benson DA. (2004). GenBank: update. *Nucleic Acids Res* **32**:23D–26.

Boyd, E.S. *et al*. (2011). An alternative path for the evolution of biological nitrogen fixation. *Frontiers in Microbiol* **2011**.00205.

Chien Y, Zinder S. (1994). Cloning, DNA sequencing, and characterization of a *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD1* from the eubacterium *Clostridium pasteurianum*. *J Bacteriol* **176**:6590.

Dedysh, S. N. (2004). NifH and NifD phylogenies: an evolutionary basis for understanding nitrogen fixation capabilities of methanotrophic bacteria. Microbiology, ***150*(5)**, 1301–1313.

Eddy SR. (1998). Profile hidden Markov models. Bioinformatics Review **9**:755-763.

Edgar RC, *et al*. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**:2194–2200.

Fani R, *et al*. (2000). Molecular evolution of nitrogen fixation: the evolutionary history of the *nifD*, *nifK*, *nifE*, and *nifN* genes. *J Mol Evol* **51**:1-11.

Finn, R. D., *et al*. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, ***39*(suppl)**, W29–W37. doi:10.1093/nar/gkr367

Gaby JC, and Buckley DH. (2011). A global census of nitrogenase diversity. *Environ Microbiol* **13**:1790–1799.

Gaby, J. C., and Buckley, D. H. (2014). A comprehensive aligned *nifH* gene database: a multipurpose tool for studies of nitrogen-fixing bacteria. *Database*, 2014.

Henson, B. J., *et al*. (2004). Molecular phylogeny of the heterocystous

cyanobacteria(subsections IV and V) based on *nifD*. *Inter J Syst Evol Microbiol* ***54***(2), 493–497.

Holmes, D. E. (2004). Comparison of 16S rRNA, *nifD*, *recA*, *gyrB*, *rpoB* and *fusA* genes within the family *Geobacteraceae* fam. nov. *Inter J Syst Evol Microbiol*, ***54***(5), 1591–1599.

Huang Y, *et al*. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**:680–682.

Igarashi RY. (2003). Nitrogen Fixation: The Mechanism of the Mo-Dependent Nitrogenase. *Critical Rev Biochem Mol Biol* **38**:351–384.

Joerger RD, *et al*. (1989). Two *nifA*-like genes required for expression of alternative nitrogenases by Azotobacter vinelandii. *J Bacteriol* **171**:3258–3267.

Kanz C. (2004). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **33**:D29–D33.

Krogh A, *et al*. (1994). Protein modeling using hidden Markov models. *J Mol Biol* **235**:1501–1531.

Kulikova T, *et al*. (2007). EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res* **35**:D16–D20.

Lahiri S, *et al*. (2008). NifH: Structural and Mechanistic Similarities with Proteins Involved in Diverse Biological Processes. *Am J Biochem Biotechnol* **4**:304.

Lehman LJ, and Roberts GP. (1991). Identification of an alternative nitrogenase system in Rhodospirillum rubrum. *J Bacteriol* **173**:5705–5711.

Lovell CR, *et al*. (2000). Molecular Analysis of Diazotroph Diversity in the Rhizosphere of the Smooth Cordgrass, *Spartina alterniflora*. *Appl Environ Microbiol* **66**:3814–3822

Lovell CR, *et al*. (2001). Recovery and Phylogenetic Analysis of *nifH* Sequences from Diazotrophic Bacteria Associated with Dead Aboveground Biomass of Spartina alterniflora. *Appl Environ Microbiol* **67**:5308–5314.

Ludwig W, *et al*. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**:1363–1371.

Marchler-Bauer A, *et al*. (2001). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* **30**:1–3.

Marchler-Bauer A, *et al*. (2011). CDD: a Conserved Domain Database for the

functional annotation of proteins. *Nucleic Acids Res* **39**:D225–D229.

Ohkuma M, *et al*. (1996). Diversity of nitrogen fixation genes in the symbiotic intestinal microflora of the termite *Reticulitermes speratus*. *Appl Environ Microbiol* **62**:2747–2752.

Parker, M. A., *et al*. (2002). Conflicting phylogeographic patterns in rRNA and *nifD* indicate regionally restricted gene transfer in *Bradyrhizobium*. *Microbiol* **148(Pt8)**, 2557–2565.

Raymond J, *et al*. (2004). The natural history of nitrogen fixation. *Mol Biol Evol* **21**:541–554.

Rodríguez-Echeverría, S. (2010). Rhizobial hitchhikers from Down Under: invasional meltdown in a plant-bacteria mutualism? *J Biogeogr* **37**:1611-1622

Rubio LM and Ludden PW. (2008). Biosynthesis of the iron-molybdenum cofactor of nitrogenase. *Annu Rev Microbiol* **62**:93–111.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* ***4*****(4)**, 406–425.

Simonet P, *et al*. (1991). Frankia genus-specific characterization by polymerase chain reaction. *Appl Environ Microbiol* **57**:3278–3286.

Staples, C.R. *et al*. (2007). Expression and Association of Group IV Nitrogenase NifD and NifH Homologs in the Non-Nitrogen-Fixing Archaeon *Methanocaldococcus jannaschii*. *J Bacteriol* **189(20)**:7392.

Steppe T, *et al* (1996). Consortial N_2 fixation: a strategy for meeting nitrogen requirements of marine and terrestrial cyanobacterial mats. *FEMS Microbiol Ecol* 149–156.

Tripp HJ, *et al*. (2011). Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res* **39**:8792–8802.

Ueda T, *et al*. (1995). Remarkable N2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences. *J Bacteriol* **177**:1414–1417.

Young, J. P. W. (1992). Phylogenetic classification of nitrogen-fixing organisms, p.43– 86. In *Biological nitrogen fixation*, G. Stacey, H. J. Evans, and R. H. Burris (ed.). Chapman and Hall, New York, N.Y.

Young J. (2005). The phylogeny and evolution of nitrogenases. In Palacios R,

Newton WE (eds). In *Genomes and Genomics of Nitrogen-Fixing Organisms*. Springer: Netherlands, pp 221–241.

Zehr JP and McReynolds LA. (1989). Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. *Appl Environ Microbiol* **55**:2522–2526.

Zehr J, *et al*. (1995). Diversity of heterotrophic nitrogen fixation genes in a marine cyanobacterial mat. *Appl Environ Microbiol* **61**:2527.

Zehr J and Capone D. (1996). Problems and Promises of Assaying the Genetic Potential for Nitrogen Fixation in the Marine Environment. *Microb Ecol* **32**:263–281.

Zehr JP, *et al*. (1997). Phylogeny of cyanobacterial *nifH* genes: evolutionary implications and potential applications to natural assemblages. *Microbiology* **143**:1443–1450.

Zehr J, *et al*. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. *Appl Environ Microbiol* **64**:3444.

Zehr J, *et al*. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**:539–554.

Supplementary Appendix A: Procedure for updating an existing ARB database with ARBitrator output

Following each ARBitrator run, the output is read into a new ARB database through a custom import filter. A custom field is created for all new sequences and populated with the date of the ARBitrator run. Sequences (or "Species" in ARB's terminology) from this new ARB database are merged into the previous ARB database version using the ARB "MERGE" function. Once in the updated database, the newly added nucleotide sequences are translated into amino acid sequences, after which the amino acid sequences from all entries in the merged database are exported for alignment using a custom export filter.

The exported nifH sequences are aligned to a Pfam-curated multiple alignment of the NifH/frxC family (Fer4_NifH; PF00142) using the HMMalign module of the HMMer software package. HMMAlign is preferred over more commonly used alignment tools such as the Clustal suite because hidden Markov models have been shown to provide the most consistent and reproducible alignments for functional genes. The start and end positions resulting from the HMM alignment output file (in A2M file format) are written to new ARB fields for each sequence using a perl script (MarkAlignPosOfA2mFile_ver2.pl), prior to importing the aligned sequences into ARB using a custom import filter, and merging with the updated ARB database.

ARB is then used to back-align nucleotide sequences according to the newly aligned amino acid data. Sequences that are unable to be back-aligned are marked and disregarded in all future steps, but very few sequences fall into this category. For example, in the December 2011 update, 122 of 30,591 (0.4%) total sequences were unable to be aligned.

In order to facilitate the user-friendly analysis of environmental nifH sequences, which now number in the 30,000s, clusters based on nucleotide and amino acid sequence similarity are determined using the CD-HIT suite (Huang et al. 2010). This requires the use of a custom export ARB filter, as well as a perl script (CDHIT_to_ARB_5dot2.pl) designed to format the CD-HIT and CD-HIT-EST output results for use in ARB. Once the cluster data is merged into the updated ARB database, a final manual transfer of user-defined NDS fields is required.

Putative chimeric sequences were first determined after the December 2010 update using the UCHIME algorithm (Edgar et al. 2011). Nucleotide alignments for all 22,579 sequences (Dec 2010 update; including those obtained from genomes) were clustered using CD-HIT-EST (Huang et al. 2010) at 98% sequence identity cut-off, and the resulting 8579 representative sequences were analyzed for chimeras using the UCHIME algorithm in de novo mode. As accurate abundance data was unavailable for many studies in this database, the number of sequences in each CD-HIT-EST cluster was used as a proxy. UCHIME was run using all the default parameters, but

the resulting chimeras were subject to the additional criteria determined empirically to reduce the number of false positives. Putative chimeric sequences were tagged in a field named "PutativeChimera" in the ARB database. Likely chimeras were further defined if the two parent sequences were recovered from the same study, and tagged in the "PutativeChimeraSameStudy" field. After each subsequent ARBitrator run and nifH database update, the most recent sequences are screened using the UCHIME algorithm, and only tagged as a likely chimera if they don't meet the criteria and the parent sequences originate from the same study.

Each updated nifH ARB database version along with supporting documentation is available for download at http://pmc.ucsc.edu/~wwwzehr/research/database/.

Figures

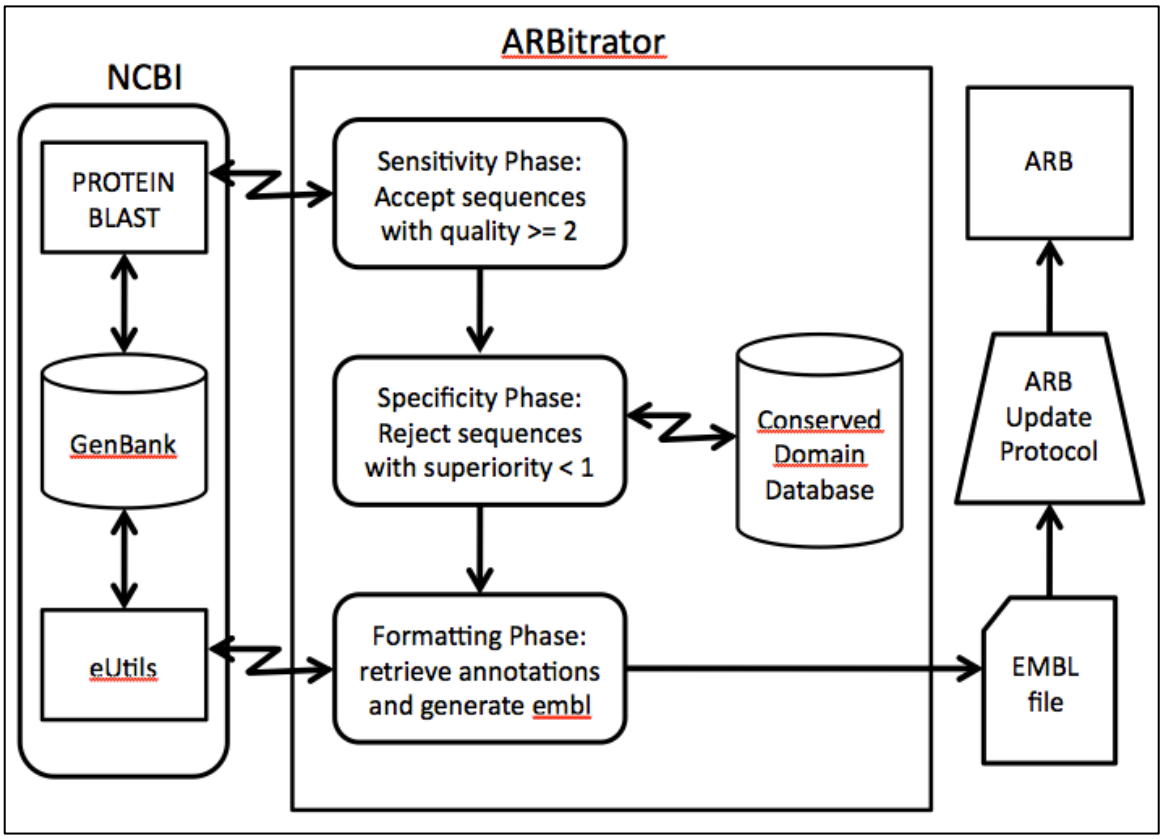


Figure 1. ARBitrator flowchart. In the sensitivity phase, representative nifH sequences are BLASTed against the GenBank database at NCBI. In the specificity phase, candidate sequences retrieved by the sensitivity phase are BLASTed against a database of conserved domains. In the formatting phase, accepted sequence are output as EMBL records. After ARBitrator executes, the protocol described in Supplementary Appendix A is used to add new records to an existing ARB database.

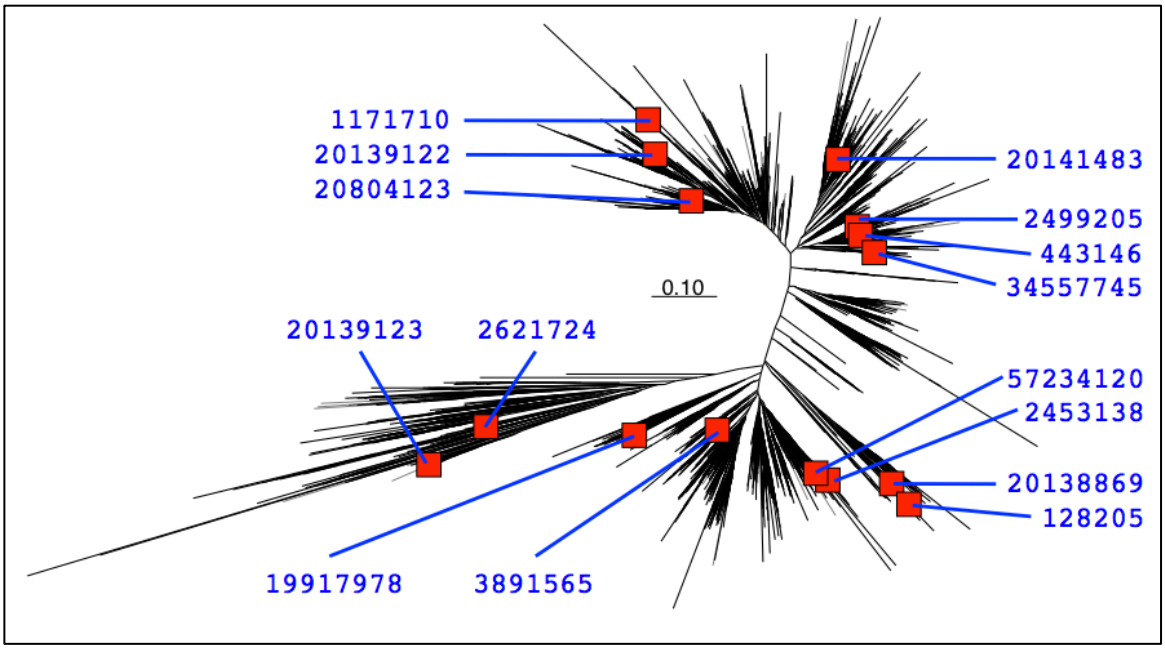


Figure 2. Neighbor-joining tree of partial nifH amino acid sequences constructed using the 15 representative sequences (red boxes) and positive training set sequences. All major clusters are represented by at least one sequence. Labels are GenBank GenInfo Identifiers (GIs).

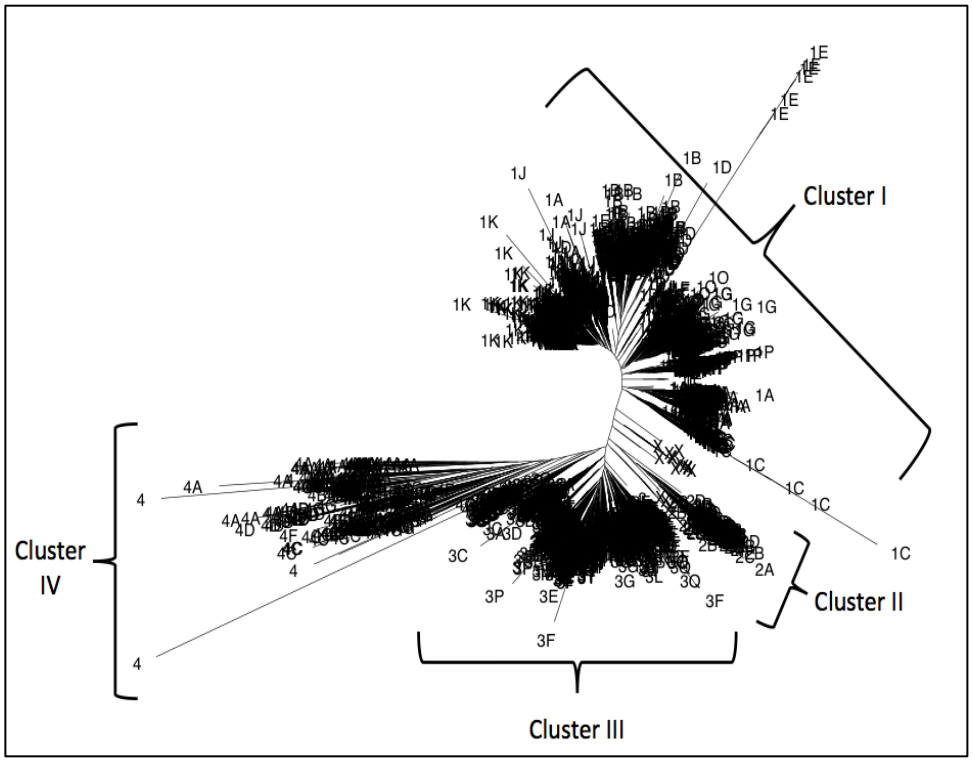


Figure 3. Neighbor-joining tree of partial nifH amino acid sequences representing the 34,420 nifH sequences acquired in the Nov 20, 2012 pipeline run. Sequences were clustered at 98% amino acid identity using CD-HIT (Huang et al., 2010). Clusters and sub-clusters are marked.

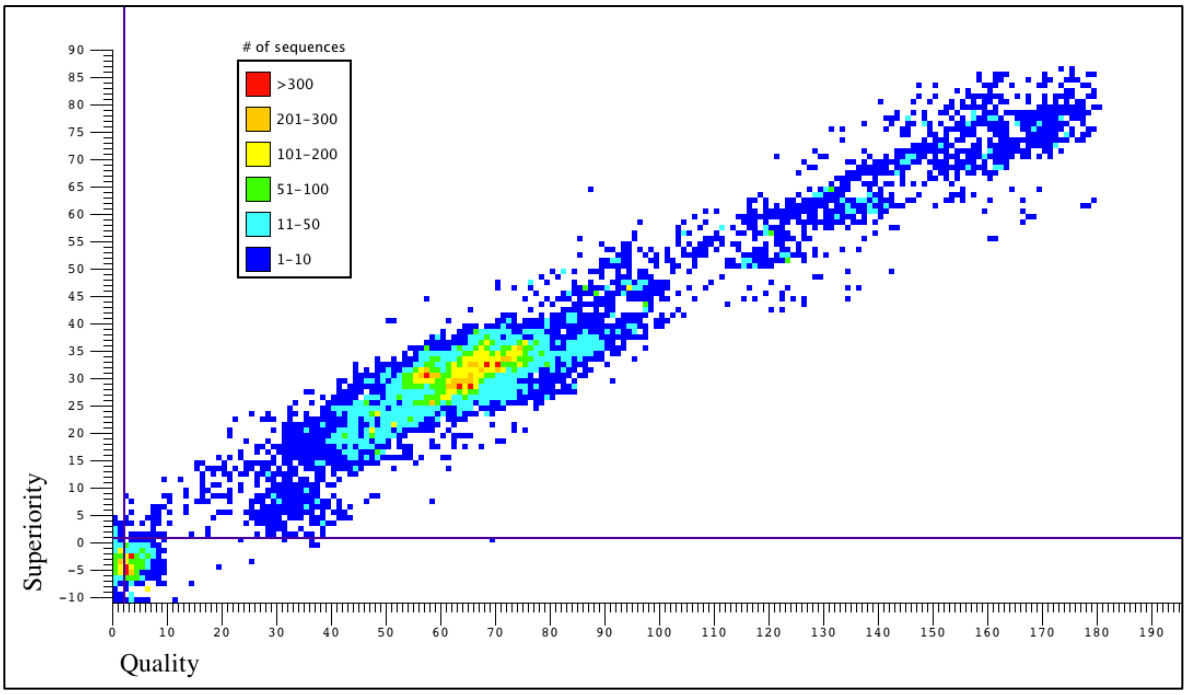


Figure 4. Distribution of sequences by quality and superiority. Thresholds are represented by purple crosshairs; sequences classified as *nifH* by ARBitrator are above and to the right of the crosshairs.

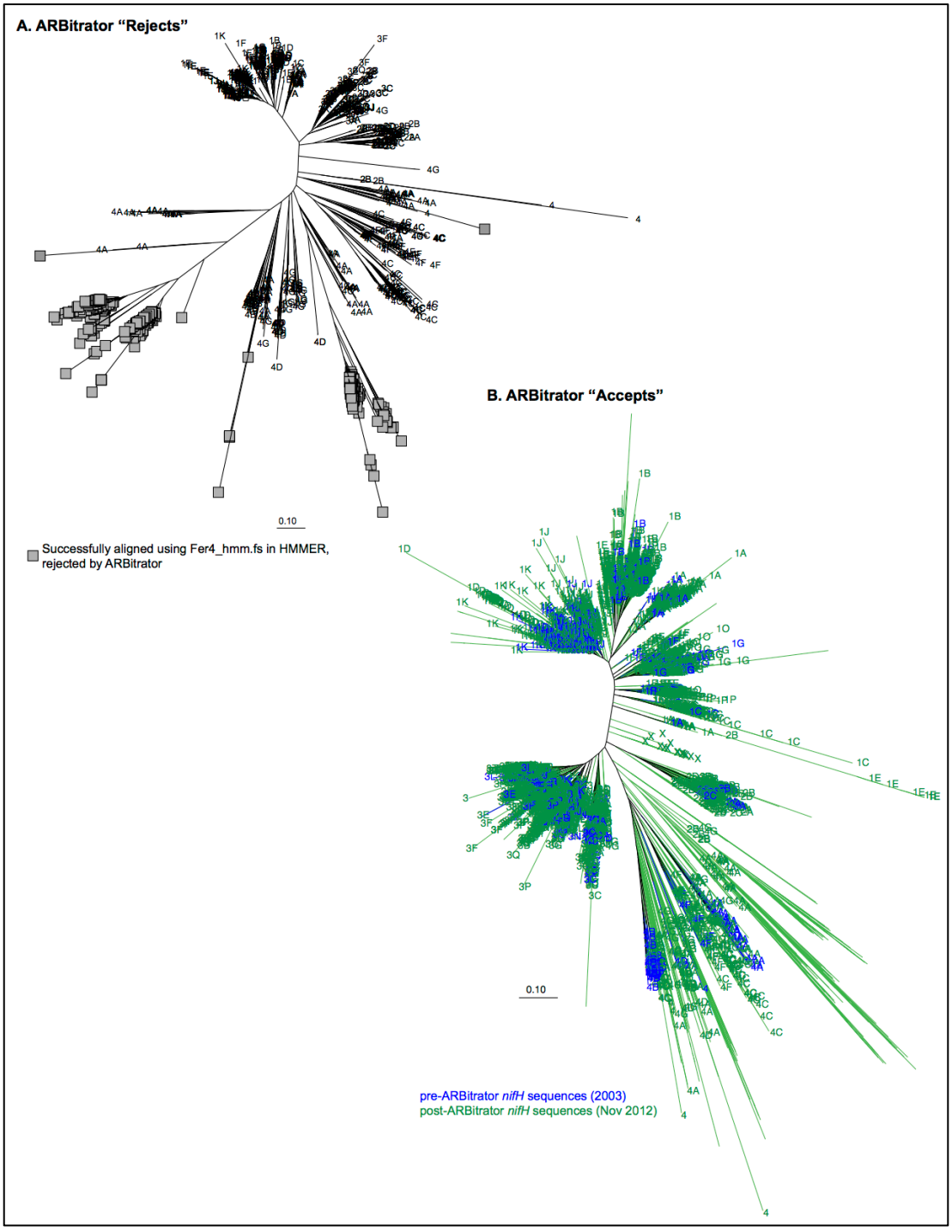


Figure S1. Phylogenetic tree of accepted and rejected sequences.

Chapter 2: Metagenomics of Uncultivated UCYN-A Cyanobacteria[1]

[1]Part of this chapter has been published as: Bombar, D., Heller, P., Sanchez-Baracaldo, P., Carter, B., and Zehr, J.P. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *The ISME Journal*, 1–13. doi:10.1038/ismej.2014.167

Preface

This chapter describes work I performed under the supervision of Dr. Deniz Bombar while he was a postdoctoral fellow working with Dr. J. P. Zehr (Professor, UCSC). The project was the assembly and analysis of the metagenome of an uncultivated marine cyanobacterium, the UCYN-A2 strain of *Candidatus* Atelocyanobacterium thalassa. *Candidatus* Atelocyanobacterium thalassa is a symbiotic nitrogen-fixing microorganism with a wide geographic distribution (Goebel et al., 2010) and possible quantitative biogeochemical significance in the ocean (Agawin et al., 2014). Prior to the work presented here, the genome of only a single strain, UCYN-A1, had been sequenced (Tripp et al., 2010). The present study aimed to compare the genomes of 2 UCYN-A strains, in order to determine how similar they are, if they contained the same unusual deletions, the evolutionary distance or conservation between nucleotide sequences, and to determine if there may be co-evolution between UCYN-A1 and UCYN-A2 and their respective hosts.

My contribution to the project was the assembly of the sequence reads and most of the subsequent bioinformatic analyses. Dr. Bombar performed the comparative analysis between the UCYN-A1 reference strain and the UCYN-A2 strain, and wrote the main body of the article, for which we share “equal

contribution" credit. I describe my contribution in Supplemental Appendix A of the article.

In addition to work presented in the article, I performed several analyses needed to complete the project, including a novel analysis that was necessary but was too detailed to include in the research publication. This part of the project is described here in Chapter 2, immediately following the published work.

Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria

ABSTRACT

Marine planktonic cyanobacteria capable of fixing molecular nitrogen (termed 'diazotrophs') are key in biogeochemical cycling, and the nitrogen fixed is one of the major external sources of nitrogen to the open ocean. *Candidatus* Atelocyanobacterium thalassa (UCYN-A) is a diazotrophic cyanobacterium known for its widespread geographic distribution in tropical and subtropical oligotrophic oceans, unusually reduced genome and symbiosis with a single-celled prymnesiophyte alga. Recently a novel strain of this organism was also detected in coastal waters sampled from the Scripps Institute of Oceanography pier. We analyzed the metagenome of this UCYN-A2 population by concentrating cells by flow cytometry. Phylogenomic analysis provided strong bootstrap support for the monophyly of UCYN-A (here called UCYN-A1) and UCYN-A2 within the marine *Crocosphaera* sp. and *Cyanothece* sp. clade. UCYN-A2 shares 1159 of the 1200 UCYN-A1 protein-coding genes (96.6%) with high synteny, yet the average amino-acid sequence identity between these orthologs is only 86%. UCYN-A2 lacks the same major pathways and proteins that are absent in UCYN-A1, suggesting that both strains can be grouped at the same functional and ecological level. Our results suggest that UCYN-A1 and UCYN-A2 had a common ancestor and diverged after genome reduction. These two variants may reflect adaptation of the host to different niches,

which could be coastal and open ocean habitats.

Introduction

Marine pelagic cyanobacteria play a major role in biogeochemical cycling of carbon and nitrogen in the ocean. *Prochlorococcus* and *Synechococcus* together are the most abundant phototrophic prokaryotes on Earth, and are responsible for a major fraction of oceanic carbon fixation (Partensky et al., 1999; Scanlan and West, 2002; Scanlan, 2003; Johnson et al., 2006). Likewise, cyanobacteria capable of fixing molecular nitrogen ('diazotrophs') dominate global oceanic N_2 fixation although they are typically orders of magnitude less abundant than *Prochlorococcus* or *Synechococcus* (Zehr and Paerl, 2008; Zehr and Kudela, 2011; Voss et al., 2013). Together with upward fluxes of deep-water NO_3^- to the surface ocean, diazotrophs supply the nitrogen requirement of primary productivity and quantitatively balance losses by sinking of organic material, which can sequester CO_2 from the atmosphere to deep waters (Karl et al., 1997; Sohm et al., 2011).

There are several groups of quantitatively significant diazotrophic cyanobacteria in the open ocean, all of which thrive mainly in tropical and subtropical latitudes (Stal, 2009). Traditionally, the filamentous, aggregate-forming cyanobacterium *Trichodesmium* sp. was viewed as the most important oceanic N_2 fixer, based on its wide distribution and direct measurements of its N_2 fixation capacity (Dugdale et al.,

1961; Capone et al., 1997; Bergman et al., 2013). Other diazotrophic cyanobacteria discovered in early microscopic studies are the filamentous heterocyst-forming types of the Richelia and Calothrix lineages, which live in symbioses with several different diatom species (Villareal, 1992; Janson et al., 1999; Foster and Zehr, 2006). More recently, the application of molecular approaches resulted in the discovery of unexpected and unusual cyanobacteria involved in oceanic N_2 fixation (Zehr et al., 1998, 2001). These have usually been grouped as 'unicellular' diazotrophic cyanobacteria, but, among them, different types have very different lifestyles, with *Crocosphaera watsonii* being photo- synthetic and mostly free-living cells (but see Foster et al., 2011), whereas UCYN-A (*Candidatus* Atelocyanobacterium thalassa) is a photoheterotroph that is symbiotic with prymnesiophyte algae (Thompson et al., 2012). While the major biogeochemical role of all diazotrophic cyanobacteria is to provide new nitrogen to the system, their different lifestyles suggest important differences regarding their distribution in the ocean, and the fate of the fixed nitrogen and carbon (Glibert and Bronk, 1994; Scharek et al., 1999; Mulholland, 2007).

As a diazotrophic cyanobacterium, UCYN-A (termed UCYN-A1 from here on) is remarkable in several ways. Although somewhat closely related to *Cyanothece* sp. strain ATCC 51142, the UCYN-A1 genome is only 1.44Mb and lacks many genes including whole metabolic pathways and proteins, such as the oxygen-evolving photosystem II and RuBisCO, that is, features that normally define cyanobacteria (Tripp et al., 2010). The recent identification of a symbiotic eukaryotic prymnesiohyte partner, to which UCYN-A1 provides fixed nitrogen while receiving

carbon in return, is the first known example of a symbiosis between a cyanobacterium and a prymnesiophyte alga (Thompson et al., 2012). Further, UCYN-A1 can be detected in colder and deeper waters compared with other major N_2 fixers like *Trichodesmium* sp. and *C. watsonii* (Needoba et al., 2007; Langlois et al., 2008; Rees et al., 2009; Moisander et al., 2010; Diez et al., 2012), and is also abundant in some coastal waters (Mulholland et al., 2012).

There is now evidence that there are at least three nifH lineages of UCYN-A in the ocean (Thompson et al., 2014). These different clades were previously unrecognized because their nifH amino-acid sequences are nearly identical, with sequence variation primarily only occurring in the third base pair of each codon (Thompson et al., 2014). It is unknown whether these strains are different metabolic variants of UCYN-A, analogous to observations in free-living cyanobacteria like *Prochlorococcus* and *Synechococcus*, which have extensive heterogeneity in their genome contents that enable them to occupy different niches along gradients of nutrients and light (Moore et al., 1998; Ahlgren et al., 2006; Kettler et al., 2007). Phylotype 'UCYN-A2' shares only 95% nifH nucleotide similarity with UCYN-A1, and was discovered to be abundant and actively expressing nifH off the Scripps Institute of Oceanography (SIO) pier. This habitat seems to generally lack UCYN-A1 and has environmental conditions that clearly differ from the tropical/subtropical oligotrophic open ocean during most times of the year (Chavez et al., 2002). UCYN-A2 is associated with a prymnesiophyte host that is closely related to, but not identical to, the UCYN-A1 host (Thompson et al., 2014). Interestingly, the known 18S rRNA gene sequences of the

UCYN-A2 host generally fall into a ‘coastal’ cluster whereas the UCYN-A1 host sequences almost exclusively cluster with sequences recovered from open ocean environ- ments (Thompson et al., 2014). Further, both UCYN-A1 and its host appear to be significantly smaller than UCYN-A2 and its host (Thompson et al., 2014). On the basis of these findings, Thompson et al. (2014) suggested that UCYN-A1 could be an oligotrophic open ocean ecotype, whereas UCYN-A2 could possibly be more adapted to coastal waters.

The present study is the first opportunity to characterize the metabolic potential of a new clade of UCYN-A, by analyzing the metagenome of a UCYN-A2 population sampled from waters off the SIO pier. This enabled us to test whether habitat differences, or a distinct symbiont–host relation- ship, are reflected in genome features that distin- guish UCYN-A2 from UCYN-A1, and whether UCYN-A2 has the same lack of genes as UCYN-A1. With the availability of the new UCYN-A2 metagenome, it was also possible to perform phylogenomic analyses (including 135 proteins), to determine whether UCYN-A2 and UCYN-A1 form a monophyletic group, and to establish how these two organisms are related to other cyanobacteria.

Materials and Methods

Sampling

After the initial detection of a new *nifH* phylotype similar to UCYN-A1 in coastal

waters off Scripps Pier and its classification as a new strain (UCYN-A2, Thompson et al., 2014), we used the previously described cell-sorting approach (Zehr et al., 2008; Thompson et al., 2012) to obtain cell sorts enriched in UCYN-A2 for genome sequencing. Surface water samples (10l) were taken at Scripps Pier with a bucket, gently poured into a polypropylene bottle and immediately transferred to the laboratory at Scripps. The sample was then concentrated by gentle vacuum filtration through a 0.22-micron-pore-size polycarbonate filter and cells resuspended by vortexing the filter in 50ml of sterile-filtered seawater. The concentrate was flash-frozen in liquid nitrogen and shipped to the University of California, Santa Cruz, USA.

Fluorescence-activated cell sorting and nifH quantitative PCR (qPCR) and genome amplification

The concentrated seawater samples were thawed at room temperature and briefly vortexed again immediately prior to cell sorting. Seawater samples were pre-filtered using 50-micron-mesh-size CellTrics filters (Partec, Swedesboro, NJ, USA) to prevent clogging of the nozzle (70 micron diameter) with large particles. Samples were analyzed in logarithmic mode with an Influx Cell Sorter (BD Biosciences, San Jose, CA, USA). Flow cytometry sorting gates were defined using forward scatter (a proxy for cell size) and chlorophyll fluorescence at 692 nm (Figure 1). Chlorophyll autofluorescence was excited using a 200-mW, 488-nm sapphire laser (Coherent, Santa Clara, CA, USA).

A UCYN-A2-specific qPCR assay (Thompson et al., 2014) was used to screen sorted events within each gate (between 100 and 200 events). Cells were sorted directly into aliquots of 10-ml 5-kDa filtered nuclease-free water, and then amended with qPCR 1 x Universal PCR master mix (Applied Biosystems, Foster City, CA, USA) to a total reaction volume of 25 ml, including UCYN-A2-specific forward and reverse primers (0.4 micromolar final concentration), as well as TaqMan (Life Technologies Corp., Grand Island, NY, USA) probes (0.2 micromolar final concentration). qPCR reactions were conducted in a 7500 real-time PCR instrument (Applied Biosystems). Reaction and thermal-cycling conditions were as described previously (Moisander et al., 2010; Thompson et al., 2014). Abundances of *nifH* gene copies were quantified relative to standard curves comprising amplification of linearized plasmids containing inserts of the target *nifH* gene, and abundances of gene copies per sample calculated as described by Short and Zehr (2005). Standards were made from serial dilutions of plasmids in nuclease-free water (range: 1–10^3 *nifH* gene copies per reaction), with 2 ml of each dilution added up to 25 ml of qPCR (total volume) mixtures. Duplicates of each standard were included with each set of samples run on the qPCR instrument, as well as at least two no-template controls.

Using this approach, we detected a sort region relatively enriched in UCYN-A2 but still containing other organisms besides the target (Figure 1). This region appears to include single UCYN-A2 cells rather than populations in the picoeukaryote-size fraction as described in Thompson et al. (2014) (Figure 1). The disruption of the

UCYN-A symbiotic association appears to be a typical result of the concentration and freezing protocol (Thompson et al., 2012), and proved advantageous for our genome amplification and assembly. A sample taken on 31 May 2011 was used to obtain a cell sort enriched in UCYN-A2 for genome amplification (Figure 1). Approximately 3.5×10^4 events were sorted into a 1.5-ml microcentrifuge tube containing 90 ml of TE buffer. Cells were pelleted at 14 000 r.p.m. (21 000 x g) for B45 min and the supernatant was discarded. We used a Qiagen REPLI-g Midi kit (Valencia, CA, USA) for cell lysis and amplification of genomic DNA, following the manufacturer's recommendations with few modifications. Briefly, the pelleted cells were resuspended in 3.5ml phosphate-buffered saline buffer and 3.5 ml buffer D2 (0.09 M dithiothreitol), incubated at 65 1C for 5min, and immediately stored on ice after adding the kit-provided 'stop buffer'. The amplification reaction was carried out in a thermal cycler at 30 1C for 6 h after addition of 40 ml Repli G mastermix to the tube. The quality, size and quantity of the amplified DNA were checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and again quantified using Pico Green (Invitrogen Corp., Carlsbad, CA, USA). The suitability of this sample for a genome-sequencing run was indicated by the presence of 10^6 *nifH* gene copies of UCYN-A2 per ml, measured by qPCR.

Illumina sequencing

Library preparation and paired-end sequencing were performed at the BioMicro Center of the Massachusetts Institute of Technology (MIT, http://

openwetware.org/wiki/BioMicroCenter:Sequencing). The DNA sample was split into two equal aliquots and prepared for sequencing using the SPRI works system (Beckman Coulter Genomics, Danvers, MA, USA) with 150–350 and 250–550 bp inserts. Ligated libraries were amplified and molecular barcodes added. Samples were pooled and sequenced on an Illumina (San Diego, CA, USA) MiSeq v1 flowcell with 151bp of sequence read in each direction. Fastq files (Illumina v1.5) were prepared and separated into the individual libraries, allowing one mismatch with the barcode sequences. Post-run quality control includes confirmation of low sequencing error rates by analyzing phiX spike sequences, checking for significant contamination from human, mouse, yeast and Escherichia coli, and confirming the presence of only the expected barcodes.

Please see the Supplementary Material section for a detailed description of sequence assembly, annotation and phylogenomic analyses. This sequencing project has been deposited at DDBJ/EMBL/GenBank under the organism name '*Candidatus* Atelocyanobacterium thalassa isolate SIO64986', accession number JPSP01000000.

Results

The aligned UCYN-A2 scaffolds to the UCYN-A1 reference chromosome covered nearly the entire UCYN-A1 sequence (Figure 2). For the majority of the adjacent pairs of scaffolds, the last gene of the upstream scaffold and the first gene of the downstream scaffold matched consecutive genes in the gene order of UCYN-A1 (30 cases), thereby conserving and extending the high synteny seen across the alignments.

In the remaining cases, adjacent scaffold ends carried partial genes that matched different parts of the same gene in UCYN-A1 (43 partial genes in UCYN-A2 matching 21 genes in UCYN-A1).

Overall, the UCYN-A2 draft genome is highly similar to UCYN-A1 in gene content, synteny and basic genome features, including GC content (31%), percent of coding DNA (79.3%), codon usage (Supplementary Figure S4) and overall gene count, including two rRNA operons (Figure 2 and Table 1). There is 99% 16S rRNA gene sequence identity between both genomes. Seven RNA genes in UCYN-A2 had very similar but unannotated sequences in UCYN-A1 (91–100% nucleotide identity over 97– 100% of the query sequence), and some annotated matching sequences exist in other cyanobacteria such as *Calothrix* sp. PCC7507 and *Cyanothece* sp. 8801, 8802 and 51142. These consist of one additional tRNA gene for methionine and six RNA genes annotated as noncoding RNA with unknown functions ('other RNA genes' in Table 1).

A total of 1159 of the 1200 UCYN-A1 proteins (Tripp et al., 2010) have closely matching sequences in UCYN-A2, that is, 96.6% of UCYN-A1's genes are shared with UCYN-A2. For these 1159 genes, the average amino-acid sequence identity is 86.3% (range 51–100%, Figure 2). The most conserved genes (X95% identity) include housekeeping genes (ribosomal proteins, NADH dehydrogenase, ATP synthase), Photosystem I subunits and proteins involved in N_2 fixation (nif cluster).

The previously described UCYN-A1 genome was unusual and had extensive genome

reduction, lacking the genes encoding Photosystem II, RuBisCO, biosynthesis pathways for several amino acids and purines, as well as the TCA cycle and other key metabolic pathways (Zehr et al., 2008; Tripp et al., 2010). The genes missing in the UCYN-A1 genome were also absent in the UCYN-A2 draft genome. In addition to the analysis of all rejected contigs, we used TBLASTN to search the full set of unassembled sequencing reads for all 114 *Cyanothece* sp. 51142 genes reported missing in UCYN-A1 (Tripp et al., 2010), to test whether some of these genes might have escaped assembly. Subject reads were compared with GenBank using BLASTN against the nt database, and taxonomy was retrieved for the top 20 hits for each read. Matching reads were found for only 13 different genes out of these 114 query genes (18 total hits, incl. 5 PSII genes). Seven hits had 98–100% identity to known organisms (*Synechococcus*, *Pelagomonas*, *Thalassiosira pseudonana*), and four hits to an uncultured marine prokaryote. The remaining seven hits had maximal identity ranging between 79% and 89% to sequences from other organisms (*Galdieria*, *Aureococcus*, *Acaryochloris*, *Flavobacterium*, *Nitrosomonas* and *Monosiga*).

Apart from the 1159 genes shared by UCYN-A2, there are 41 UCYN-A1 genes (including 25 hypothetical proteins) that appear to be pseudogenes in UCYN-A2. These pseudogenes were either neighboring partial genes that aligned consecutively to a full open reading frame of a UCYN-A1 gene, with interrupting stop codons and/or insertions between them (a total of 21 partial genes in UCYN-A2 matching eight genes in UCYN-A1, not counting genes at scaffold ends; Table 2), or short, unannotated sequences that match only parts of UCYN-A1 genes (the remaining 33

UCYN-A1 genes). Although the evidence for pseudogenes was strong, as the UCYN-A2 sequences were from good assemblies that yielded high-coverage scaffolds, we additionally used PCR to amplify across nine random examples of these pseudogenes, confirming that the interrupting stop codons were present and were not artifacts of assembly (see Supplementary Material for details). The genome comparison revealed that such pseudogenes also exist in UCYN-A1 (Table 2).

An interesting difference between both genomes is that for all UCYN-A1 genes at least short, unannotated remnants or pseudogenes can be found in UCYN-A2, while in turn UCYN-A2 possesses 31 genes, of which 15 are hypothetical proteins, for which no traces (pseudogenes or gene remnants) were found in UCYN-A1, indicating that they have been completely lost from the genome (Table 2). The loss of these genes has in most cases resulted in further genome compaction in UCYN-A1, that is, they appear fully excised instead of being replaced by noncoding DNA (examples shown in Figure 3). The majority of these unique UCYN-A2 genes had top BLASTP similarity to genes in different *Cyanothece* sp. (16 genes) or in other Cyanobacteria (five genes), whereas 10 short hypothetical proteins (27–63 amino acids) had no clear phylogenetic affiliation.

In addition to interrupted genes, we note 132 genes that show differences in amino-acid length compared with orthologs in the other genome, that is, they appear truncated at either the C- or N-terminal end of the protein. For UCYN-A2, this was also confirmed for a few examples by PCR amplification (Supplementary Material).

Some of these truncated genes might be pseudogenes as well. Thirteen genes in UCYN-A1 and 14 genes in UCYN- A2 had o75% of the amino acids in the comparable protein sequence in the other strain. A comparison of the ortholog pairs of UCYN-A1 and UCYN-A2 with orthologs in *Cyanothece* sp. 51142 showed that the truncated versions of the genes almost exclusively occur in one of the UCYN-A strains, but not in *Cyanothece* sp. 51142, whereas the gene length of the longer ortholog in UCYN-A1/A2 correlated well with the gene length in *Cyanothece* sp. 51142 (Figure 4a). Interestingly, UCYN-A1 generally possessed the shortest versions of the gene among these three genomes (Figure 4b).

Overall, both genomes show extremely similar genome reduction, but there are some differences regarding which genes have become pseudogenes, and UCYN-A1 appears to have a higher level of reduction, with fully excised genes at several loci and overall greater truncation of genes. Functions affected by gene deletions or pseudogenization differ for UCYN-A1 and UCYN-A2 (Table 2), with the latter genome, for example, retaining genes involved in cell wall synthesis, vitamin import and detoxification of active oxygen species such as H_2O_2.

Maximum likelihood analyses confirmed that both UCYN-A strains belong to a well-supported monophyletic group of marine planktonic cyanobacteria containing *Crocosphaera* sp., *Cyanothece* sp. and other unicellular N_2 fixing cyanobacteria (Sanchez- Baracaldo et al., 2014). The results of the analyses strongly support that UCYN-A2 and UCYN-A1 form a monophyletic group, that is, a sister group to

Crocosphaera sp. and *Cyanothece* sp. (Bootstrap support 100; Figure 5). This clade of marine unicellular N_2 fixers belongs to the previously described SPM group (Sanchez-Baracaldo et al., 2005) containing *Synechocystis*, *Pleurocapsas* and *Microcystis* (Figure 5).

Discussion

UCYN-A is likely one of the major oceanic N_2 fixers given that it has a wider geographic distribution than *Trichodesmium* sp., diatom symbionts or *Crocosphaera* sp., and can be highly abundant at certain times and places (Church et al., 2009; Moisander et al., 2010). The symbiotic relationship of UCYN-A with a eukaryotic, possibly calcifying, prymnesiophyte raises many important questions about the variability and regulation of N_2 fixation in UCYN-A, the fate of the fixed nitrogen (and carbon) in the planktonic food web, the role of UCYN-A in element export to the deep ocean, and its susceptibility to ocean acidification (Thompson et al., 2012). Further, the recently recognized *nifH* sequence diversity in the UCYN-A clade suggests that there could be different ecotypes of UCYN-A in the ocean, which could potentially be very different in terms of genome composition and physiology (Thompson et al., 2014). The genome comparison in this study addresses this question, with the surprising discovery that both types have very similar gene content, genome reduction, but also substantially divergent DNA sequences.

UCYN-A2 has very similar gene content to UCYN-A1 and also lacks photosystem II genes, RuBisCO, TCA cycle components and other pathways. It therefore is a second,

independently verified example of this kind of genome reduction in UCYN-A symbionts. Together with the highly conserved gene order, which implies gene function conservation, this suggests that UCYN-A1 and UCYN-A2 have similar functions and metabolic interactions in the symbiosis with their haptophyte hosts.

Although it can be difficult to confirm that genes are missing in unclosed genomes, we base the claim on several independent lines of evidence. (1) Many scaffolds ended with partial genes that mapped to a single UCYN-A1 gene, or ended with full genes that matched and preserved the gene order in UCYN-A1, suggesting that breaks between scaffolds were not due to missing sequence. (2) Even though there is variability in genome sequence coverage (26.7 on average, Supplementary Figure S2), it is highly unlikely that there would be no coverage at all for the long stretches of target genome needed to contain the many missing genes in UCYN-A1. (3) The rejected contigs had a GC content of 44.7% (very different from the 31% found in UCYN-A1 and UCYN-A2), sparse BLAST hits to UCYN-A1 or *Cyanothece* sp. (even at a very relaxed e-value threshold), and any detected hits to UCYN-A1- or *Cyanothece* sp.-like sequences were redundant, with genes already present in the UCYN-A2 draft genome; this ascertains that no UCYN-A2 genes were missed. (4) Searching the sequence reads by TBLASTN for all 114 *Cyanothece* sp. 51142 genes that appeared to be missing in UCYN-A1 returned only 13 of the query genes, of which most had highest similarity values to different organisms. (5) Recently obtained field data show peaks in *nifH* expression of UCYN-A2 during daytime, closely matching the temporal patterns of *nifH* expression determined for UCYN-A1

in the open-oligotrophic ocean around Hawaii (Church et al., 2005; Thompson et al., 2014). This may be viewed as further confirmation for the absence of oxygen-evolving PSII in UCYN-A2, given the oxygen sensitivity of the nitrogenase enzyme.

Each UCYN-A strain has only a handful of genes that are either absent or disrupted in the other genome (Table 2). The loss of genes in symbiont genomes is a gradual process, and highly reduced genomes characteristically exhibit slow gene loss in the form of erosion of individual genes or operons, rather than larger deletions via chromosomal rearrangements (Moran and Mira, 2001; Wernegreen et al., 2002; Moran, 2003). The pattern of lost, disrupted or truncated genes seen in the UCYN-A strains examined here appears consistent with such slow gene decay.

Gene inactivation and loss in symbionts mainly occurs because genes become functionally redundant and therefore non-essential, for example, due to metabolite exchange with the host. Many of the functions encoded by pseudogenes in UCYN-A1/A2 indeed appear dispensable when considered in the context of the symbiont–host relationship, such as restriction endonucleases, pyrimidine synthesis or cell motility (Table 2). However, the intact versions of those genes in the other genome, and the unique genes in UCYN-A2, raise the question whether they have been retained because their function is still important, or whether they are also non-essential/ redundant but have so far escaped inactivation and elimination. Noteworthy examples are the genes involved in cell wall biogenesis and cell shape determination in UCYN-A2. The latter genes occur in rod-shaped cells and also in *Cyanothece* sp.

51142. These genes could indicate that UCYN-A2 has a different morphology than UCYN-A1, and could point to differences in how it is structurally associated with its host, which might also influence the fragility of the association. Interestingly, genes involved in cell wall biogenesis, which have become pseudogenes in UCYN-A1, are also among disrupted genes in the obligate cyanobacterial endosymbiont of the diatom *Rhopalodia gibba* (Kneip et al., 2008). Another interesting case is the UCYN-A2 peroxidase gene 2528848519. Peroxidases act in detoxifying active oxygen species such as H_2O_2, for example, the thioredoxin peroxidase in *Synechocystis* PCC6803 (68% nucleotide identity to UCYN-A2 gene) (Yamamoto et al., 1999). Active oxygen species are formed during respiration and photosynthesis, but also during many other processes (Miyake and Yokota, 2000). The presence of a peroxidase could indicate that UCYN-A2 experiences higher intracellular oxygen concentrations than UCYN-A1. UCYN-A2 would then have to respire more oxygen in order to fix N_2, and in the process would generate more reactive oxygen species, thus potentially relying on this peroxidase gene.

On the basis of searches in metagenomic and metatranscriptomic data sets, the UCYN-A1 genome was initially assumed to be a global population with very similar genome sequences (X97% nucleotide-sequence identity, Tripp et al., 2010), analogous to the low sequence diversity seen in *C. watsonii* (Zehr et al., 2007; Bench et al., 2011). While the phylogenomic analysis strongly supports the two UCYN-A strains to be sister species (Figure 5), one of the striking results from our genome comparison is the relatively large range of sequence similarity seen among shared

genes in UCYN-A1 and UCYN-A2 (Figure 2). The combination of this sequence divergence with the extremely high similarity in basic genome features, gene content and synteny suggests that the genome reduction occurred prior to the speciation event and genetic divergence. It is therefore likely that the common ancestor of UCYN-A1 and UCYN-A2 was already a symbiont. Vicariance might have triggered the genetic divergence in the course of speciation of the prymnesiophyte host into strains that possibly are slightly better adapted to different oceanic realms. This would have allowed the cyanobacterial genomes to accumulate gene sequence mutations after driving forces causing large genome rearrangements were no longer significant, which appears typical for symbiont genomes that have already been highly reduced (Tamas et al., 2002; Moran, 2003; Silva et al., 2003). Interestingly, genes involved in N_2 fixation were among the most conserved orthologs, likely reflecting the importance of this process in maintaining the symbiosis, as it arguably represents the function most beneficial to the host and which must have been vital in the initial formation of the symbiotic relationship.

Small, conserved and highly syntenic genomes exhibiting high amino-acid divergence can also be found in the free-living heterotrophic SAR11 clade (Wilhelm et al., 2007; Grote et al., 2012). SAR 11 is an example for genome reduction due to 'streamlining', whereas the genome reduction seen in UCYN-A appears typical for symbiont genomes (Giovannoni et al., 2014). The amino-acid divergence between the UCYN-A strains lies within the range seen in the SAR11 Ia cluster (which have 2% 16S rRNA divergence, Grote et al., 2012). However, UCYN-A1 and UCYN-A2 have

even more conserved genome content than SAR11 Ia and are considerably more conserved than members of the cyanobacterial *Prochlorococcus* group (Kettler et al., 2007), which appears typical for obligate intracellular organisms (Grote et al., 2012). This evolutionary pattern is unusual and suggests that the genomes of these UCYN-A strains are under strong selection, as they are highly specialized symbionts of eukaryote algae.

Although *nifH* sequences of UCYN-A1 and UCYN-A2 can co-occur in some samples from around the world, the question has been raised whether these two different strains could be adapted to different nutrient regimes, and could therefore have overlapping, but different distributions in the ocean (Thompson et al., 2014). However, we find no evidence in the genomes of UCYN-A1 and UCYN- A2 that would resemble genetic differentiation analogous to that in, for example, the high-light or low-light ecotypes of *Prochlorococcus* sp. (Moore et al., 1998; Kettler et al., 2007), or the 'coastal' ecotypes of *Synechococcus* sp. (Ahlgren and Rocap, 2006; Palenik et al., 2006). This lack of genetic differentiation, and the overall level of genome reduction, is characteristic for genomes of obligate symbionts with high dependency on their host (Moran, 2003; Hilton et al., 2013), and suggests that UCYN-A may not be directly exposed to, or affected by the external environment. Analyzing the genomes of the host algae and other UCYN-A strains will be necessary to identify genes that might represent adaptation to different environmental conditions.

While the two strains show no immediately apparent gene adaptations to cope with horizontal nutrient gradients or light quality, it is interesting that UCYN-A1 appears to be smaller than UCYN-A2 (Thompson et al., 2014), has fully excised genes compared with UCYN-A2 (Figure 3) and greater truncation of genes (Figure 4). The genomic signatures in UCYN-A point to typical genome reduction in a symbiont via genetic drift, a mechanism that is particularly enhanced under small effective population sizes (van Ham et al., 2003; Giovannoni et al., 2014). However, the further reduced genome of UCYN-A1 could also reflect an adaptation to the open ocean environment with very low levels of nutrients. Comparative genomics and ecological studies (Scanlan et al., 2009), as well as trait evolution analyses (Larsson et al., 2011), have shown a trend in genome reduction among cyanobacteria adapted to oligotrophic environments. For the host of UCYN-A, the ecological advantage of hosting a 'diazoplast' would come at the cost of having to sustain it with carbon energy, nutrients and a range of metabolites. Thus, it appears possible that more severe nutrient deprivation (especially for phosphorus, Scanlan et al., 2009) experienced by an open ocean ecotype of the host would also induce more extensive genome compaction (that is, stream- lining) in the symbiont. Further studies are necessary to fully understand these observations.

Conclusions

The genomes of the two UCYN-A strains show considerable divergence at the amino-acid and nucleotide levels along with high conservation of genome structure, gene

content and basic genome features, suggesting that they had a common symbiotic ancestor and then were separated spatially in the course of speciation. While there is some evidence for unequal distribution and possibly habitat-specific genomic streamlining in these two strains, it remains unclear whether they occupy different or overlapping niches. The genome size and the number of pseudogenes not yet fully excised from the genome of both strains might suggest that UCYN-A is still in a relatively early stage of symbiotic association with the eukaryotic host, analogous to, for example, the diazotrophic spheroid bodies found in rhopalodiacean diatoms (Kneip et al., 2008; Nakayama et al., 2014). Genome sequencing of additional UCYN-A strains and of host genomes will show whether the small differences in genetic potential reflect environmental adaptation in these organisms, and whether genetic material from UCYN-A has migrated into the host genome, as found in organelle-like stages of symbiosis (Nakayama and Ishida, 2009). The existence of different UCYN-A strains associated with different prymnesiophytes has implications for the trophic transfer and vertical export of nitrogen and carbon, and for the distribution and regulation of N_2 fixation in the ocean. Further studies are needed for a better understanding of symbiotic N_2 fixation and the genomic basis for UCYN-A's role as a globally important N_2 fixer.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank F Malfatti and F Azam at SIO for assistance with sampling and for providing lab facilities, Kendra Turk-Kubo for carrying out PCR reactions to confirm pseudo-gene sequences, S Biller at MIT for assistance with sample handling, S Bench for bioinformatics assistance, and A Thompson, J Tripp and J Hilton for valuable discus- sions. Comments from three anonymous reviewers greatly improved the paper. This work was supported by a Gordon and Betty Moore Foundation Marine Investigator Award (JZ), the MEGAMER facility (supported by GBMF) and the Center for Microbial Oceanography: Research and Education (NSF grant 0424599).

References

Ahlgren N, Rocap G. (2006). Culture isolation and culture-independent clone libraries reveal new marine Synechococcus ecotypes with distinctive light and N physiologies. *Appl Environ Microbiol*. **72**: 7193–7204.

Ahlgren NA, Rocap G, Chisholm SW. (2006). Measurement of prochlorococcus ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environ Microbiol* **8**: 441–454.

Bench SR, Ilikchyan IN, Tripp HJ, Zehr JP. (2011). Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features. *Front Microbiol* **2**: 261.

Bergman B, Sandh G, Lin S, Larsson J, Carpenter EJ. (2013). Trichodesmium - a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **37**: 286–302.

Blank CE, Sanchez-Baracaldo P. (2010). Timing of morphological and ecological innovations in the cyanobacteria - a key to understanding the rise in atmospheric oxygen. *Geobiology* **8**: 1–23.

Capone DG, Zehr JP, Paerl HW, Bergman B, Carpenter EJ. (1997). *Trichodesmium*, a globally significant marine cyanobacterium. *Science* **276**: 1221–1229.

Chavez FP, Pennington JT, Castro CG, Ryan JP, Michisaki RM, Schlining B et al.

(2002). Biological and chemical consequences of the 1997-98 El Nino in central California waters. *Progr Oceanogr* **54**: 205–232.

Church MJ, Mahaffey C, Letelier RM, Lukas R, Zehr JP, Karl DM. (2009). Physical forcing of nitrogen fixation and diazotroph community structure in the North Pacific subtropical gyre. *Global Biogeochem Cycles* **23**: GB2020.

Church MJ, Short CM, Jenkins BD, Karl DM, Zehr JP. (2005). Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. *Appl Environ Microbiol* **71**: 5362–5370.

Diez B, Bergman B, Pedros-Alio C, Anto M, Snoeijs P. (2012). High cyanobacterial *nifH* gene diversity in Arctic seawater and sea ice brine. *Environ Microbiol Rep* **4**: 360–366.

Dugdale RC, Menzel DW, Ryther JH. (1961). Nitrogen fixation in the Sargasso Sea. *Deep-Sea Research* **7**: 297–300.

Foster RA, Kuypers MMM, Vagner T, Paerl RW, Musat N, Zehr JP. (2011). Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *ISME J* **5**: 1484–1493.

Foster RA, Zehr JP. (2006). Characterization of diatom-cyanobacteria symbioses on the basis of *nifH*, *hetR* and 16S rRNA sequences. *Environ Microbiol* **8**: 1913–1925.

Giovannoni SJ, Thrash JC, Temperton B. (2014). Implications of streamlining theory for microbial ecology. *ISME J* **8**: 553–565. doi:10.1038/ismej.2014.60.

Glibert PM, Bronk DA. (1994). Release of dissolved organic nitrogen by marine diazotrophic cyanobacteria Trichodesmium spp. *Appl Environ Microbiol* **11**: 3996–4000.

Grote J, Thrash C, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3**: e00252–12.

Hilton JA, Foster RA, Tripp JH, Carter BJ, Zehr JP, Villareal TA. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat Commun* **4**: 1767.

Janson S, Wouters J, Bergman B, Carpenter EJ. (1999). Host specificity in the *Richelia*-diatom symbiosis revealed by *hetR* gene sequence analysis. *Environ Microbiol* **1**: 431–438.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW.

(2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.

Karl D, Letelier R, Tupas L, Dore J, Christian J, Hebel D. (1997). The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **388**: 533–538.

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

Kneip C, Vobeta C, Lockhart P, Maier U. (2008). The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC. Evol Biol* **8**: 30.

Langlois RJ, Hummer D, LaRoche J. (2008). Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. *Appl Environ Microbiol* **74**: 1922–1931.

Larsson J, Nylander J, Bergman B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC. Evol Biol* **11**: 187.

Miyake C, Yokota A. (2000). Determination of the rate of photoreduction of O2 in the water-water cycle in watermelon leaves and enhancement of the rate by limitation of photosynthesis. *Plant Cell Physiol* **41**: 335–343.

Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS, Carlson CA et al. (2010). Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327**: 1512–1514.

Moore LR, Rocap G, Chisholm SW. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**: 464–467.

Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6**: 512–518.

Moran NA, Mira A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol* **2**: RESEARCH0054.

Mulholland MR. (2007). The fate of nitrogen fixed by diazotrophs in the ocean. *Biogeosciences* **4**: 37–51.

Mulholland MR, Bernhardt PW, Blanco-Garcia JL, Mannino A, Hyde K, Mondragon E et al. (2012). Rates of dinitrogen fixation and the abundance of diazotrophs in North American coastal waters between Cape Hatteras and Georges Bank. *Limnol*

Oceanogr **57**: 1067–1083.

Nakayama T, Ishida K-i. (2009). Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. *Curr Biol* **19**: R284–R285.

Nakayama T, Kamikawa R, Tanifuji G, Kashiyama Y, Ohkouchi N, Archibald JM et al. (2014). Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proc Natl Acad Sci USA* **111**: 11407–11412.

Needoba JA, Foster RA, Sakamoto C, Zehr JP, Johnson KS. (2007). Nitrogen fixation by unicellular diazotrophic cyanobacteria in the temperate oligotrophic North Pacific Ocean. *Limnol Oceanogr* **52**: 1317–1327.

Palenik B, Ren QH, Dupont CL, Myers GS, Heidelberg JF, Badger JH et al. (2006). Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *ProcNatlAcad Sci USA* **103**: 13555–13559.

Partensky F, Blanchot J, Vaulot D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic water: a review. *Bull Inst Oceanogr Special* **19**: 457–475.

Rees AP, Gilbert JA, Kelly-Gerreyn BA. (2009). Nitrogen fixation in the western English Channel (NE Atlantic Ocean). *Mar Ecol Prog Ser* **374**: 7–12.

Ruby JG, Bellare P, DeRisi JL. (2013). PRICE: software for the targeted assembly of components of (meta) genomic sequence data. *G3: (Bethesda)* **3**: 865–880.

Sanchez-Baracaldo P, Ridgwell A, Raven JA. (2014). A neoproterozoic transition in the marine nitrogen cycle. *Curr Biol* **24**: 652–657.

Sanchez-Baracaldo P, Hayes PK, Blank CE. (2005). Morphological and habitat evolution in the cyanobacteria using a compartmentalization approach. *Geobiology* **3**: 145–165.

Scanlan DJ. (2003). Physiological diversity and niche adaptation in marine *Synechococcus*. In Advances in Microbial PhysiologyVol. 47. Academic Press Ltd: London, pp 1–64.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR et al. (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev*. **73**: 249–299.

Scanlan DJ, West NJ. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol Ecol* **40**: 1–12.

Scharek R, Tupas L, Karl DM. (1999). Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station ALOHA. *Mar Ecol Prog Ser* **182**: 55–67.

Short SM, Zehr JP. (2005). Quantitative analysis of *nifH* genes and transcripts from aquatic environments In: Jared RL (eds) *Methods Enzymology* Vol. 397. Academic Press, pp 380–394.

Silva FJ, Latorre A, Moya A. (2003). Why are the genomes of endosymbiotic bacteria so stable? *Trends genet* **19**: 176–180.

Sohm JA, Webb EA, Capone DG. (2011). Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol* **9**: 499–508.

Stal LJ. (2009). Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? *Environ Microbiol* **11**: 1632–1645.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ et al. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**: 2376–2379.

Thompson AW, Carter BJ, Turk-Kubo KA, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol*; doi:10.1111/ 1462-2920.12490.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**: 1546–1550.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F et al. (2010). Metabolic streamlining in an open ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90–94.

van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U et al. (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* **100**: 581–586.

Villareal TA. (1992). Marine nitrogen-fixing diatom - cyanobacteria symbioses. In: Carpenter EJ, Capone DG, Rueter JG (ed) *Marine Pelagic Cyanobacteria: Trichodesmium and Other Diazotrophs*. Kluwer Academic Publishers: The Netherlands, pp 163–175.

Voss M, Bange HW, Dippner JW, Middelburg JJ, Montoya JP, Ward B. (2013). The

marine nitrogen cycle: recent discoveries, uncertainties and the potential relevance of climate change. *Philos TransR Soc B* **368**: 20130121–20130121.

Wernegreen JJ, Lazarus AB, Degnan PH. (2002). Small genome of *Candidatus* Blochmannia, the bacterial endosymbiont of *Camponotus*, implies irreversible specialization to an intracellular lifestyle. *Microbiology* **148**: 2551–2556.

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**: 27.

Yamamoto H, Miyake C, Dietz K-J, Tomizawa K-I, Murata N, Yokota A. (1999). Thioredoxin peroxidase in the Cyanobacterium *Synechocystis* sp. PCC 6803. *FEBS Lett* **447**: 269–273.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T et al. (2008). Globally distributed uncultivated oceanic N_2 -fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**: 1110–1112.

Zehr JP, Bench SR, Mondragon EA, McCarren J, DeLong EF. (2007). Low genomic diversity in tropical oceanic N_2-fixing cyanobacteria. *Proc Natl Acad Sci USA* **104**: 17807–17812.

Zehr JP, Kudela RM. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. *Annu Rev Mar Sci* **3**: 197–225.

Zehr JP, Mellon MT, Zani S. (1998). New nitrogen fixing microorganisms detected in oligotrophic oceans by the amplification of nitrogenase (*nifH*) genes. *Appl Environ Microbiol* **64**: 3444–3450.

Zehr JP, Paerl HW. (2008). Molecular ecological aspects of nitrogen fixation in the marine environment. In: Kirchman DL (ed) *Microbial Ecology of the Oceans*. Wiley-Liss, Inc.: Durham, NC, pp 481–525.

Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward GF et al. (2001). Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* **412**: 635–638.

Supplemental Methods

Sequence assembly and annotation

The sequencing run yielded a total of 4,387,138 paired-end 151-bp reads. These reads were initially assembled using Newbler (Chaisson and Pevzner, 2008) release 2.6, with default settings. This resulted in 18,336 contigs, ranging in length between 152,740 nt and around 500 nt. The 36 longest contigs of this assembly (1.3 Mb) all had high BLASTN similarity to the UCYN-A1 reference genome. In total, we obtained 167 contigs (1.47 Mb) with high similarity to UCYN-A1 from this initial de novo assembly. These 167 contigs were then taken as "seeds" in the second assembly step, using revision 0.18.2 of the PRICE (Paired-Read Iterative Contig Extension) assembler (Ruby et al., 2013). This resulted in a final set of 52 scaffolds (1,485,499 nucleotides), ranging in length between 249,164 and 675 nucleotides. The mean scaffold length was 28,567, with lengths distributed as shown in supplemental Fig. S1. The mean estimated coverage depth over all 52 scaffolds in this UCYN-A2 draft genome is 26.7, with a standard deviation of 26.4. (supplemental Fig. S2). Supplemental Fig. S3 shows a histogram of the number of positions by coverage depth.

Both UCYN-A1 and UCYN-A2 have two identical rRNA operons. This presents a challenge during assembly, because the assembly program is not able to correctly assign the 151 bp reads to two identical regions that are thousands of nucleotides long (and thus not spanned by paired-end information). This was resolved by generating the contigs containing the rRNA operons in two separate

assemblies, in which PRICE was seeded with Newbler-assembled reads that recruited to the flanking regions of the rRNA operons in the UCYN-A1 genome. This was sufficient to build two separate UCYN-A2 rRNA contigs.

Genes were identified using GeneMark (Lomsadze et al., 2005) with the annotation pipeline in the Integrated Microbial Genomes Expert Review (IMG-ER) system (Markowitz, 2009). Candidate gene homologs were identified using BLASTP with a 1e-2 value cutoff. Genes were assigned various annotations based on functional resources such as COG, Pfam, TIGRfam, Gene Ontology, and SwissProt. Further characterizations of the functional annotations were obtained from associating the genes with functional classifications including COG functional categories and the KEGG and MetaCyc pathway collections. Identification of tRNAs were performed using tRNAScan-SE-1.23 (Lowe and Eddy, 1997), and ribosomal RNAs were identified using BLAST searches against the database of non-redundant rRNAs from complete, finished IMG genomes. Other RNA genes were found by searching the genome for Rfam profiles using the program INFERNAL v.0.81 (Griffiths-Jones et al., 2005). Ortholog genes were identified by searching for bidirectional best BLASTP matches between predicted protein coding sequences of both genomes (highest found e-value 9 E-12). Partial genes at scaffold ends or parts of pseudogenes, i.e. UCYN-A2 genes that aligned to only a part of the UCYN-A1 sequence, were checked to match a particular ortholog based upon visual examination of the alignments. The larger total number of 1246 protein coding genes in the UCYN-A2 draft genome,

compared to 1200 in UCYN-A1, is due to unique UCYN-A2 genes, partial genes at scaffold ends and parts of pseudogenes, but also due to 5 UCYN-A2 genes (incl. 2 hypothetical proteins) that have near full-length matches in UCYN-A1 but were not annotated in the latter genome (Table 2 in main document).

The initial assembly process also generated 15,156 contigs which were not included in the draft, on the basis of redundancy or lack of similarity to UCYN-A1 or other closely related cyanobacteria. These contigs represent other organisms present in the population of sorted cells, and ranged in length from between 23 to 56,164 nucleotides, with a mean GC content of 44.7 ± 0.08 %. In order to confirm their rejection, we ran these contigs by BLAST against the nt database with a loose cutoff of 1e +1. 13,158 rejected contigs had no matches, 1153 contigs had 1 match, and 843 contigs had 2 or more matches (usually to different organisms, suggesting that the contigs contained chimeric or otherwise erroneous reads). Of 3998 total matches, 4 subjects were viruses, 3 subjects were Archaea; 68 could not be taxonomically assigned; the remaining 3923 were Bacteria. The predominant bacterial taxa were Flavobacteria (872 hits), Alphaproteobacteria (468 hits), Planctomycetes (213 hits), and Gammaproteobacteria (187 hits). 172 matches to UCYN-A1 and *Cyanothece* sp. 51142 were also detected, 160 of which were 100% identical to sequences already in draft contigs; the gene associated with each the remaining 12 hits (length 13 to 30 nucleotides) was determined by BLASTX against the nr protein

database, and in all cases it was ascertained that the gene was already present in the draft assembly.

Phylogenomic analyses

A multiple gene approach was used to identify the phylogenetic relationships of UCYN-A2 and UCYN-A1within the cyanobacterial tree. Sequence data for 57 cyanobacterial genomes were obtained from GenBank (http:// www.ncbi.nlm.nih.gov). Taxa were included to represent well-supported monophyletic groups previously described by recent phylogenomic studies (Sánchez-Baracaldo et al. in press; Shih et al., 2012). The 135 protein sequences analyzed are highly conserved and have undergone a minimum number of gene duplications (Blank and Sanchez-Baracaldo, 2010). These genes also represent a wide diversity of cellular functions. A detailed list and description of the genes can be found in Blank and Sánchez-Baracaldo (2010). Individual genes were aligned using SATé 2.2.3 (Liu et al., 2009). Single alignments were later concatenated into a phylip format matrix with a total of 56,251 amino acids. ProTest v.2.4 (Abascal et al., 2005) was used to estimate the best model of evolution for this protein set. Protein sequences we analyzed implementing the LG model + I (estimation of a proportion of invariable sites) and + G (gamma-distribution with 4 rate categories). Maximum likelihood analyses and bootstrap values were performed using RAxML 7.4.2 (Stamatakis, 2006).

Confirmation of pseudogenes and gene remnants by PCR amplification

A subset of the pseudogenes and gene remnants identified in the UCYN-A2 metagenome were PCR amplified to confirm that their presences are not artifacts e.g. of low coverage in certain regions of the assembly. PCR primers were designed using Primer3 (Supplemental Table 1; http://primer3plus.com/cgi-bin/dev/primer3plus.cgi; Untergasser et al., 2012), and specificity was confirmed using both Primer-BLAST and via BlastN of all UCYN-A2 contigs. All PCR reactions were carried out in replicate using 2 μL of 1:100 dilutions of the UCYN-A2 whole genome amplification product in 25 μL total volumes, with 1.5 U Platinum® Taq DNA Polymerase (Invitrogen, Carlsbad, CA, USA), 1X PCR buffer, 4 mM $MgCl_2$, 400 μM dNTPs mix, and 0.5 μM of each primer. Thermocycling parameters used a two step protocol of 30 cycles of denaturation at 95°C for 30 sec and annealing at 59.1°C for 30 sec. PCR amplicons were cleaned prior to sequencing or cloning using the QIAquick Gel Extraction kit (Qiagen, Valencia, CA, USA) according to manufacturer's guidelines. Cleaned amplicons were either sent directly for bidirectional sequencing (using the PCR primers as sequencing primers), or ligated overnight and cloned using the pGEM®-T Vector Systems (Promega, Madison, WI, USA), according to manufacturer's guidelines. All PCR amplicons were bidirectionally sequenced using Sanger technology at the UC Berkeley DNA Sequencing Facility.

All raw sequences were trimmed of vector contamination and poor quality base calls using Sequencher® 5.1 (Gene Codes Corporation, Ann Arbor, MI, USA), and aligned to the target sequences from the UCYN-A2 metagenome for confirmation in MEGA v5.2 (Tamura et al., 2011).

Supplemental References

Abascal F, Zardoya R, Posada D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**(9):2104-2105.

Blank CE, Sanchez-Baracaldo P. (2010). Timing of morphological and ecological innovations in the cyanobacteria - a key to understanding the rise in atmospheric oxygen. *Geobiol*. **8**(1):1-23.

Chaisson MJ, Pevzner PA. (2008). Short read fragment assembly of bacterial genomes. *Genome Research* **18**(2):324-330.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. **33**(suppl 1):D121-D124.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. (2009). Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science* **324**(5934):1561-1564.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. **33**(20):6494-6506.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res*. **25**(5):0955-0964.

Markowitz VM (2009). IMG ER: A System for Microbial Genome Annotation Expert Review and Curation. *Bioinformatics* **25**(17): 2271-2278.

Ruby JG, Bellare P, DeRisi JL. (2013). PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3: Genes|Genomes|Genetics* **3**(5):865-880.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, et al. (2012). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *PNAS* **110**(3): 1053-1058.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21):2688-2690.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**: 2731-2739.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M and Rozen SG. Primer3--new capabilities and interfaces. *Nucleic Acids Res*. 2012 Aug 1;**40**(15):e115.

Figures and Tables

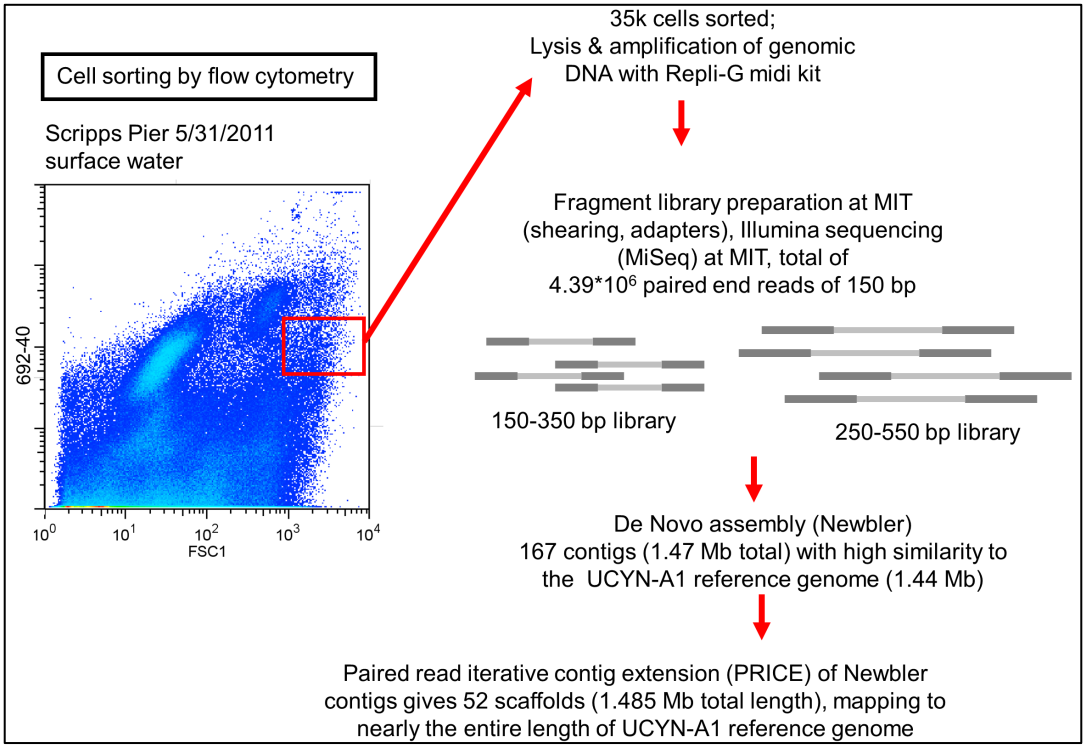


Figure 1. Work flow diagram describing the cell-sorting, genome-sequencing and assembly approach used in this study. The chosen FCM sort gate was determined in earlier experiments by screening different sorted populations for the presence of UCYN-A2 nifH by qPCR, as described previously. The PRICE assembly was carried out as described in Ruby et al. (2013).

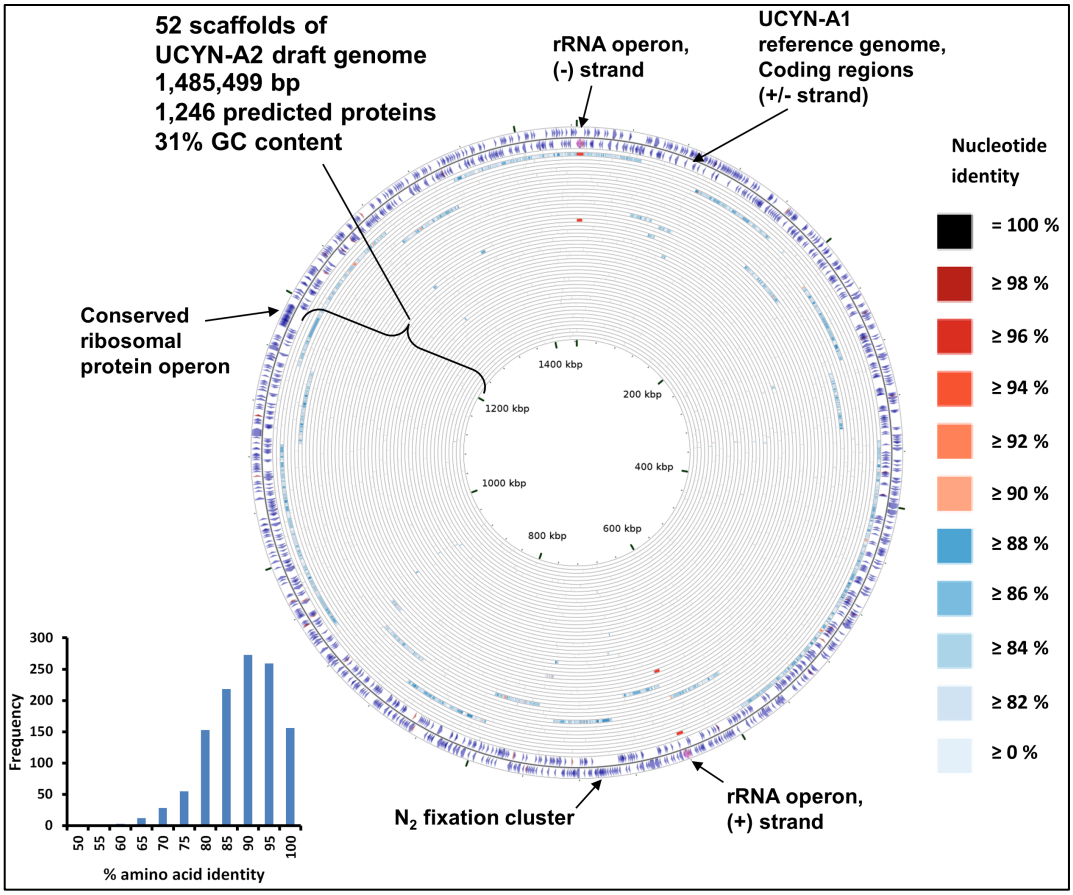


Figure 2. Circular map showing all 52 scaffolds of the UCYN-A2 draft genome aligned to the UCYN-A1 chromosome. Each concentric ring represents a scaffold, with the color code representing percent nucleotide identity. The scaffolds are sorted by length, with the longest scaffold (249 164 nt) on the outermost ring, and decreasing in length towards the center ring (shortest scaffold of 675 nt). The inlet graph is a histogram of percent amino-acid identity for all 1159 ortholog genes.

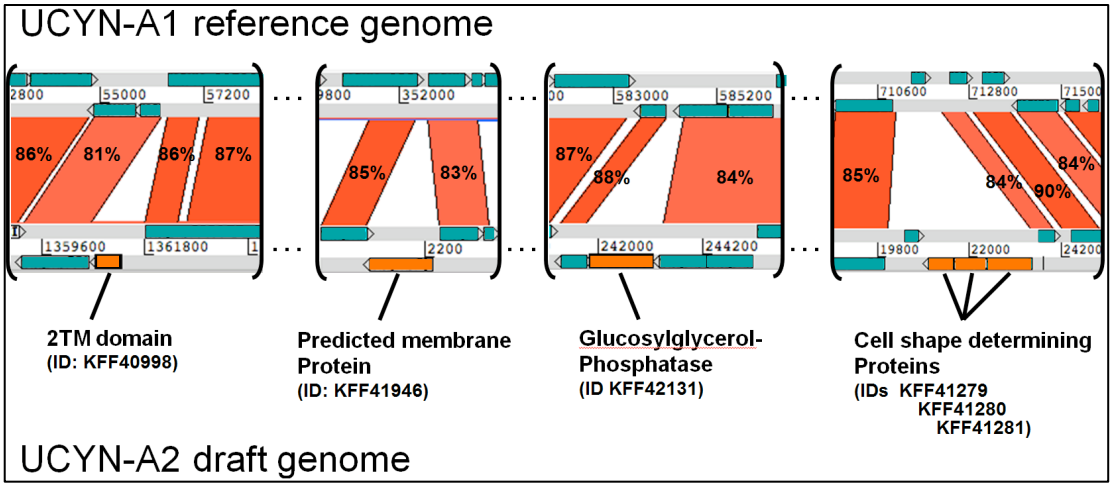


Figure 3. Examples of missing genes in UCYN-A1, demonstrating the resulting genome compaction. A total of 31 genes was found to be unique in UCYN-A2. The alignment was done using the Artemis Comparison Tool (http://www.sanger.ac.uk/) and shows closely matching gene neighborhoods apart from the missing genes (percent nucleotide identity given for aligned genes).

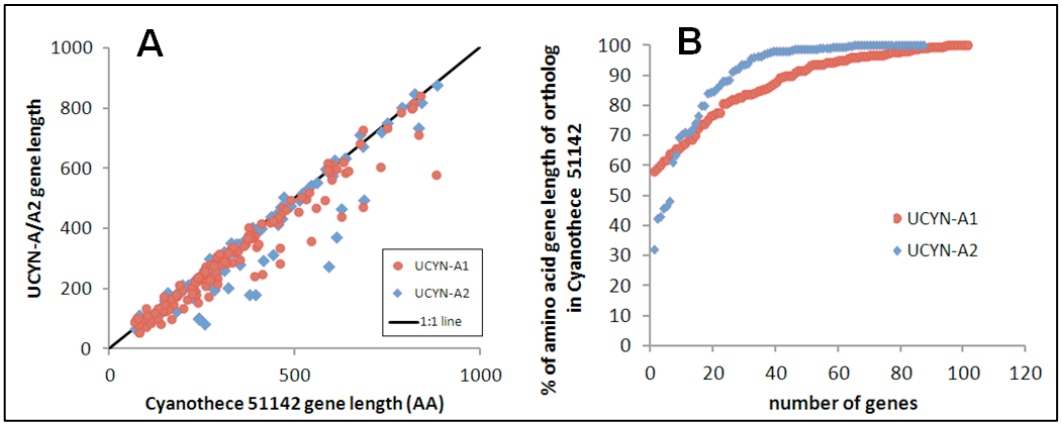


Figure 4. (a) Comparison of amino-acid lengths of ortholog genes in UCYN-A1, UCYN-A2 and Cyanothece sp. 51142. (b) The range of percent gene length of the UCYN-A1 and UCYN-A2 orthologs compared with the Cyanothece sp. 51142 orthologs.

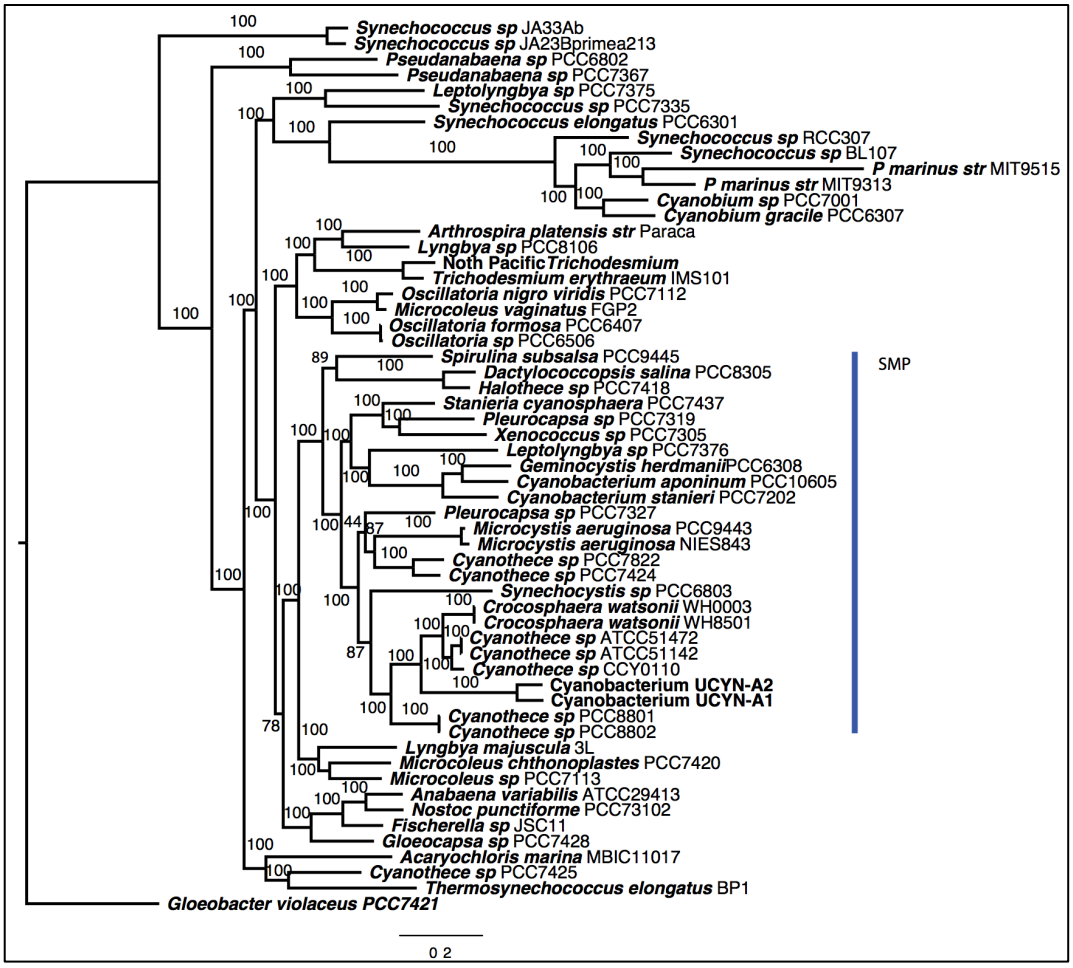


Figure 5. Phylogeny of 57 cyanobacteria based on a concatenated alignment of 135 highly conserved protein sequences. A detailed list and description of the genes can be found in Blank and Sanchez-Baracaldo (2010). Maximum likelihood analyses were performed using RAxML 7.4.2 (Stamatakis, 2006). Bootstrap values are indicated above branches. The vertical bar marks sequences belonging to a strongly supported clade of marine unicellular N_2 fixers previously described as the SPM group.

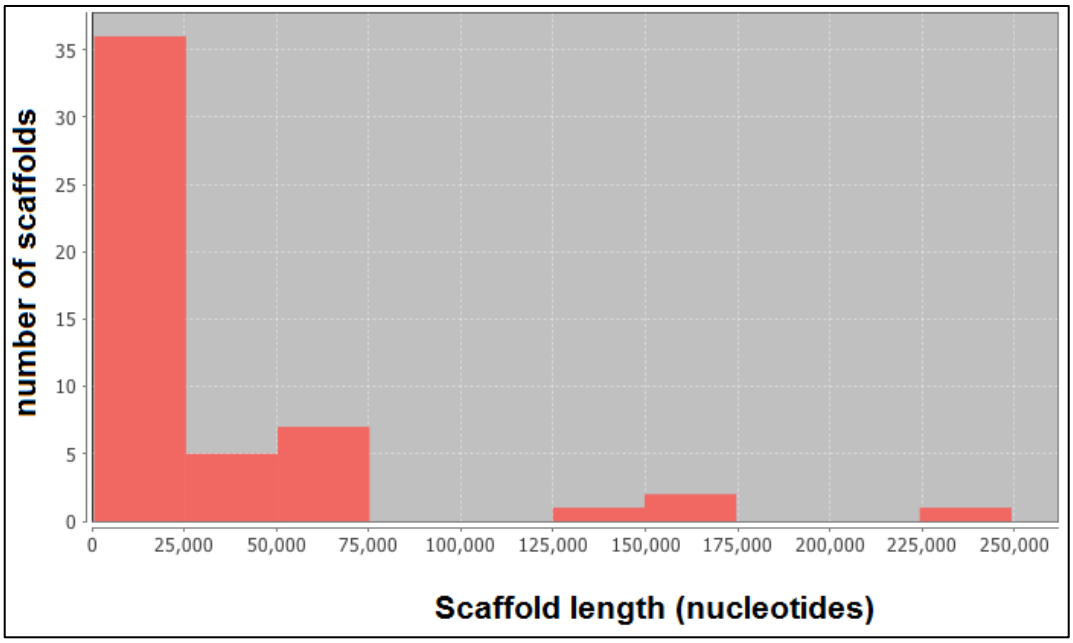


Figure S1: Frequency of UCYN-A2 draft scaffolds by length.

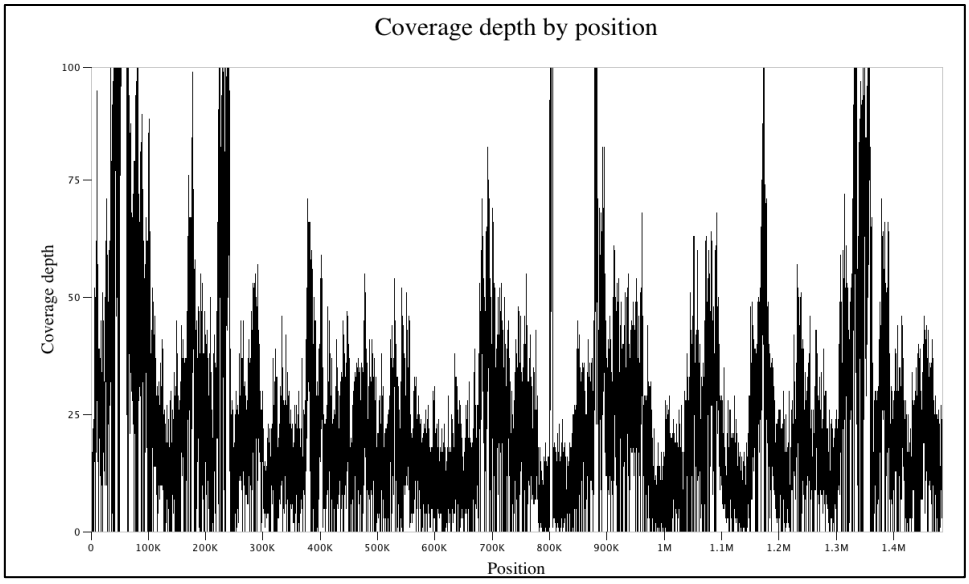


Figure S2. Estimated coverage depth for each nucleotide position on the 52 scaffolds of the UCYN-A2 draft genome

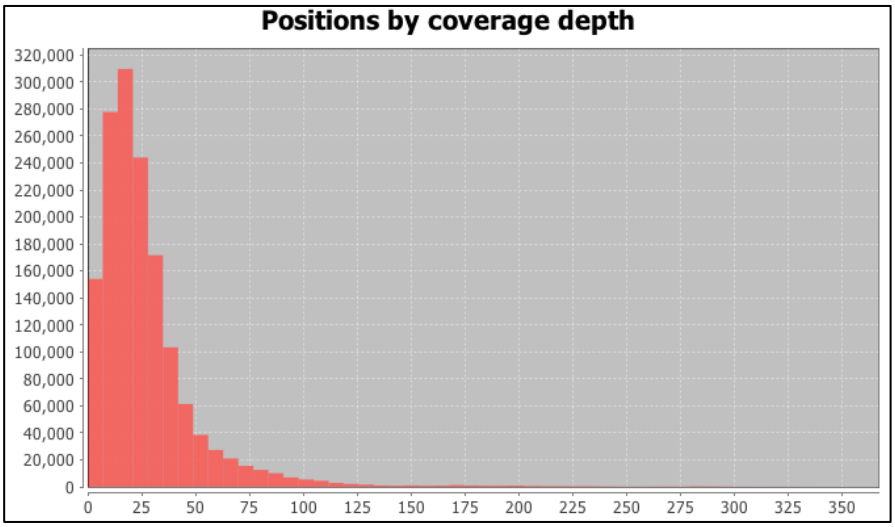


Figure S3. Histogram of the number of nucleotide positions having a particular depth of coverage on the 52 scaffolds of the UCYN-A2 draft genome.

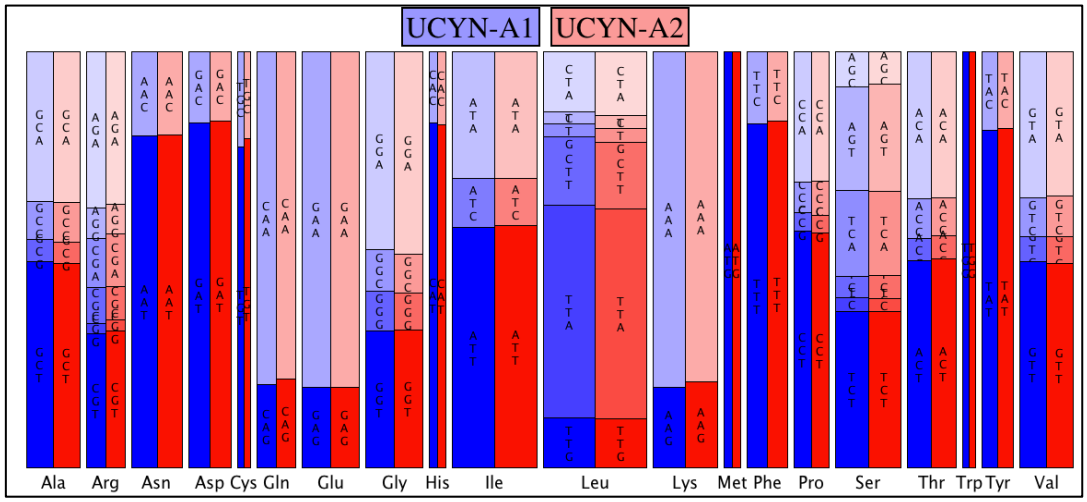


Figure S4. Codon usage of protein-coding genes in UCYN-A1 (blue) and UCYN-A2 (red). For each amino acid, bar width is proportional to prevalence of that amino acid. Height of each vertical segment within each bar is proportional to prevalence of the corresponding codon.

	UCYN-A1	UCYN-A2
Location	HOT station, 22. January 2008	Scripps Pier, 31. May 2011
Genome size	1443806	1485499
Number of scaffolds	1	52
GC %	31	31
Coding base count %	81.42	79.32
Protein coding genes	1200	1246
RNA genes	42	49
rRNA genes	6	6
5S RNA genes	2	2
16S RNA genes	2	2
23S RNA genes	2	2
tRNA genes	36	37
Other RNA genes		6

Table 1. Genome statistics of UCYN-A1 and UCYN-UCYN-A2

Category	*GenBank accession*	Gene length (AA)	Annotation	Function Description
UCYN-A1	YP_003421868	159	Peroxiredoxin	protein related to alkyl hydroperoxide reductase
genes that	YP_003421145	167	Restriction endonuclease	Defense
are possible	YP_003421558	207	HAS barrel domain protein	Domain in ATP synthases
pseudogenes in UCYN-A2	YP_003421659	398	NurA domain-containing protein	NurA domain, endo- and exonucleases
	YP_003421689	103	NifZ domain-containing protein	N_2 fixation, nif operon
	YP_003422000	318	Transcriptional regulator, GntR family	transcription factors, possibly regulation of primary metabolism
UCYN-A2	KFF41946	371	Predicted membrane protein	Function unknown
genes	KFF42131	430	Glucosylglycerol phosphatase (EC 3.1.3.69)	Osmoprotectant synthesis
absent in	KFF41831	236	Tellurite resistance protein	Contains C-terminal domain of Mo-dependent nitrogenase
UCYN-A1	KFF41325	208	Thymidylate kinase	Pyrimidine metabolism, DNA synthesis
	KFF41279	347	Cell shape-determining protein, MreB/Mrl family	Cytoskeleton synthesis, cell shape determination
	KFF41280	248	Rod shape-determining protein MreC	Cytoskeleton synthesis, cell shape determination
	KFF41281	186	Rod shape-determining protein MreD	Cytoskeleton synthesis, cell shape determination
	KFF41062	427	Folate/biopterin transporter	Membrane transport
	KFF40998	165	2TM domain	Function unclear, transmembrane alpha helixes
	KFF41013	56	Sigma-70, region 4	DNA directed RNA polymerase
	KFF41656	344	Folate-binding protein YgfZ	Predicted aminomethyltransferase, possibly glycine synthesis
	KFF41014	63	Sigma-70 region 3	DNA directed RNA polymerase
	KFF40922	215	Peroxiredoxin	Detoxification of active oxygen species such as H2O2
	KFF41590	231	Zn-dependent hydrolases, including glyoxylases	Pyruvate metabolism
	KFF40927	277	Tetratricopeptide repeat/TPR repeat	Unclear function- involved in chaperone, cell-cycle, transciption, and protein transport complexes
	KF41183	94	RNA-binding proteins (RRM domain)	Function unclear
UCYN-A2 genes that	KFF41758	38	Cytochrome B6-F complex subunit 5	Photosynthesis, connects PSI and PSII in e^- transport chain
match unannotated	KFF41141	64	LSU ribosomal protein L33P	Structural constituent of ribosome
ORFs in UCYN-A1	KFF41382	470	Hemolysins and related proteins containing CBS domains	Membrane protein, regulate activity of associated enzymatic transporters

UCYN-A2 genes that are possible pesudogenes	KFF41284	211	Uncharacterized protein, similar to the N-terminal domain of Lon protease	Proteolysis
in UCYN-A1	KFF41208	165	Predicted RNA-binding protein	General function prediction only
	KFF41109	86	Glutaredoxin-like domain (DUF836)	Domain of unknown function
	KFF41037	267	Helix-turn-helix domain	DNA binding, gene expression regulation
	KFF40997 (2)	461	Domain of unknown function (DUF697)	Function unknown
	KFF41565 (2)	301	CAAX protease self-immunity	Probably protease, transmembrane protein
	KFF41236 (2)	396	Glycosyltransferases involved in cell wall biogenesis	Cell wall/membrane/envelope biogenesis
	KFF42055 (2)	350	UDP-N-acetylglucosamine-*N*-acetylmuramylpentapeptide *N*-acetylglucosamine transferase	Cell wall/membrane/envelope biogenesis
	KFF41265 (2)	294	Competence/damage-inducible protein CinA C-terminal domain	Transformation
	KFF41488 (2)	196	Putative translation factor (SUA5)	Translation, ribosomal structure and biogenesis
	KFF41875(2)	140	Predicted endonuclease involved in recombination (possible Holliday junction resolvase in *Mycoplasma* and *Bacillus subtilis*)	Replication, recombination, and repair
	KFF41338 (2)	600	Subtilisin-like serine proteases	Proteolysis or cell motility
	KFF42033 (2)	385	Phosphate ABC transporter substrate-binding protein, PhoT family (TC 3.A.1.7.1)	Inorganic ion transport and metabolism

Table 2. Annotated genes absent or possibly pseudogenes in other genome. The table also shows three annotated genes in UCYN-A2 that match unannotated regions in UCYN-A1.

Pseudogene/gene remnant target	Confirmed feature	Forward primer (5' - 3')	Reverse primer (5' - 3')
UCYN-A2_contig_1.1 191000..192205	Two partial genes that align with UCYN-A1 gene 646530270 with the interruption of a stop codon	GTCCCATCGTTCGCCCTTAT	GCAGGAGCATTAAATTTCCGTCA
UCYN-A2_contig_6.6 52398..54883 (reverse complement)	Three partial genes that align with CYN-A1 gene 646530499 with the interruption of stop codons	CTATAAAAAGCAAATGATAAGGCAAC	GCTAGTAAATGTGGACAACCTCTT
UCYN-A2_contig_2.2 39750..40980	Single gene that aligns with UCYN-A1 gene 646530695 along 86% of the length	ATTAAGTTTGATGTATTGAAGCTAGT	GGATTCAACTATACGGCCCA
UCYN-A2_contig_7.7 12430..13690	Single gene that aligns with UCYN-A1 gene 646530039 along 88% of the length	CTTCTGGAGCAAAGGGACGT	TCTTCTGCTGCTTCTGCACC
UCYN-A2_contig_13.13 9472..10659	Single gene that aligns with UCYN-A1 gene 646530553 with an early stop codon	TGTCAGGAGAGTAAGAATGAAATGG	AGGATCTCGATCAAAAATTATACGA
UCYN-A2_contig_4.4 26223..27277	Single gene that aligns with UCYN-A1 gene 646530886 with different start codon	TCGCCAAGAGGATGATCTGT	CCTTAAGCTCTTCAGGGGTCA
UCYN-A2_contig_2.2 19385..20631	Single gene that aligns with UCYN-A1 gene 646530711 with an early stop codon	AGTCAATATAGAGCAGTGCAGGA	CCGCACCATTAGGACCAAGT
UCYN-A2_contig_16.16 5100..6751	Remnant of UCYN-A1 gene 646530363	GCAATAGCTAGCGATGTGTT	ACCTAAGGGAAGAAGCAACT
UCYN-A2_contig_3.3 131872..133802	Remnant of UCYN-A1 gene 646530983	ACCTCACCAATGTCTATGCTGA	GACTTTATCTTCCAAAACCTCCACA

Table S1. PCR primers.

Addendum to Chapter 2: 16S Ribosomal RNA Assembly

The remainder of this chapter describes research I performed during the assembly phase of the UCYN-A2 project described in the preceding article. The original assembly was unable to correctly recover the UCYN-A2 genome's two 16S ribosomal RNA sequences due to inherent limitations in paired-end sequencing technology. I developed a computational approach that recovered both 16S rRNA sequences.

Addendum Introduction

Cyanobacteria are Bacteria that are oxygenic phototrophs and fix carbon dioxide. They typically contain two copies of the 16S ribosomal RNA gene in their genome in opposite orientations. The duplication is also present in chloroplasts (Palmer, 1983). The divergence of chloroplasts from cyanobacteria approximately 800 – 900 million years ago (Shih & Matzke, 2013; McFadden and van Dooren, 2004) was first proposed by Mereschowsky in 1905 (Martin & Kowallik, 1999), and has been verified by molecular methods (Giovannoni et al., 1988).

Given the ubiquity and age of the rRNA duplication and its presence in UCYN-A1 (Tripp et al., 2010) and the closely related *Cyanothece* sp. ATCC 51142, it was reasonable to expect that UCYN-A2 also contains the duplication. However, neither the initial assembly of the UCYN-A2 metagenome using Newbler (Chaisson & Pevzner, 2008) nor the subsequent Newbler+PRICE (Ruby et al., 2013) assembly produced a second copy of 16S rRNA. The assembly is difficult because of the nearly identical sequences of the rRNA genes which are bracketed by different sequences, causing difficulty in determining the correct assembly of the two copies with the bracketing regions of the genome. For the assembly in UCYN-A1, the duplication had to be confirmed by PCR. The procedure described below was developed to confirm whether UCYN-A2 contains the duplication and, if so, to accurately assemble both copies, without the need for PCR.

Addendum Methods

After confirming the likelihood of a16S rRNA duplication in UCYN-A2 by analysis of the stoichiometry of reads and genome coverage, the overall approach (Figure A1) was to assemble a consensus UCYN-A2 16S rRNA sequence. The reads affiliated with the flanking regions of the forward-strand 16S rRNA gene in UCYN-A1 were collected and assembled with the 16S rRNA consensus sequence. The process was repeated for the reverse-strand 16S rRNA gene (Figure A2).

Stoichiometric analysis was performed by comparing Newbler coverage depth of sequence reads in the 16S rRNA gene region to the average coverage depth of reads across all contigs excluding the 16S rRNA gene.

The UCYN-A1 16S rRNA sequence was submitted as query for a BLASTN search (Altschul 1990) of a database of all metagenomic sequence reads in the study. Subject reads that were identified by BLASTN with at least 97% sequence similarity across at least 97% of their length were collected and assembled using Newbler to produce a consensus UCYN-A2 16S rRNA sequence. The upstream and downstream flanking sequences of the UCYN-A1 forward-strand 16S rRNA gene were then submitted as queries in a BLASTN search of the database of reads. Subject reads identified by BLAST E-values <= 0.001 were collected and assembled using PRICE (Ruby 2013), using the UCYN-A2 consensus 16S rRNA sequence as a seed. Assembly produced a contig containing the consensus UCYN-A2 16S rRNA gene as well as upstream and downstream flanking regions for the forward strand. The process was repeated using the flanking sequences of the UCYN-A1 reverse-strand 16S rRNA gene, to produce a contig containing the consensus UCYN-A2 16S rRNA gene as well as upstream and downstream flanking regions for the reverse strand. Finally the two contigs were validated by BLASTN comparison against the GenBank nt database (Benson 2004). The set of seed contigs described in the “Supplemental Methods” section of the article was

modified by replacing the original 16S rRNA contig with the two contigs generated by the procedure described here. A final assembly was performed using PRICE and the modified set of seed contigs.

Addendum Results

The stoichiometric analysis showed that average coverage of the original (single) 16S rRNA sequence was 109 reads, compared to 27 reads for the overall genomic sequences.

For assembly of the consensus UCYN-A2 16S rRNA sequence, the BLAST search of the UCYN-A1 16S rRNA sequence against all reads produced 3220 reads with >=97% identity across >=97% of read length (Figure A1). These were assembled into a consensus sequence of 4460 nucleotides, with 99% identity to the UCYN-UCYN-A1 16S rRNA gene. 47 and 348 reads (respectively) recruited to the UCYN-A1 upstream and downstream flanking regions of the forward-strand gene; assembling these with PRICE using the UCYN-A2 consensus 16S rRNA sequence as a seed produced a contig of length 6603. 167 and 613 reads (respectively) recruited to the UCYN-A1 reverse-strand upstream and downstream flanking regions; assembling these with PRICE using the UCYN-A2 consensus 16S rRNA sequence as a seed produced a contig of length 8766.

BLASTN comparison of the forward-strand contig against GenBank's nt database returned a best match to the UCYN-A1 genome, with 95% identity over 93% of query length. The second-best match was to *Cyanothece* sp. ATCC 51142, with 93% identity over 73% of its length. BLASTN comparison of the reverse-strand contig against the GenBank nt database returned a best match to the UCYN-A1 genome, with 95% identity over 98% of query length. The second-best match was to *Cyanothece* sp. ATCC 51142, with 93% identity over 55% of length.

The original set of seed contigs contained 167 contigs generated by Newbler as described in the article. The single original contig containing the 16S rRNA sequence was removed from the seed set and the two new contigs were added. PRICE assembly with 168 seeds resulted in a final group of contigs from which the 52 contigs submitted as the draft genome were selected.

Addendum Discussion

Using the method presented here, two duplicate UCYN-A2 16S rRNA contigs were successfully assembled from metagenomic short paired-end reads. Assembly of such duplicated regions is challenging because assembly will only produce the correct copy number of a repeated motif if the motif is shorter than the amplicon length. Since there was a duplicated 16S rRNA region in the UCYN-A genome, only a single 16S rRNA copy was computed, and it was necessary to

use the reference-based approach described here to correctly assemble the two copies.

The stoichiometric analysis showed 4x deeper coverage of the 16S rRNA gene than the average coverage of the rest of the UCYN-A2 genome. While this is consistent with a16S rRNA duplication, it is also consistent with an even higher copy number, and it could be possible that additional copies are present in the genome but absent from the assembly. The possibility, however, is remote, since the additional copy or copies would have to be located in the regions between contigs of the UCYN-A2 assembly; these regions have been investigated by bioinformatic analysis and *in vitro* (see Discussion section of the article) and no evidence of additional 16S rRNA gene copies has been detected in UCYN-A.

The existence of a single 16S rRNA gene in the initial assembly is the result of an inherent limitation of short-read assembly. If the original genome contains repeats of a motif that is longer than the reads, the short-read library is uninformative regarding the number of copies of the repeated motif (Green 2002; Alkan et al., 2010). Paired-end technology can overcome the problem when repeated motif lengths are shorter than the amplicon length; however the UCYN-A2 16S rRNA gene is too long for paired-end assembly to detect and resolve the duplication. The approach presented here takes advantage of *a priori* knowedge of the UCYN-A1 reference genome to assign reads that flank the copies of 16S rRNA to the correct contig prior to assembly. The identity between

corresponding flanking regions in UCYN-A1 compared to UCYN-A2 is high (95%, similar to the identity between corresponding *nifH* regions), suggesting that the UCYN-A2 contigs are valid.

The consensus 16S rRNA sequence (green sequence "D" in Figure A1 and "F" in Figure A2) was used in the assembly of both contigs ("I" in Figure A1 and "G" in Figure A2). Thus, the draft UCYN-A2 genome contains two 100% identical 16S rRNA sequences. This identity, which may reflect reality (the two copies in UCYN-A1 are identical), has not been confirmed by a biomolecular approach and for now should be considered an artifact of the assembly methodology. UCYN-A is an important source of fixed nitrogen in the oceans (Montoya et al. 2004) and has a broader latitudinal range than *Trichodesmium* (Moisander et al., 2010). It is widely distributed: in the Pacific Ocean it has been detected in the oligotrophic North Pacific Ocean (Zehr et al., 2008), the North Pacific Transitional Zone (Church et al., 2008), the North Equatorial Counter Current (Church et al., 2008), the tropical and subtropical oligotrophic South Pacific Ocean (Moisander et al., 2010), the Great Barrier Reef (Hewson et al. 2007), and the coastal waters of Japan (Hagino et al., 2013) and California (Bombar et al., 2014). In the Atlantic Ocean, UCYN-A has been detected in western tropical oligotrophic waters (Langlois et al., 2008), eastern tropical coastal and oligotrophic waters (Turk et al., 2011), and eastern subtropical oligotrophic waters (Langlois et al., 2008; Agawin et al., 2014). However, little is known about the genetic diversity of

UCYN-A. To date, three distinct clades of UCYN- A have been identified, and UCYN-A *nifH* sequences not affiliated with any of these clades have been observed (Thompson et al., 2014). This study showed that these clades are distinctly different based on genome sequences, which has implications for their ecology and evolution of the symbiosis. The comparison was dependent on assembling a genome with duplications of the 16S rRNA gene which presented a challenge for the short reads from current high-throughput DNA sequencing techniques. The method described here enabled the comparison of these two genomes even though they had identical duplicated 16S rRNA genes. Further studies are likely to reveal additional strains of UCYN-A, for which the metagenomes can be analyzed by the methods described in this chapter.

Addendum References

Agawin, N. S. R., Benavides, M., Busquets, A., Ferriol, P., Stal, L. J., & Arístegui, J. (2014). Dominance of unicellular cyanobacteria in the diazotrophic community in the Atlantic Ocean. *Limnology and Oceanography*, *59*(2), 623–637. doi:10.4319/lo.2014.59.2.0623.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1), 61–65. doi:10.1038/nmeth.1527.

Altschul S, et al. (1990). Basic Local Alignment Search Tool. *J Mol Biol* 215, 403-410. Benson DA. (2004). GenBank: update. *Nucleic Acids Res* **32**:23D–26.

Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B., & Carpenter, E. J. (1997). Trichodesmium, a Globally Significant Marine Cyanobacterium. *Science*, *276*(5316), 1221–1229. doi:10.1126/science.276.5316.1221.

Chaisson, M. J., & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome Research*, *18*(2), 324–330. doi:10.1101/gr.7088808.

Giovannoni, S. J., Turner, S., Olsen, G. J., Barns, S., Lane, D. J., & Pace, N. R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. *Journal of Bacteriology*, *170*(8), 3584–3592.

Green, P. (2002). Whole-genome disassembly. *Proceedings of the National Academy of Sciences*, *99*(7), 4143–4144. doi:10.1073/pnas.082095999.

Mahaffey, C., Michaels, A. F., & Capone, D. G. (2005). The conundrum of marine N2 fixation. *American Journal of Science*, *305*(6-8), 546–595. doi:10.2475/ajs.305.6- 8.546.

Martin, W., & Kowallik, K. (1999). Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *European Journal of Phycology*, *34*(3), 287–295. doi:10.1080/09670269910001736342.

McFadden, G.I. and van Dooren, G.G.(2004). Evolution: Red Algal Genome Affirms a Common Origin of All Plastids. *Current Biology* Vol. 14, R514–R516 DOI 10.1016/j.cub.2004.06.041.

Montoya, J., Joll, C.M., Zehr, J.P., Hansen, A., Villareal, T.A., & Capone, D.G. (2004). High rates of N2 fixation by unicellular diazotrophs in the oligotrophic Pacific

Ocean. *Nature 430* (1027-1031).

Moisander, P. H., Beinart, R. A., Hewson, I., White, A. E., Johnson, K. S., Carlson, C. A., et al. (2010). Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. *Science*, *327*(5972), 1512–1514. doi:10.1126/science.1185468.

Palmer, J. D. (1983). Chloroplast DNA exists in two orientations. *Nature*, *301*(5895), 92–93. doi:10.1038/301092a0.

Ruby, J. G., Bellare, P., & Derisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G*enes* | *Genomes*| *Genetics*, *3*(5), 865–880. doi:10.1534/g3.113.005967.

Shih, P. M., and Matzke, N. J. (2013). Primary endosymbiosis events data to the later Proterozeric with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences, 11(30), 12355-12360.* doi:10.1073/pnas.1305813110.

Thompson, A., Carter, B. J., Turk-Kubo, K., Malfatti, F., Azam, F., & Zehr, J. P. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environmental Microbiology*, n/a–n/a. doi:10.1111/1462- 2920.12490.

Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., et al. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, *464*(7285), 90–94. doi:10.1038/nature08786.

Addendum Figures

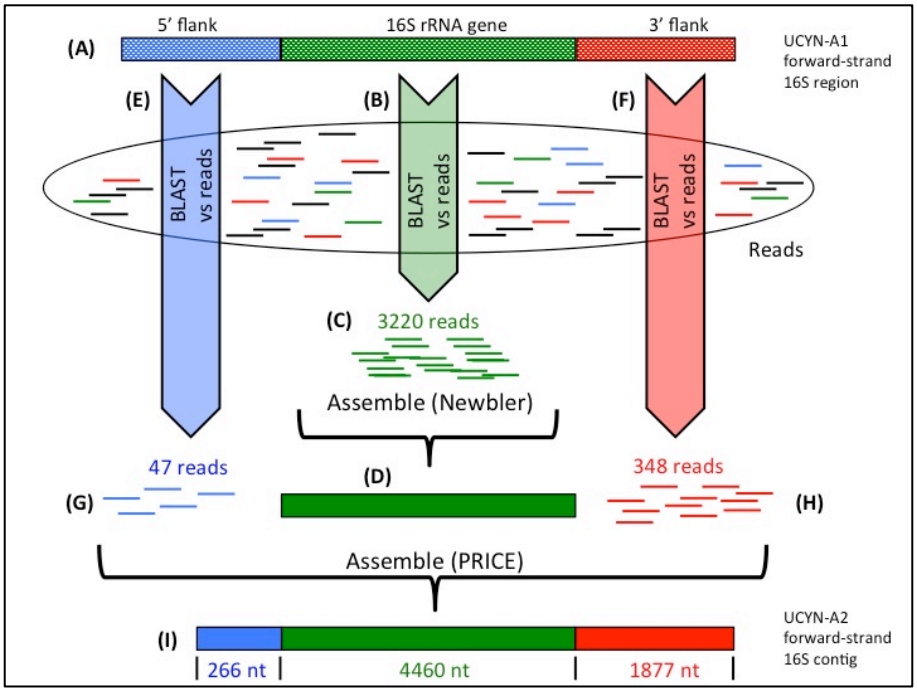


Figure A1. Assembly of forward-strand 16S rRNA contig. (A) Sequences for UCYN-A1 forward-strand 16S rRNA gene (green) and flanking regions (blue, red) were extracted from the UCYN-A1 assembly. (B) The UCYN-A2 16S rRNA gene was submitted as a BLASTN query against all reads, and matching reads (C) were assembled to produce the UCYN-A2 16S rRNA consensus sequence (D). UCYN-A1 flanking sequences (E, F) were submitted as BLASTN queries against all reads; matching reads (G, H) were assembled using PRICE with the UCYN-A2 16S rRNA consensus sequence (D) as a seed, to produce the final UCYN-A2 forward-strand 16S rRNA contig (I).

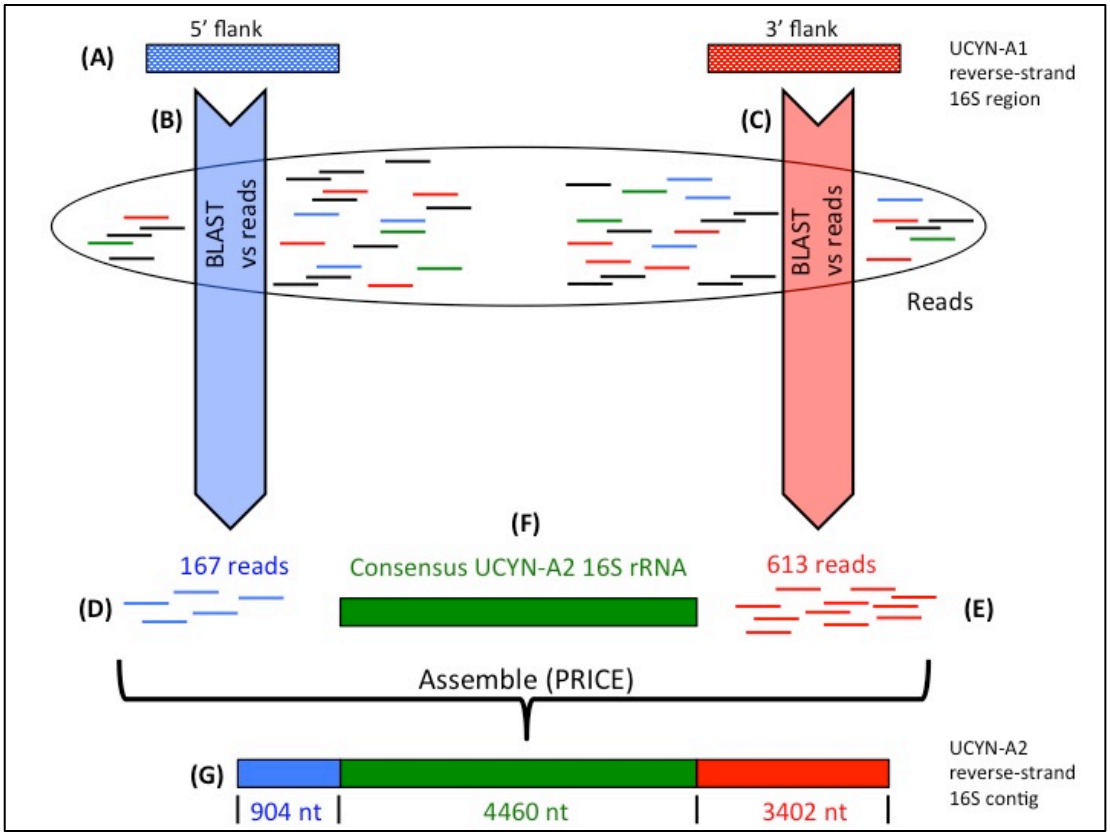


Figure A2. Assembly of reverse-strand 16S rRNA contig. (A) Sequences for UCYN-A1 reverse-strand 16S rRNA gene flanking regions (blue, red) were submitted as BLASTN queries against all reads (B, C); matching reads (D, E) were assembled using PRICE with the UCYN-A2 16S rRNA consensus sequence (F) as a seed, to produce the final UCYN-A2 reverse-strand 16S rRNA contig (G).

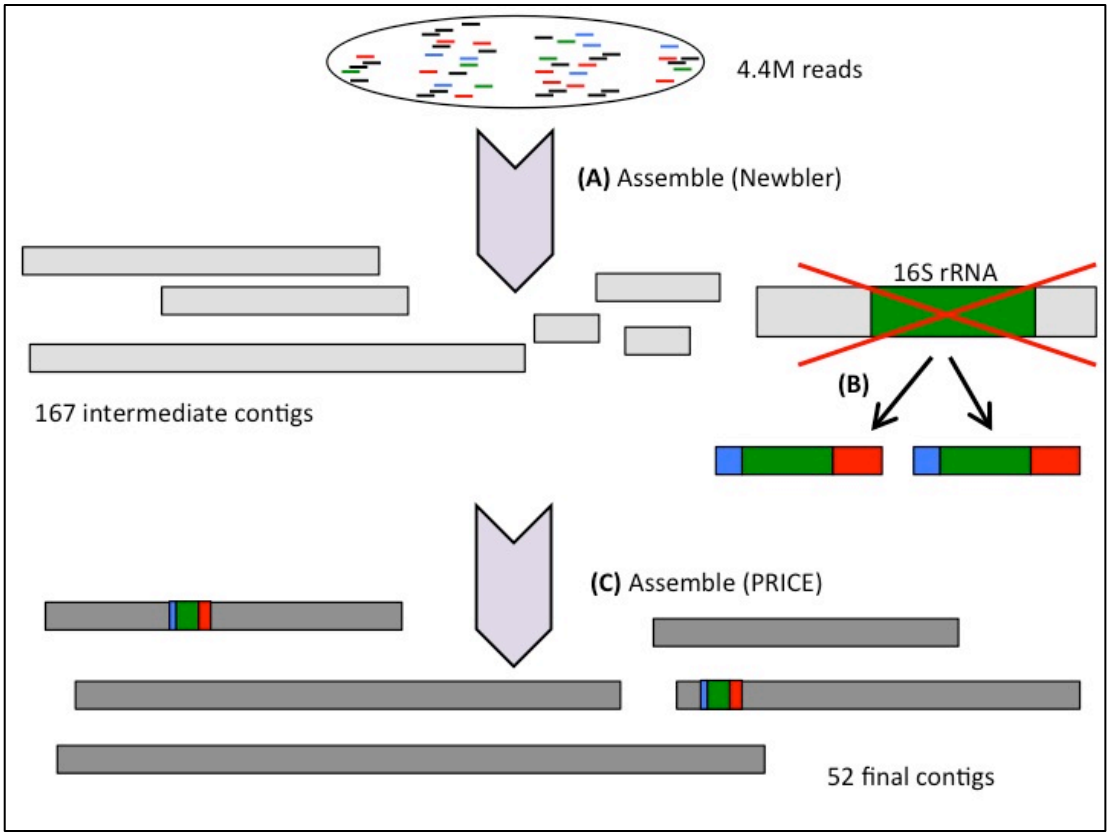


Figure A3. Overall assembly procedure. (A) Newbler assembly produces contigs (light gray), including one contig containing a 16S rRNA sequence. (B) the 16S rRNA contig is replaced by the two contigs generated by the procedure described in this addendum. (C) PRICE assembly produces final contigs (dark gray).

Chapter 3 – Dexter: A tool for exploring diel expression data sets

Abstract

Gene expression in cells can fluctuate over time in response to internal or external stimuli. For example, phototrophic cyanobacteria, whose metabolism is tightly coupled to the daily cycle of sunlight, have evolved complex daily patterns of gene expression that may derive from optimizing growth by coordinating photosynthesis and protein synthesis. The underlying diel gene expression patterns, which can be a result of circadian rhythms and/or shifts in metabolism driven by light, can provide much information on how microorganisms grow and respond to the environment. However, it is difficult to analyze and compare complex information from whole-genome expression studies involving thousands of genes from multiple organisms and multiple time points. Novel computational approaches are required, with support for visual exploration and analysis of data in order to detect related expression signatures.

A Java application called Dexter was developed to facilitate analysis of single or multiple expression time-series data sets. Dexter makes possible a wide variety of analyses; the value of the program was demonstrated by using it to explore gene expression similarity within predicted operons in the genomes of several cyanobacteria for which diel expression studies are available: *Crocosphaera*,

Prochlorococcus, and *Trichodesmium.* Current operon predictions appear to be correct but possibly incomplete. Analysis using Dexter demonstrated the existence of many pairs of short adjacent predicted operons, with highly similar expression profiles among all genes of both operons. A statistical framework was developed for estimating the probability that such pairs of predicted operons actually constitute single operons. Statistical evidence was found to support merging 35 pairs of neighboring predicted operons.

Background

Daily cycling of behavior, metabolic processes, and gene expression have been observed in all three domains of life. For example, in (Bergersen, 1962) mammals, daily periods of high levels of movement locomotion are often timed so as to optimize contact with food or avoid contact with predators (Panda et al., 2002). In plants and bacteria production of photosystem II proteins, whose half-lives range from 30 minutes to 12 hours (Yao, Brune, Vavilin, & Vermaas, 2012); Renger, coincides with available light (Dodd et al., 2005). Diazotrophic cyanobacteria, which both photosynthesize and reduce atmospheric nitrogen (N_2), need to segregate nitrogenase enzymes from the oxygen evolved by photosynthesis (Parker & Scutt, 1960; Bond, 1961; Bergersen, 1962; Fay, 1992); segregation is sometimes temporal, with nitrogenase component proteins

produced hours out of phase from photosystem II proteins (Tuit et al., 2004).

In all these examples, daily cycling is orchestrated not only by external cues such as light or temperature but by internal circadian biological clocks that can continue to function for some time in the absence of light or temperature triggers (Panda et al., 2002). Circadian clocks confer advantage by optimizing the interactions between external dark/light conditions and internal metabolic processes (Dodd et al., 2005). Cells whose internal clock mechanisms have been disrupted are not competitive with wild-type cells (Johnson 1998; Mori and Johnson 2001; Woelfle2004; Woelfle & Johnson, 2006). Intriguingly, circadian control systems have evolved independently at least 4 different times (Young 2001), suggesting a definite fitness benefit.

In cyanobacteria, circadian cycle timing is partially provided by the interactions of the KaiA, KaiB, and KaiC proteins, which phosphorylate and de-phosphorylate one another in a sustainable 24-hour rhythm (Dong & Golden, 2008). KaiC appears to upregulate SasA and RppA in a pathway that influences DNA topology, possibly making parts of the chromosome more accessible to transcription machinery (Axmann et al., 2009). In *Prochlorococcus*, where the genome has been extensively streamlined (Dufresne et al., 2003), the *kaiA* gene is absent and the clock, based on KaiB and KaiC proteins alone, is less robust than in other cyanobacteria, and requires external cues such as light stimulus to

maintain stable diel patterns (Holtzendorff et al., 2008). The control pathway linking the Kai clock to individual gene expression patterns, and the role of DNA topology, are not yet fully understood.

Diel patterns in cyanobacteria were first elucidated in experiments that focused on individual processes such as photosynthesis (Bruyant et al., 2005), nitrogen fixation (Capone et al., 1990; BermanFrank et al., 2001), carbon and nitrogen metabolism (Mohr et al., 2010), optical properties (Claustre {Claustre et al., 2002), and DNA topology (Pennebaker et al., 2010; Vijayan et al., 2011); or on the expression patterns of small sets of genes, primarily cell-cycle genes and *nifH*, which encodes the iron protein subunit of nitrogenase (ColonLopez et al., 1997; Holtzendorff et al., 2001; Steunou et al., 2008; Wyman et al., 1996). These studies, which consistently reported diel oscillations of important metabolic functions, have been performed on *Crocosphaera watsonii* WH8501 (Mohr et al., 2010; Pennebaker et al., 2010), *Cyanothece* sp. Strain ATCC 51142 (Schneegurt et al., 1994; Colón-López et al., 1997), *Prochlorococcus* sp. PCC 9511 (Holtzendorff et al., 2001; Claustre et al., 2002; Bruyant et al., 2005), *Synechococcus* sp. PCC 7942 and various ecotypes (Steunou et al., 2008; Vijayan et al., 2009), *Trichodesmium thibautii* (Wyman et al., 1996; Capone et al., 2005), and *Trichodesmium erythraeum* IMS101 (Berman-Frank et al., 2001). Recent advances in microarray technology have enabled time-series expression studies of whole or nearly whole genomes, including *Crocosphaera watsonii* WH8501

(Shi et al., 2010), *Cyanothece* sp. Strain ATCC 51142 (Stöckel et al., 2008; Toepel et al., 2008), *Prochlorococcus marinus* MED4 (Zinser et al., 2009; Waldbauer et al., 2012), and *Trichodesmium erythraeum* IMS101 (I. Shilova and J. Zehr, unpub. data). These studies have detected diel cycling (defined as >= 2x change in transcript abundance over 24 hours) in 79%, 30%, 41%, and 63% of analyzed genes respectively.

Many data sets generated by whole-genome time-series expression studies have not been deeply mined or compared, primarily due to a lack of adequate software tools. Development of such tools is hampered by several challenges, including 1) the need to accommodate variations in experimental design, 2) the structure of datasets, and 3) the difficulty of creating effective user interfaces for visualization and analysis. If these challenges can be overcome, mining of individual data sets can help elucidate gene function and predict regulation pathways; comparison of data for multiple organisms can provide insight into relationships between gene expression and adaptation to habitat.

Single-organism whole-genome expression studies are designed with varying duration and starting time of the artificial dark-light cycles, as well as differences in sampling schedules. Thus datasets from different sources are unlikely to be easy to compare. For example, in a *Prochlorococcus* whole-genome study expression study (Zinser et al., 2009), measurements began in the light period and were taken every 2 hours, with a 14-hour light period and 10-hour dark

period. In contrast, a *Crocosphaera* study (Shi et al., 2010) began in the dark period, the dark and light period were both 12 hours, and samples were collected at the start, middle, and the end of the periods. The *Trichodesmium* study (I. Shilova and J. Zehr, unpub. data) began at the end of the dark phase, dark and light phases were both 12 hours, and sampling was every 3 hours. Apparently no current tool can automatically reconcile these incompatible dark/light and sampling aspects of experiment design.

The structure of datasets is also a challenge for a gene expression comparative data analysis tool. Diel gene expression datasets are often characterized by a small number of time points (6-12 per day) and a large number of time series (1 for each of several thousand genes). The small number of time points eliminates the feasibility of using established statistical or mathematical time-series analysis techniques such as Fourier analysis, cosinors, periodograms, Haar wavelets, or Bayesian time series analysis, all of which require a large number of time points per gene to provide enough information to model the expression pattern. Thus analysis software tools based on these approaches are not applicable to typical diel expression datasets, and novel approaches are required.

A third challenge for software development stems from the size and complexity of whole-genome datasets, which make it difficult to detect similarity patterns without software tools that can cluster genes based on expression signature

similarity and facilitate subsequent data exploration. A number of tools for time-expression analysis, including SAM (Tusher et al., 2001), LIMMA (Smyth, 2004.), and BATS (Angelini et al., 2008), have been used in prokaryote expression studies (Spellman et al., 1998; Claridge-Chang et al., 2001; Whitfield et al., 2002; Glynn et al., 2006); however, they are primarily designed for medical applications where the goal is to detect differential expression between two time series for the same gene, for example, to compare healthy versus diseased tissue, and are not readily adaptable to analysis of single or multiple data sets. A number of recent programs provide visualization and graphical user interfaces, but of 16 tools presented in a recent review article ("Studying and modelling dynamic biological processes using time-series gene expression data," Bar-Joseph et al., 2012), 11 support searching or clustering by expression profile similarity, only 2 can present more than 2 data sets at a time, and none provides extensive visual exploration. An opportunity therefore exists to provide tools that will advance diel expression investigation by supporting clustering, visualization, exploration, and simultaneous comparisons of multiple time series data sets.

In this study, a Java application called Dexter ("Diel Expression Terminal") was developed and applied to compare and analyze gene expression time series datasets. Dexter supports visualization, exploration, and clustering of gene expression signatures, integrates the expression data with gene annotations, and

reconciles otherwise incompatible experimental designs. Dexter avoids the use of analysis approaches that require large numbers of timepoints, and is therefore applicable to typical diel expression datasets.

One useful application of Dexter is to improve operon prediction. Operons are groups of consecutive genes, controlled by a single promoter, that are expressed as a single bicistronic or polycistronic transcript. Individual operons have been identified by wetlab techniques in many cyanobacteria, including *Anabaena* (Kuritz et al., 1997), *Crocosphaera* (Pade et al., 2012), *Microcystis* (Mikalsen et al., 2003), *Prochlorococcus* (Vogel et al., 2003; Klein et al., 2009; Osburne et al., 2010), *Synechococcus* (Kutsuna et al., 2005; Shen et al., 2007; Omata,et al., 2001), *Synechocystis* (Vinnemeier et al., 1998; Midorikawa et al., 2012) and *Trichodesmium* (Wang, 2005). Identification of operons can provide important clues regarding inference of regulatory pathways, can support interpretation of operon transcriptome experiments (Moreno-Hagelsieb & Collado-Vides, 2002), and informs computational predictions of *cis*-regulatory elements (Price et al., 2005). The expense of wetlab discovery has prompted the development of algorithms for deducing operons in silico from assembled genomes (Moreno-Hagelsieb & Collado-Vides, 2002; Sabatti et al., 2002; Zheng et al., 2002; Novichkov et al., 2010; Price et al., 2005). Although it is not always possible to predict whole operons, Price (Price et al., 2005) has published an algorithm that predicts whether consecutive genes are members of a common operon. There

are operon predictions from this algorithm for 1336 microbial organisms, including several for which diel expression data sets have been published (http://www.microbesonline.org/operons/OperonList.html). However, few of these predictions have been experimentally verified and we suspected that operon sizes were underestimated. Since genes in an operon are expressed as a unit, they should have very similar or identical diel expression signatures. Therefore a high degree of expression similarity among adjacent genes might predict operon membership and might improve upon prior predictions. We used Dexter to explore expression similarity among genes of pairs of neighboring predicted operons, and identified 35 such pairs in 3 cyanobacterial genomes that we believe are actually single operons.

System and Methods

Requirements

In order to design Dexter, the following six requirements were identified as essential for a single integrated tool targeted at analyzing diel whole- genome gene expression datasets:

- Flexibility with respect to differing experimental designs. Published data set spreadsheets should be imported without modification, and the tool should reconcile experimental design differences with minimal user interaction to facilitate comparison of overall patterns.

- Visualization. Data should be displayed graphically rather than numerically. Graphics should be intuitive and interactive.
- Exploration. The tool should support searching the data space using a number of different gene features, including expression profile, name, pathway, chromosomal proximity, and annotated function.
- Clustering. The tool should be able to cluster any or all of its genes based on similarity of expression profile, using commonly accepted distance metrics.
- Export of results. Interesting data subsets and results of analysis should be readily exportable in standard formats for further analysis or publication.
- Functional extensibility. Since not all desirable features can be anticipated, especially as regards search terms, clustering metrics, and clustering algorithm, the software should be easily extended by users. To this end, software should be open source and should provide polymorphic implementations of features that users might want to extend.

Workflow

Dexter begins in an input wizard mode (Figure 1) that guides the user through the process of describing data set files and importing supplemental information. When all data has been imported, the main Dexter exploration/analysis screen

(Figure 2) presents small thumbnail graphs of expression data that serve as starting points for exploration. During the data exploration phase, the user generates ad-hoc "experiment" graphs (Fig. 2.4, 4^{th} column) that display genes of interest and serve as starting points for analysis.

Input Wizard

The Input Wizard imports raw data spreadsheet files from NCBI, prompts the user to specify spreadsheet structure and experimental setup, determines how to resolve different experimental setups, and imports optional additional information including lists of orthologous genes and operon predictions. (Figure 2.1). In Step A, the user selects one or more dataset spreadsheet files. The first several rows of each spreadsheet file are displayed (A1), and the user selects columns of interest (A1, upper). Each of those columns is then assigned a role (A1, lower), which may be a timepoint, gene i.d., gene name, gene function, or KEGG pathway. In Step B, the user specifies durations of dark and light phases that will constitute a "reference background" on which all expression signatures will be displayed (B1). In Step C the timepoints of each dataset are then mapped to the reference background. In screenshot C1 the reference is at the top of the screen. Timepoints from a spreadsheet ("D11", "L2", etc.) are laid out in the blue strip below the reference. The user drags the mouse to draw lines connecting spreadsheet timepoints to points in the reference.

In Steps D and E (screenshots not shown), the user has the option to import

orthologous genes and/or operon predictions. Orthologous genes may be specified by csv text files or by tabular BLAST result files. Operon predictions are specified by csv text files. After Step E, all data has been imported and the main screen opens to facilitate exploration and analysis.

Main Screen: Exploration and Analysis

The main screen contains scrolling vertical strips of thumbnail graphs (Figure 2). The leftmost strips show the expression profiles of the imported data sets, grouped vertically by pathway or operon. In the figure, three studies have been imported: "Shi_Croco" (Shi et al., 2010), "Shilova_Tery" (I. Shilova and J. Zehr, unpub. data), and "Zinser_Pro" (Zinser et al., 2009); the genes in the study columns can be grouped by order of appearance in the imported spreadsheet, by KEGG pathway, or by operon prediction. The next column ("Experiments") contains graphs of ad-hoc collections that the user assembles using the various Dexter search functions. The remaining columns contain trees and subtrees built with the clustering tool in Dexter.

Any thumbnail can be expanded to a full-screen view. Figure 3 shows a full-screen expansion of the glyoxylate and dicarboxylate metabolism pathway from the *Crocosphaera* study. Expanded views provide a mechanism for identifying genes of interest and collecting them into experiments. The purple, blue, and red profiles in the figure are similar enough to suggest a possible relationship among the genes; this possibility can be explored by checking the genes in the legend

and then clicking “Selected to experiment”.

Figure 4 shows an experiment derived from Figure 3: the three selected genes have been collected and colored red. In an experiment screen, genes can designated as search terms; other genes can be added to the experiment on the basis of expression profile similarity, pathway membership, orthology, or operon membership. In the figure, a search has been performed for genes in other organisms that are orthologous to the three original (red) genes; the two blue genes were found in the *Trichodesmium* data. The dissimilarity between the red and blue profiles suggests that similarity of gene function does not necessarily imply similarity of expression.

Genes in the experiment screen can be selected for display by proximity (Figure 5). The left side of the proximity screen shows a schematic of the genes and their expression profiles, ordered by gene locus. The colors of the vertical separating bars indicate the number of intervening genes. Operon predictions, if available, appear as green bars to the right of the gene names. The figure shows 8 genes from the ATP synthase KEGG pathway. The upper 5 and lower 2 genes are adjacent; there is one intervening gene between atpD and atpC; atpC and atpE are separated by 3633 genes. One predicted operon contains atpI through atpC; a second contains atpE and atpB. A high degree of expression similarity is seen not only within the operons but among all 8 genes of the ATP synthase pathway.

Gene expression patterns in individual thumbnails and/or entire datasets can be clustered using the Neighbor-Joining (Saitou) distance-based tree building algorithm with Euclidean or Pearson Correlation Coefficient distance metrics. Simple tree exploration is supported (Figure 6). Subtrees may be selected and copied into a column in Dexter's main screen (see rightmost column in Figure 2). Hovering the mouse over a node in the tree pops up a thumbnail graph of the gene expressions represented by the node. In the figure, the thumbnail on the left contains 5 apparently similar profiles (black, pink, and purple, all with prominent maxima near the light-dark transition) as well as a large number of profiles with dissimilar expression and lower amplitude. The thumbnail on the right contains just the 5 high-amplitude profiles, which can be copied to an experiment screen for further investigation. For more sophisticated exploration by 3rd-party tools, trees can be exported in Newick format. Newick trees built by 3rd-party tools or exported from previous Dexter sessions can be imported and explored.

Application to Operon Prediction

To demonstrate the power of Dexter as a tool for analyzing expression data, gene expression data from three cyanobacteria diel experiments were supplemented with data on predicted operons to determine whether analysis of gene expression profiles with Dexter can be used to test and possibly improve computationally derived operon predictions that have not been experimentally

verified and may be optimally sensitive.

During the wizard phase, Dexter imported the results of experiments with cultures of *Crocosphaera watsonii* WH 8501 (Shi et al., 2010), *Prochlorococcus marinus* MED4 (Zinser et al., 2009), and *Trichodesmium erythraeum* IMS 101 (I. Shilova and J. Zehr, unpub. data). A reference background was designed, starting with a 2-hour dark phase, followed by 12-hour light and dark phases, and ending with a 2-hour light phase; the light/dark timing designs for the three experiments were mapped to this reference. Operon predictions for the three organisms were obtained from *http://www.microbesonline.org/operons/* and imported into Dexter.

Initial exploration indicated that genes within predicted operons often demonstrate diel expression cycling and have highly similar expression profiles; this was verified by comparing the mean Euclidean expression distance within each predicted operon to the mean expression distance within each organism. Further study showed that genes in neighboring pairs of predicted operons sometimes exhibit highly similar expression patterns, suggesting that the predicted operons are not distinct. To facilitate observing relationships between prior predicted operons and expression profiles, genes in the main screen were grouped into thumbnails by operon prediction, resulting in display of 845 operon thumbnails for *Crocosphaera*, 257 for *Prochlorococcus*, and 835 for *Trichodesmium*. 10 thumbnails with prominent differential expression were

selected for each organism and expanded; in all cases high similarity was visually prominent among all expression profiles of the predicted operon. Each thumbnail was then copied to a new experiment screen, and one gene with typical expression profile was identified and used as a search term for finding the 5 genes with most similar expression. Genes of each experiment were displayed in the proximity screen. The original operon was frequently within 2 genes of another operon, with high expression similarity among all genes of both operons and any intervening genes.

These observations suggested the hypothesis that two neighboring predicted operons should be merged along with any intervening genes when all of the following criteria are met:

1. All genes are on the same strand.

2. There are at most 2 intervening genes.

3. All genes have highly similar expression profiles. That is, Euclidean distances between expression profiles are significantly closer than distances between profiles of pairs of genes that are known not to be in a common operon.

The 3 genomes were scanned computationally to verify that all prior predicted operons met criteria 1 and 3, and to identify pairs of prior predicted operons

that meet all the criteria and are therefore candidates for merging. To estimate the likelihood that recommended merges are correct, a negative training set was collected for each organism, consisting of pairs of adjacent genes known not to be in the same operon by virtue of being on opposite strands. For each negative training set, expression differences between gene pairs were computed and a Gaussian distribution was fitted to the differences. A statistic for estimating the quality of merge recommendations was developed as follows. Let dG_1G_2 be the Euclidean expression distance between any two genes G_1 and G_2. Let $op(G_1G_2)$ signify that genes G_1 and G_2 are in a common operon; conversely let $nop(G_1G_2)$ signify that genes G_1 and G_2 are not in a common operon. Let A, B, C, and D be consecutive genes on the same strand, where the Price algorithm predicts that A and B are in a common operon, and C and D are in a common operon. By analogy to E-values produced by BLAST searches (Altschul et al, 1990), define the E-value for the four genes as the probability of observing a degree of expression similarity at least as close at that among A, B, C, and D if there is no common operon. This is the product, for every pair of genes G_1 and G_2 not predicted to be in the same operon, of $P(d <= dG_1G_2 \mid nop(G_1G_2))$. This term is the cumulative termperature termperature probability, across the Gaussian distribution for the organism's negative training set, of observing two genes whose expression is less than or equal to dG_1G_2.

$E= P(d<dAC|nop(AC)) * P(d<dAD|nop(AD)) * P(d<dBC|nop(BC)) *$

$P(d<dBD|nop(BD))$

The formula can be extended to evaluate recommended merges of genes A, B, C, D, and E, all with observed similarity of diel expression profile, where there is a single intervening gene C between two pairs of predicted operons A, B and D, E. Again, the E-value is the product of probabilities for all pairs of genes for which there is no prior prediction of containment in a common operon:

$E = \Pi\ P(d<dXY \mid nop(XY))$ for $XY\ \varepsilon$ (AC, AD, AE, BC, BD, BE, CD, CE)

and for recommended merges of genes A, B, C, D, E, and F, all with observed similarity of diel expression profile, where there are two intervening genes C and D between two pairs of predicted operon partners A, B and E, F:

$E= \Pi\ P(d<dXY \mid nop(XY))$ for $XY\ \varepsilon$ (AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF)

The E-value formulation was tested on all predicted operons containing at least 4 genes. All such operons were computationally split into two subsets of genes, and an E-value for merging the splits was computed. E-values were then computed for all previously identified candidate pairs. Pairs with E-values less than 5% are reported.

Results and Discussion

Diel gene expression studies are hampered by the lack of software tools to support exploration and analysis of time-series data. Dexter, the novel application presented here, was used to explore time-series data for the cyanobacteria *Crocosphaera watsonii* WH 8501, *Prochlorococcus marinus* MED4, and *Trichodesmium erythraeum* IMS 101. Visually prominent similarity of expression profiles of genes within predicted operons prompted analysis of neighboring predicted operons; additional statistical analysis resulted in recommendations for merging 35 pairs of adjacent predicted operons.

Operon Prediction

48% of *Crocosphaera* genes, 40% of *Prochlorococcus* genes, and 55% of *Trichodesmium* genes were in operons according to the predictions downloaded from *http://www.microbesonline.org/operons/*. Among these genes, diel expression (defined as >= 2x change in transcript abundance over 24 hours) was observed in 48% of *Crocosphaera* genes, 14% of *Prochlorococcus* genes, and 28% of *Trichodesmium* genes; 62% of *Crocosphaera* genes, 27% of *Prochlorococcus* genes, and 50% of *Trichodesmium* genes predicted to be in operons exhibited >= 1.5x change in transcript abundance over 24 hours. These high fractions attest to the importance of operon study and suggest a relationship between operon membership and diel gene expression.

Visualization of expression patterns using Dexter revealed similar expression patterns for genes within operons. Figure 7 shows thumbnail graphs of 10 predicted operons from each of 3 studies. Mean Euclidean expression distance within these operons was 6.2 for *Crocosphaera*, 3.3 for *Prochlorococcus*, and 3.3 for *Trichodesmium*. Mean Euclidean expression distance across all pairs of genes was 19.7 for *Crocosphaera*, 8.7 for *Prochlorococcus*, and 12.3 for *Trichodesmium*. This suggests that expression similarity within operons in general may be significantly higher than between genes outside of operons. Statistical analysis was performed on all operons to determine the validity of this conjecture.

To support statistical analysis, negative training sets of adjacent genes that are not in a common operon were created for each organism under study by collecting pairs of adjacent genes on opposite strands. Euclidean distance between gene expression profiles was computed for each pair of genes, and a Gaussian distribution was fitted to the negative training set for each organism as shown in Table 3.1. The distributions estimate the distribution of expression distance between genes that are adjacent by chance rather than by virtue of being in a common operon.

To determine whether expression similarity is higher between genes in a common operon than among unrelated neighboring genes, mean expression similarity was computed for each prior predicted operon and compared to mean similarity in the negative training sets. Similarity within predicted operons was

greater than mean training set similarity in 98% of predicted operons for *Crocosphaera*, 98% of predicted operons for *Prochlorococcus*, and 79% of predicted operons for *Trichodesmium*. These results support the conjecture of higher expression similarity within operons. The results also support the accuracy (though not the completeness) of the previously untested informatic operon predictions.

To further evaluate expression similarity within predicted operons, and to test the validity of the E-value formulation presented in "System and Methods" above, all predicted operons of length >= 4 were computationally split into two subsets, and the E-value for merging the subsets was calculated. In all cases the E-value was <= 0.003, i.e. the probability of observing the same degree of expression similarity in genes not belonging to a common operon was 0.3%. Statistical analysis thus supported the specificity of the original predictions.

To investigate the completeness of the original predictions, Dexter was used to display expression profiles of genes in neighboring predicted operons. This exploration indicated that when predicted operons were adjacent on the chromosome or separated by at most 2 intervening genes, expression similarity was high among all genes in both the operons and any intervening genes (see for example Figure 10). The visual prominence of this similarity suggested that the operon predictions were incomplete, and that some neighboring pairs of predicted operons should be merged into a single operon. Further evidence for

the incompleteness of operon predictions was obtained by examining operon length. The mean predicted operon length for all 1336 organisms available at http://www.microbesonline.org/operons/ is 3.1 genes, with standard deviation of 2.1, whereas the mean predicted length for the three organisms analyzed here was μ=2.4, σ=.88 for *Crocosphaera,* μ=2.7, σ=1.4 for *Prochlorococcus*, and μ=2.6, σ=1.5 for *Trichodesmium*. These lengths are significantly shorter than the overall mean operon length, suggesting that some operon predictions are incomplete and should be extended to include adjacent genes.

The possibility that actual operons extend beyond predicted operons, implies that some pairs of neighboring predicted operons may in fact belong to the same operon and the neighboring predictions should be merged. To identify such pairs, all predicted operons were evaluated computationally. Neighboring pairs of predicted operons on the same strand, separated by at most 2 intervening genes, were identified as possible candidates for merging. (The upper bound of 2 intervening genes was initially selected arbitrarily to simplify computation; later analysis showed that there are no neighboring pairs of predictions on the same strand that are separated by 3 or more intervening genes.) No neighboring operon pairs on the same strand were separated by more than 2 intervening genes. 124 candidate pairs were identified for *Crocosphaera*, 36 were identified for *Prochlorococcus*, and 144 were identified for *Trichodesmium* (Table 2). The E-value for merging each pair was computed.

Table 3 lists 35 candidate pairs of predicted operons (9 in *Crocosphaera*, 10 in *Prochlorococcus*, and 16 in *Trichodesmium*) that meet the criteria presented in "System and Methods" above: both operons in each pair are on the same strand, operons are separated by at most 2 intervening genes, and expression similarity among all genes is high as determined by E-value <= 5%. We propose that for each of these candidate pairs, the genes of both predicted operons together with any intervening genes are actually in a single operon.

Further investigation into expression similarity within operons may eventually provide support for genome annotation. Since coregulation implies a strong functional connection, information about the function of unknown genes in an operon might be deduced from the function of known genes in the same operon, provided at least one gene in the operon is annotated. A large fraction of any genome lacks functional assignment; indeed, most genes in the recommended merged operons presented here are not annotated. In the set of prior predicted operons that we recommend for merging, 3 *Crocosphaera* genes and 7 *Trichodesmium* genes of unknown function are in operons containing at least one annotated gene; when these operons are merged as recommended, 17

Crocosphaera genes and 20 *Trichodesmium* genes of unknown function are in operons containing at least one annotated gene. Thus merging operons increases the number of unknown genes whose function might be deduced from the function of other genes in the same operon.

Our operon predictions are recommendations to merge adjacent prior predictions and intervening genes; at present no prediction is made regarding extension before the start of the first operon or after the end of the second operon. The technique presented here could be refined to predict entire operons. For example, suppose genes X, A, B, C, D, E, and Y are consecutive genes on the same strand, and our technique recommends merging predicted operons AB and DE along with intervening gene C. If the expression signatures of X and Y are substantially different from the signatures of the other genes, then ABCDE could be a complete operon. This hypothesis could be supported by estimating the probability of observing expression distances greater than or equal to dXA and dEY, conditioned on X and Y actually being operon mates with ABCDE. The probability estimate would be computed from the distribution of expression distances among pairs of known operon-mate genes.

The analysis presented here focused on individual organisms; the results support computational operon predictions that have not been experimentally verified, and can potentially provide insight into regulatory pathways and improve prediction of *cis*-regulatory elements. Since some operons are known to

be conserved among organisms (for example *kaiABC* among non-minimal cyanobacteria, or *nifHDK* among nitrogen fixers), a further application of Dexter to operon study could be the exploration and comparison of expression signatures of conserved operons. Such a study could provide insight into how differences in operon expression confer adaptive benefit in different habitats.

The observations that led to the operon prediction analysis were made possible by a number of features of Dexter, including integration of operon predictions, the ability to visualize the positions of selected genes on their chromosomes, clustering by expression profile similarity, the ability to explore data by using an expression profile as a search term, and the ability to simultaneously display time course expression graphs for the genes of all 3 data sets. Individually these features may not be unique to Dexter, but we know of no other software tool that integrates all of them. Of the 16 applications described in a recent review article ("Studying and modelling dynamic biological processes using time-series gene expression data," Bar-Joseph et al., 2012; see also Table 4), many are designed for specific analyses rather than for general exploration. None accesses operon predictions or presents genes in the context of chromosome position, Only 5 (BETR, LIMMA, RESTS, SAM, and PESTS) can display more than one data set at a time, and of these only two (RESTS and SAM) can display more than two data sets. BETR has no graphics. Biggests cannot manage multiple data sets and supports only limited exploration. CAGED provides graphical display of heat

maps but not time course data. DynaMiteC has no graphics. EDGE provides graphical display of box plots and eigengenes but not time course data. GATE's graphics are limited to hexagonal heat maps, and exploration is not supported. GQL does not support exploration. LIMMA does not provide graphics. MVQueries manages single data sets and is not graphical; it is specifically intended for classification of response to toxins. PESTS provides limited graphics and exploration; it is specifically designed for detection of genes that show differential expression between two phenotypes (e.g. healthy and diseased) of the same organism. PRIISM is not graphical and is specifically designed for detection of stress response. REST provides graphics but not per-gene time course expression graphs; its purpose is to provide statistical analysis of genes that are differentially expressed under different stimuli. SAM provides limited graphics and is only intended for the identification of genes that are significantly related to a response variable. STEM provides time course graphics but horizontal scales are not linear, clustering is based on similarity to a predefined set of model profiles, and exploration is not supported. TimeClust's graphics include time course graphs and self-organizing maps but exploration is not supported; several clustering algorithms are provided. TRAM is designed for the classification of diseased tissues. The review article also discusses 5 additional tools: Inferelator and TSNI, which are no longer available; Network Component Analysis and DREM, which infer dynamic regulatory networks, and GeneNetWeaver, which generates data sets for benchmarking other tools. To

summarize, PESTS, which supports comparison of the response of an organism under different treatments, may come closest to Dexter in terms of functionality; however no available tool other than Dexter supports both graphical exploration and simultaneous analysis of

Further applications for Dexter

Little is understood about the mechanisms that link circadian clocks to the cyclic expression patterns of individual genes, or about the endogenous or exogenous causes of variety among expression patterns. In cyanobacteria, the adaptive benefit of observed patterns is only clear in the case of a few genes, notably those involved in photosynthesis and, for some diazotrophs, nitrogen fixation. The expression patterns of the great majority of cyanobacterial genes may confer benefit, or may simply be neutral consequences of the circadian control systems that optimize photosynthesis and sequester nitrogen fixation. Time series studies could shed light on these issues, but mining time series data sets, and comparison of expression patterns across data sets, are hampered by the lack of software tools that can reconcile different experiment designs, cluster expression patterns, and support visualization and exploration. Traditional time-series analysis techniques, which lack statistical strength without hundreds or thousands of time points, cannot be applied. Thus, there exists an opportunity to facilitate a greater understanding of regulation of gene expression through the

development of novel software tools,. There is a particular need for tools that enable visualization and graphical exploration, so that (as with the operon prediction analysis presented here) visually prominent relationships can support new hypotheses.

Dexter, the application presented here, has the potential to elucidate a number of issues concerning cell metabolism. Detection of expression similarity patterns led to the operon prediction work described here; expression similarity might also be used to characterize non-operon pathways. With incorporation of non-expression data such as DNA topology, it might be possible to investigate the control mechanisms that link a single circadian clock to the observed variety among individual expression patterns.

Comparison of expression patterns of genes or pathways among multiple organisms might elucidate relationships between expression and habitat adaptation. For example, the *Prochlorococcus* strain studied here, MED4, is adapted to high light conditions. At present there are no published diel expression studies of low-light *Prochlorococcus* strains such as MIT 9313; if a such a study were conducted, its data could easily be imported into Dexter and compared to the MED4 data, notwithstanding any differences in experiment design. The most prominent expression pattern differences between orthologous genes could then be identified, and possibly correlated to habitat adaptation.

The most promising application of Dexter could be the analysis of environmental data sets, which can be assessed using microarrays or by RNA-Seq technology (Wang et al., 2009). Microarray studies have greatly benefitted from the GeoChip array (He et al., 2007; He et al., 2010), and more recently from the MicroTOOLs array (Shilova et al., 2014), which is specifically designed for pelagic marine environments. To date, neither array has been used in a diel expression study. However, the initial proof-of-concept MicroTOOLs application included phosphorus and iron amendment experiments; although Dexter was designed for time series analysis, it could also be used to detect patterns in amendment series. Communities can also be assessed by RNA sequencing (for example Ottesen, E. et al., 2013). A recent study using this approach (Ottesen, E. et al., 2014) detected intriguing coordination of expression patterns between primary producers and heterotrophs. Deeper data exploration using Dexter could help to characterize these patterns. Similar expression coordination might exist between bacterial symbionts and their hosts. For example, recent work has revealed that *Candidatus* Atelocyanobacterium thalassa (UCYN-A), a diazotroph of possibly global ecological importance, is a symbiont that lacks many common metabolic pathways (Tripp et al., 2010); different strains have differing genomes that may reflect adaptation to habitat (Bombar et al., 2014). If expression time series data were available for both UCYN-A and its host, Dexter could be used to study the coordinated dynamics of the relationship.

Gene expression time series studies are generally characterized by a small number of time points across a large number of genes. Technological improvements are driving down the cost per probe of microarrays and the cost per RNA base of RNA-Seq experiments; however, the cost per time point remains high. Thus, for the foreseeable future data sets will be well suited for analysis by Dexter, which facilitates exploration of large numbers of expression profiles while providing analysis that does not require the statistical strength provided by large numbers of time points. However, if future technological developments enable collection of data sets with many time points, Dexter will still be applicable and useful, and most of its functions will perform without appreciable loss of speed.

Acknowledgements

The authors are grateful to Rex Malmstrom, Laurence Nedelec, Irina Shilova, and Josh Stuart for valuable discussions.

Funding: This work was supported by a first-phase Gordon and Betty Moore Marine Investigator grant (J.Z.) and the Microbial Environmental Genomics Applications: Modeling, Experimentation, and Remote Sensing (MEGAMER) facility of the University of California, Santa Cruz.

Conflict of Interest: None declared.

References

Altschul S, et al. (1990). Basic Local Alignment Search Tool. *J Mol Biol* 215, 403-410.

Angelini, C., Cutillo, L., De Canditiis, D., Mutarelli, M., & Pensky, M. (2008). BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, *9*(1), 415. doi:10.1186/1471-2105-9-415

Axmann, I. M., Dühring, U., Seeliger, L., Arnold, A., Vanselow, J. T., Kramer, A., & Wilde, A. (2009). Biochemical evidence for a timing mechanism in *prochlorococcus*. *Journal of Bacteriology*, *191*(17), 5342–5347. doi:10.1128/JB.00419-09

Bar-Joseph, Z., Gitter, A., & Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nature* 2012 vol. 13:552-564.

Bergersen, F. J. (1962). The Effects of Partial Pressure of Oxygen upon Respiration and Nitrogen Fixation by Soybean Root Nodules. *Microbiology*, *29*(1), 113–125. doi:10.1099/00221287-29-1-113

Berman-Frank, I., Lundgren, P., Chen, Y. B., Küpper, H., Kolber, Z., Bergman, B., & Falkowski, P. (2001). Segregation of nitrogen fixation and oxygenic photosynthesis in the marine cyanobacterium Trichodesmium. *Science*, *294*(5546), 1534–1537. doi:10.1126/science.1064082

Bombar, D., Heller, P., Sánchez-Baracaldo, P., Carter, B. J., & Zehr, J. P. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *The ISME Journal*, 1–13. doi:10.1038/ismej.2014.167

Bond,G.1961.The oxygen relations of nitrogen fixation in root nodules. *Z. Allg. Mikrobiol*. **1**:93-99

Bruyant, F., Babin, M., Genty, B., Prasil, O., & Behrenfeld, M. J. (2005). Diel variations in the photosynthetic parameters of Prochlorococcus strain PCC 9511: Combined effects of light and cell cycle. *Limnol Oceanogr*. 2005, **50** (3) 850-863

Capone, D.G., O'Neil, J.M., Zehr, J., & Carpenter, E.J. (1990). Basis for Diel Variation in the Marine Planktonic Cyanobacterium *Trichodesmium* thiebautii. *Applied and Environmental Microbiology*, Nov 1990, p. 3532-3536.

Capone, D. G., Burns, J. A., Montoya, J. P., Subramaniam, A., Mahaffey, C., Gunderson, T., et al. (2005). Nitrogen fixation by Trichodesmium spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochemical Cycles*, *19*(2), n/a–n/a. doi:10.1029/2004GB002331

Claridge-Chang, A., Wijnen, H., Naef, F., Boothroyd, C., Rajewsky, N., & Young, M. W. (2001). Circadian regulation of gene expression systems in the Drosophila head. *Neuron*, *32*(4), 657–671.

Claustre, H., Bricaud, A., Babin, M., Bruyant, F., Guillou, L., Le Gall, F., et al. (2002). Diel variations in Prochlorococcus optical properties. *Limnology and Oceanography*, *47*(6), 1637–1647. doi:10.4319/lo.2002.47.6.1637

Colón-López, M. S., Sherman, D. M., & Sherman, L. A. (1997). Transcriptional and translational regulation of nitrogenase in light-dark- and continuous-light-grown cultures of the unicellular cyanobacterium Cyanothece sp. strain ATCC 51142. *Journal of Bacteriology*, *179*(13), 4319–4327.

Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., et al. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science*, *309*(5734), 630–633. doi:10.1126/science.1115581

Dong, G. & Golden, S. (2008). How a cyanobacterium tells time. *Current Opinion in Microbiology.* **11**: 541-546.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., et al. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences*, *100*(17), 10020–10025. doi:10.1073/pnas.1733211100

Fay, P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiological Reviews*, *56*(2), 340–373.

Glynn, E. F., Chen, J., & Mushegian, A. R. (2006). Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics*, *22*(3), 310–316. doi:10.1093/bioinformatics/bti789

He, Z., Deng, Y., Van Nostrand, J. D., Tu, Q., Xu, M., Hemme, C. L., et al. (2010). GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *The ISME Journal*, *4*(9), 1167–1179. doi:10.1038/ismej.2010.46

He, Z., Gentry, T. J., Schadt, C. W., Wu, L., Liebich, J., Chong, S. C., et al. (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal*, *1*(1), 67–77. doi:10.1038/ismej.2007.2

Holtzendorff, J., Partensky, F., Jacquet, S., Bruyant, F., Marie, D., Garczarek, L., et al. (2001). Diel expression of cell cycle-related genes in synchronized cultures of Prochlorococcus sp. strain PCC 9511. *Journal of Bacteriology*, *183*(3), 915–920. doi:10.1128/JB.183.3.915-920.2001

Holtzendorff, J., Partensky, F., Mella, D., Lennon, J.-F., Hess, W. R., & Garczarek, L. (2008). Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. *Journal of Biological Rhythms*, *23*(3), 187–199. doi:10.1177/0748730408316040

Houmard, J., Capuano, V., Cousin, T., and Tandeau de Marsac, N. (1988) Genes encoding core components of the phycobilisome in the cyanobacterium Calothrix sp. strain PCC 7601: occurrence of a multigene family. *J.Bacteriol*. 170, 5512–5521

Klein, M. G., Zwart, P., Bagby, S. C., Cai, F., Chisholm, S. W., Heinhorst, S., et al. (2009). Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport. *Journal of Molecular Biology*, *392*(2), 319–333. doi:10.1016/j.jmb.2009.03.056

Kuritz, T., Bocanera, L. V., & Rivera, N. S. (1997). Dechlorination of lindane by the cyanobacterium Anabaena sp. strain PCC7120 depends on the function of the nir operon. *Journal of Bacteriology*, *179*(10), 3368–3370.

Kutsuna, S., Nakahira, Y., Katayama, M., Ishiura, M., & Kondo, T. (2005). Transcriptional regulation of the circadian clock operon kaiBCby upstream regions in cyanobacteria. *Molecular Microbiology*, *57*(5), 1474–1484. doi:10.1111/j.1365-2958.2005.04781.x

P. W. Ludden, G. P. Roberts, in Anoxygenic Photosynthetic Bacteria, R. E. Blankenship, M. T. Madigan, C. E. Bauer, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1995), pp. 929–947.

Midorikawa, T., Narikawa, R., & Ikeuchi, M. (2012). A deletion mutation in the spacing within the psaA core promoter enhances transcription in a cyanobacterium Synechocystis sp. PCC 6803. *Plant & Cell Physiology*, *53*(1), 164–172. doi:10.1093/pcp/pcr159

Mikalsen, B., Boison, G., Skulberg, O. M., Fastner, J., Davies, W., Gabrielsen, T. M.,

et al. (2003). Natural variation in the microcystin synthetase operon mcyABC and impact on microcystin production in Microcystis strains. *Journal of Bacteriology*, *185*(9), 2774–2785. doi:10.1128/JB.185.9.2774-2785.2003

Mohr, W., Intermaggio, M. P., & LaRoche, J. (2010). Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium Crocosphaera watsoniiWH8501. *Environmental Microbiology*, *12*(2), 412–421. doi:10.1111/j.1462-2920.2009.02078.x

Moreno-Hagelsieb, G., & Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, *18 Suppl 1*, S329–36

Muramatsu, M. and Hihara, Y. (2006) Characterization of high-light-responsive promoters of the *psaAB* genes in *Synechocystis* sp. PCC 6803. *Plant Cell Physiol*. 47: 878–890.

Novichkov, P. S., Rodionov, D. A., Stavrovskaya, E. D., Novichkova, E. S., Kazakov, A.E., Gelfand, M. S., et al. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Research*, *38*(Web Server issue), W299–307. doi:10.1093/nar/gkq531

Omata, T., Gohta, S., Takahashi, Y., Harano, Y., & Maeda, S. (2001). Involvement of a CbbR homolog in low CO2-induced activation of the bicarbonate transporter operon in cyanobacteria. *Journal of Bacteriology*, *183*(6), 1891–1898. doi:10.1128/JB.183.6.1891-1898.2001

Osburne, M. S., Holmbeck, B. M., Frias-Lopez, J., Steen, R., Huang, K., Kelly, L., et al. (2010). UV hyper-resistance in *Prochlorococcus* MED4 results from a single base pair deletion just upstream of an operon encoding nudix hydrolase and photolyase. *Environmental Microbiology*, *12*(7), 1978–1988. doi:10.1111/j.1462-2920.2010.02203.x

Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., & Delong, E. F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), E488–97. doi:10.1073/pnas.1222099110

Pade, N., Compaoré, J., Klähn, S., Stal, L. J., & Hagemann, M. (2012). The marine cyanobacterium Crocosphaera watsonii WH8501 synthesizes the compatible solute trehalose by a laterally acquired OtsAB fusion protein. *Environmental Microbiology*, *14*(5), 1261–1271. doi:10.1111/j.1462-2920.2012.02709.x

Panda, S., Antoch, M. P., Miller, B. H., Su, A. I., Schook, A. B., Straume, M., et al.

(2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, *109*(3), 307–320.

Parker, C. A., & Scutt, P. B. (1960). The effect of oxygen on nitrogen fixation by Azotobacter. *Biochimica Et Biophysica Acta*, *38*, 230–238.

Pennebaker, K., Mackey, K. R. M., Smith, R. M., Williams, S. B., & Zehr, J. P. (2010). Diel cycling of DNA staining and nifH gene regulation in the unicellular cyanobacterium *Crocosphaera watsonii* strain WH 8501 (Cyanophyta). *Environmental Microbiology*, *12*(4), 1001–1010. doi:10.1111/j.1462-2920.2010.02144.x

Price, M. N., Huang, K. H., Alm, E. J., & Arkin, A. P. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research*, *33*(3), 880–892. doi:10.1093/nar/gki232

Sabatti, C., Rohlin, L., Oh, M.-K., & Liao, J. C. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research*, *30*(13), 2886–2893.

Schneegurt, M. A., Sherman, D. M., Nayar, S., & Sherman, L. A. (1994). Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the cyanobacterium *Cyanothece* sp. strain ATCC 51142. *Journal of Bacteriology*, *176*(6), 1586–1597.

Shen, G., Balasubramanian, R., Wang, T., Wu, Y., Hoffart, L. M., Krebs, C., et al. (2007). SufR coordinates two [4Fe-4S]2+, 1+ clusters and functions as a transcriptional repressor of the *sufBCDS* operon and an autoregulator of *sufR* in cyanobacteria. *Journal of Biological Chemistry*, *282*(44), 31909–31919. doi:10.1074/jbc.M705554200

Shi, T., Ilikchyan, I., Rabouille, S., & Zehr, J. P. (2010). Genome-wide analysis of diel gene expression in the unicellular N2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *The ISME Journal*, 1–12. doi:10.1038/ismej.2009.148

Shilova, I. N., Robidart, J. C., Tripp, H. J., Turk-Kubo, K., Wawrik, B., Post, A. F., et al. (2014). A microarray for assessing transcription from pelagic marine microbial taxa, *8*(7), 1476–1491. doi:10.1038/ismej.2014.1

Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1). doi:10.2202/1544-6115.1027

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., et al.

(1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, *9*(12), 3273–3297.

Steunou, A.-S., Jensen, S. I., Brecht, E., Becraft, E. D., Bateson, M. M., Kilian, O., et al. (2008). Regulation of nif gene expression and the energetics of N2 fixation over the diel cycle in a hot spring microbial mat. *The ISME Journal*, *2*(4), 364–378. doi:10.1038/ismej.2007.117

Stöckel, J., Welsh, E. A., Liberton, M., Kunnvakkam, R., Aurora, R., & Pakrasi, H. B. (2008). Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 6156–6161. doi:10.1073/pnas.0711068105

Studying and modelling dynamic biological processes using time-series gene expression data. (2012). Studying and modelling dynamic biological processes using time-series gene expression data, *13*(8), 552–564. doi:10.1038/nrg3244

Toepel, J., Welsh, E., Summerfield, T. C., Pakrasi, H. B., & Sherman, L. A. (2008). Differential transcriptional analysis of the cyanobacterium Cyanothece sp. strain ATCC 51142 during light-dark and continuous-light growth. *Journal of Bacteriology*, *190*(11), 3904–3913. doi:10.1128/JB.00206-08

Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., et al. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, *464*(7285), 90–94. doi:10.1038/nature08786

Tuit, C., Waterbury, J., & Ravizza, G. (2004). Diel variation of molybdenum and iron in marine diazotrophic cyanobacteria. *Limnology and Oceanography*.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116–5121. doi:10.1073/pnas.091062498

Vijayan, V., Zuzow, R., & O'Shea, E. K. (2009). Oscillations in supercoiling drive circadian gene expression in cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(52), 22564–22568. doi:10.1073/pnas.0912673106

Vinnemeier, J., Kunert, A., & Hagemann, M. (1998). Transcriptional analysis of the *isiAB* operon in salt-stressed cells of the cyanobacterium *Synechocystis* sp. PCC 6803. *FEMS Microbiology Letters*, *169*(2), 323–330.

Vogel, J., Axmann, I., Herzel, H., & Hess, W. Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4 (2003). *Nucleic Acids Research*, 2003, vol. 31, No. 11.

Waldbauer JR, Rodrigue S, Coleman ML, Chisholm SW (2012) Transcriptome and Proteome Dynamics of a Light-Dark Synchronized Bacterial Cell Cycle. *PLoS ONE 7*(8): e43432. doi:10.1371/journal.pone.0043432

Wang, J. (2005). Recent Cyanobacterial Kai Protein Structures Suggest a Rotary Clock. *Structure*, *13*(5), 735–741. doi:10.1016/j.str.2005.02.011

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, *13*(6), 1977–2000. doi:10.1091/mbc.02-02-0030

Woelfle, M. A., & Johnson, C. H. (2006). No promoter left behind: global circadian gene expression in cyanobacteria. *Journal of Biological Rhythms*, *21*(6), 419–431. doi:10.1177/0748730406294418

Wyman, M., Zehr, J. P., & Capone, D. G. (1996). Temporal Variability in Nitrogenase Gene Expression in Natural Populations of the Marine Cyanobacterium *Trichodesmium thiebautii*. *Applied and Environmental Microbiology,*, *62*(3), 1073–1075.

Yao, D. C. I., Brune, D. C., Vavilin, D., & Vermaas, W. F. J. (2012). Photosystem II component lifetimes in the cyanobacterium *Synechocystis* sp. strain PCC 6803: small Cab-like proteins stabilize biosynthesis intermediates and affect early steps in chlorophyll synthesis. *The Journal of Biological Chemistry*, *287*(1), 682–692. doi:10.1074/jbc.M111.320994

Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R. & Kasif, S. (2012). *Genome Research*.**12**: 1221-1230.

Zhong Wang, Mark Gerstein & Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (January 2009) | doi:10.1038/nrg2484

Zinser, E. R., Lindell, D., Johnson, Z. I., Futschik, M. E., Steglich, C., Coleman, M. L., et al. (2009). Choreography of the Transcriptome, Photophysiology, and Cell Cycle of a Minimal Photoautotroph, *Prochlorococcus*. *PLoS ONE*, *4*(4), e5135. doi:10.1371/journal.pone.0005135

Figures and Tables

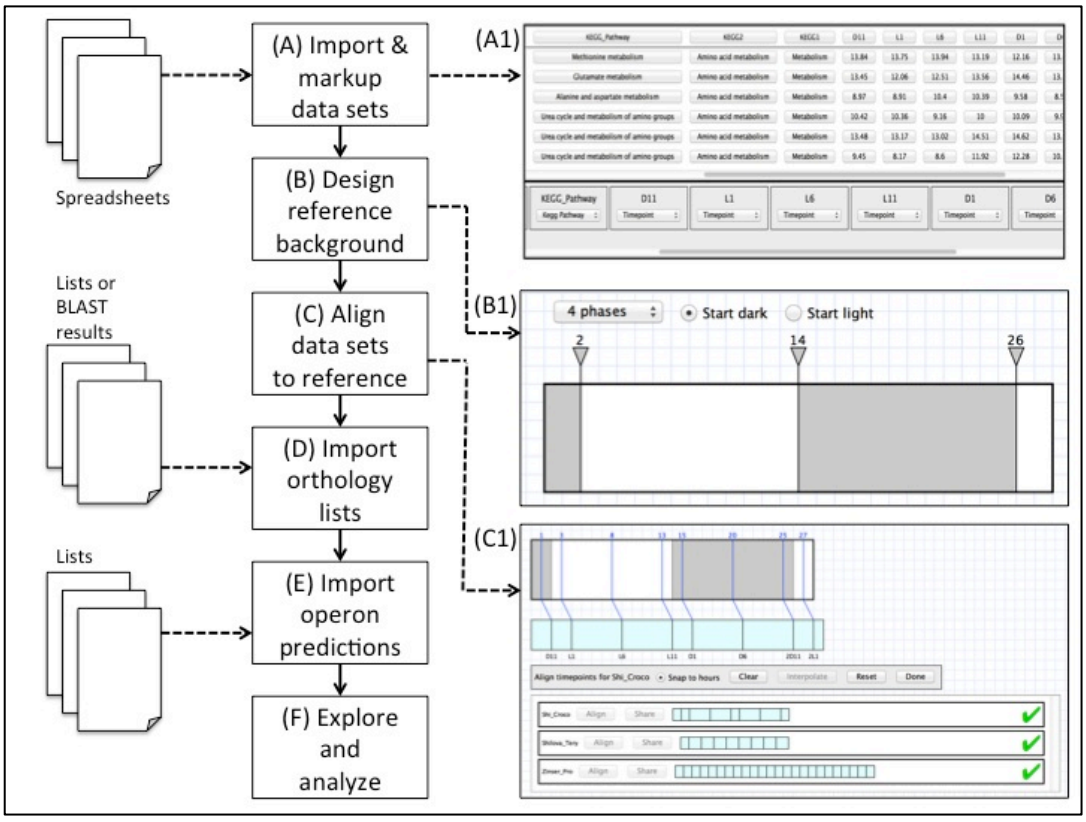


Figure 1. Input Wizard. The wizard guides the user through 5 steps prior to exploration and analysis: (A) specifying dataset spreadsheet structure; (B) designing a "reference background" onto which experiment timepoints will be mapped (C); and optionally importing lists of orthologous genes (D) and operon predictions (E).

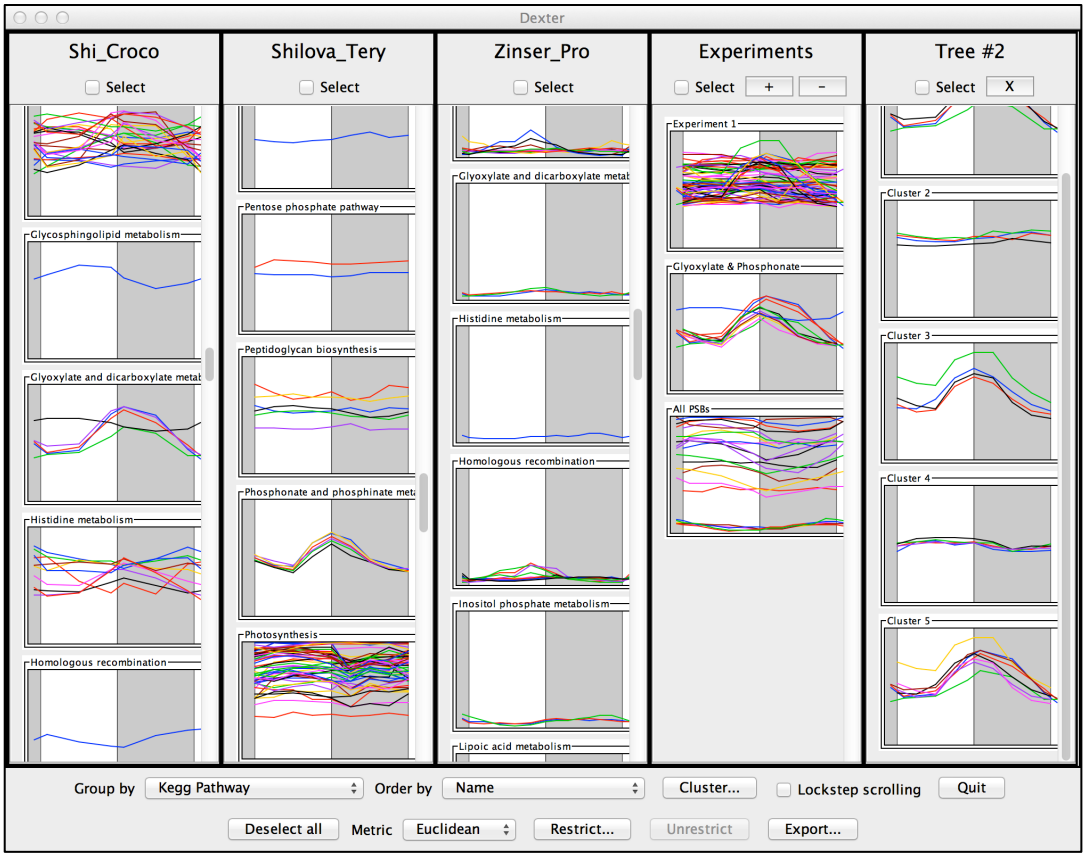


Figure 2. Main screen. The leftmost strips are the original data sets, grouped by KEGG pathway. The fourth strip contains ad-hoc experiments. The rightmost strip contains subtrees from a clustering operation.

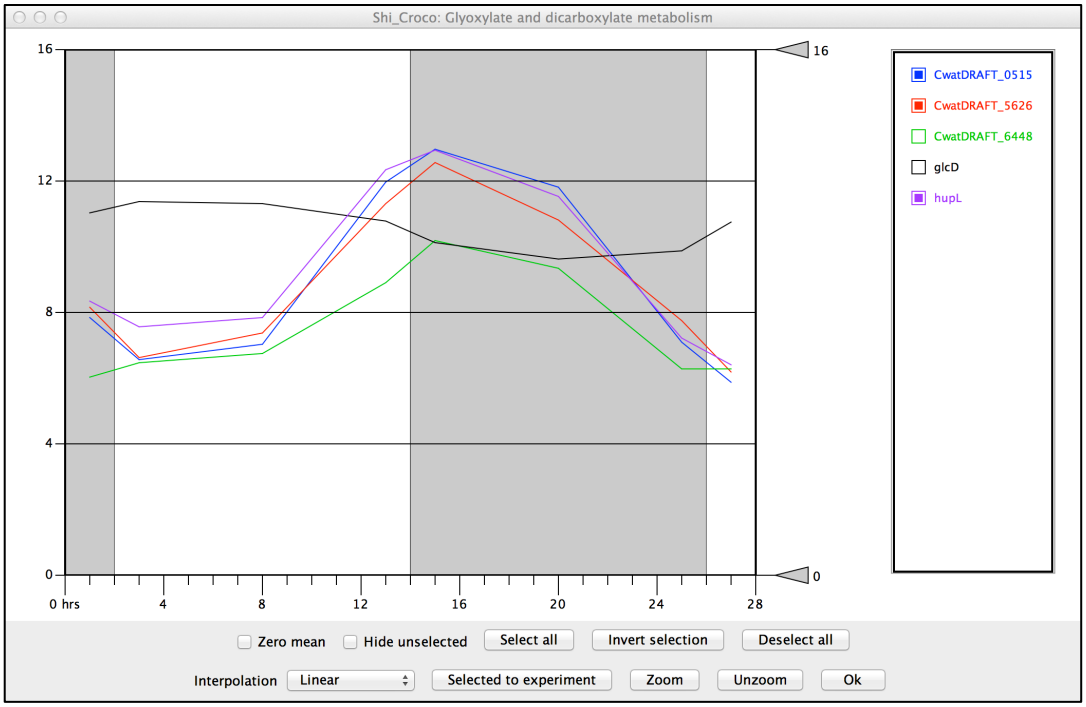


Figure 3. Expanded View of glyoxylate and dicarboxylate pathway from the *Crocosphaera* data set.

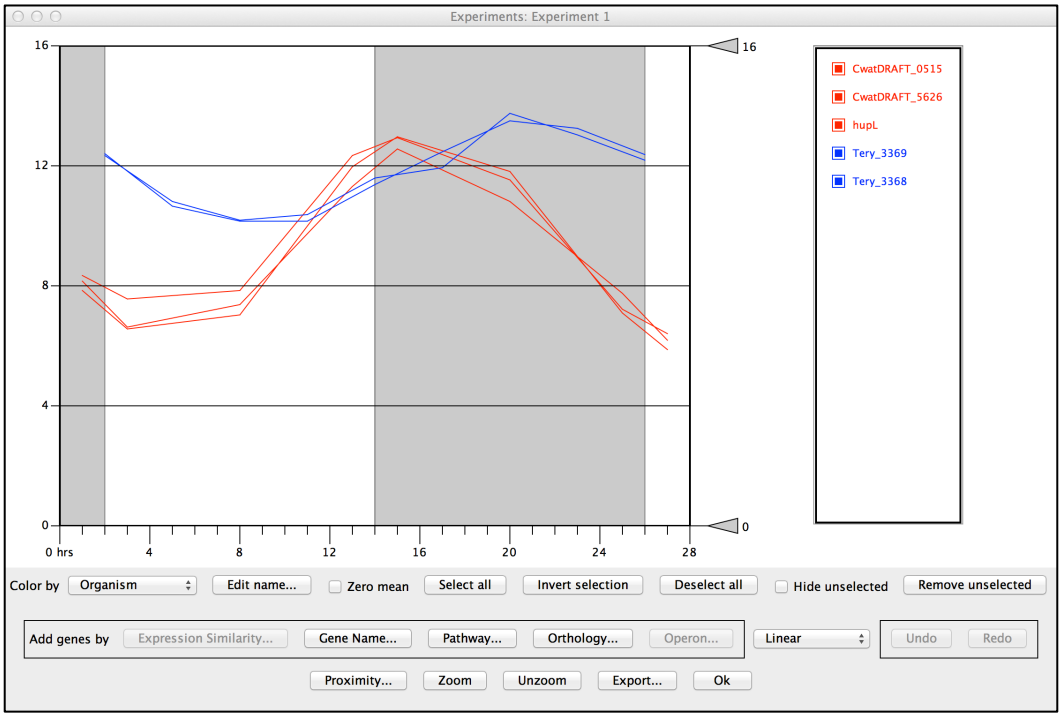


Figure 4. Expanded experiment view. The 3 red genes have been selected from the screen in Figure 2.3. The blue genes have been selected by searching for orthologs of the red genes.

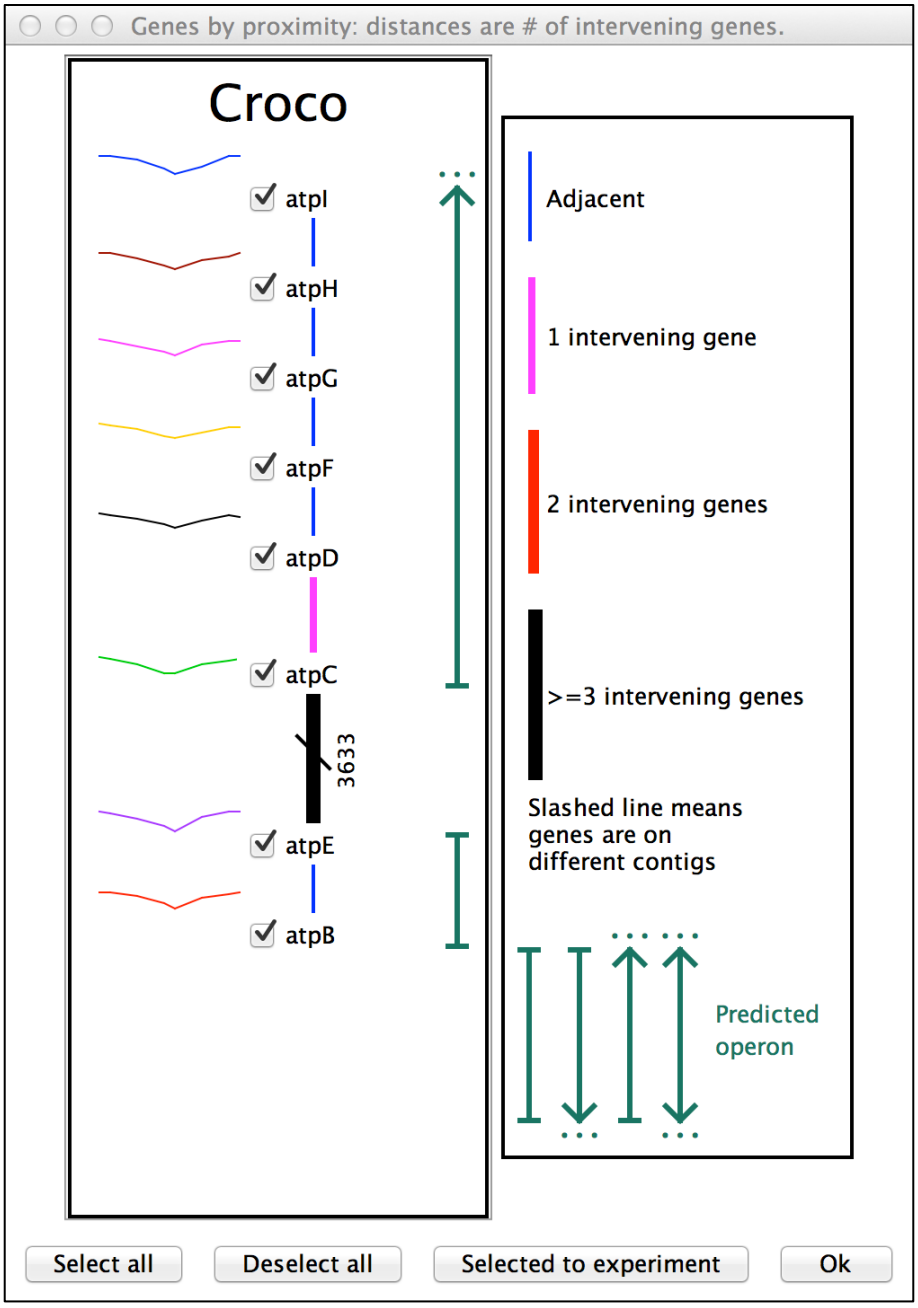


Figure 5. Proximity screen. Genes previously selected in an experiment screen are shown with their expression profiles, number of intervening genes between neighbors, and operon predictions.

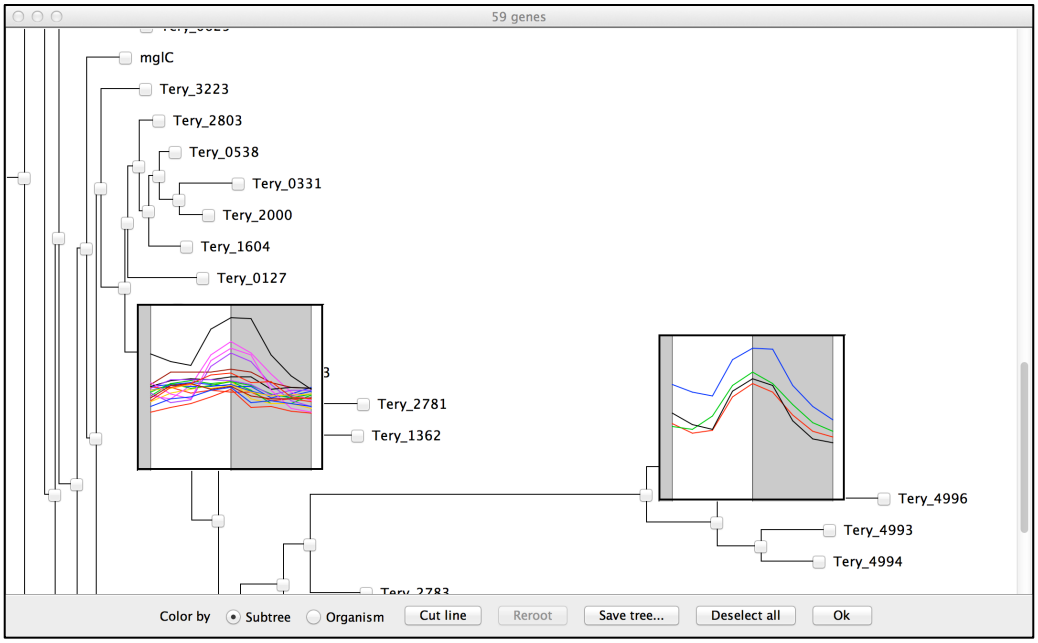


Figure 6. Tree exploration screen. Mousing over a node displays a popup graph of the associated subtree.

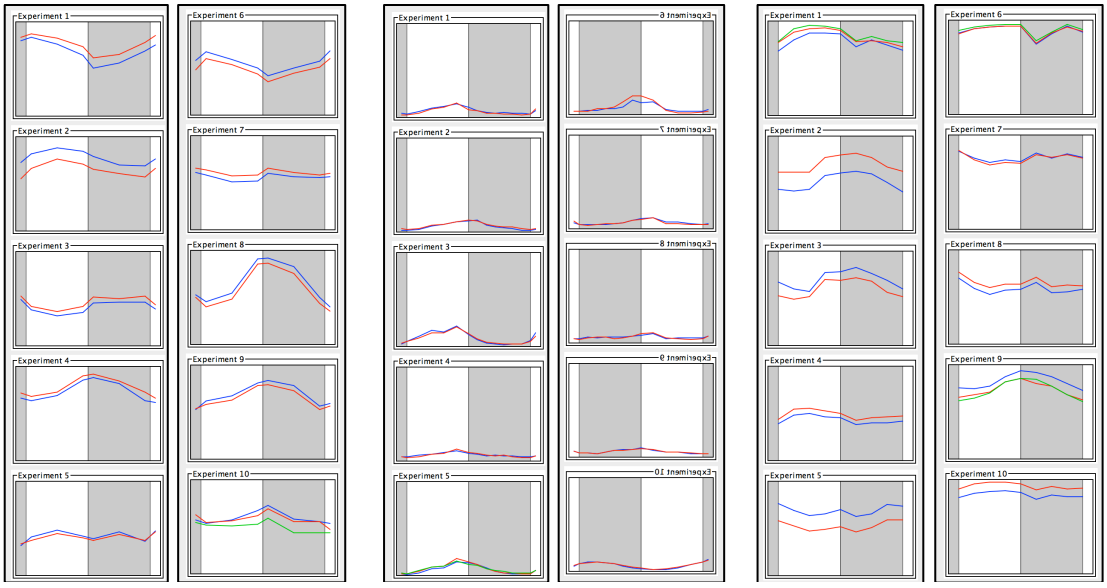


Figure 7. 10 Sample predicted operons from each organism. Left: *Crocosphaera*. Center: *Prochlorococcus*. Right: *Trichodesmium*.

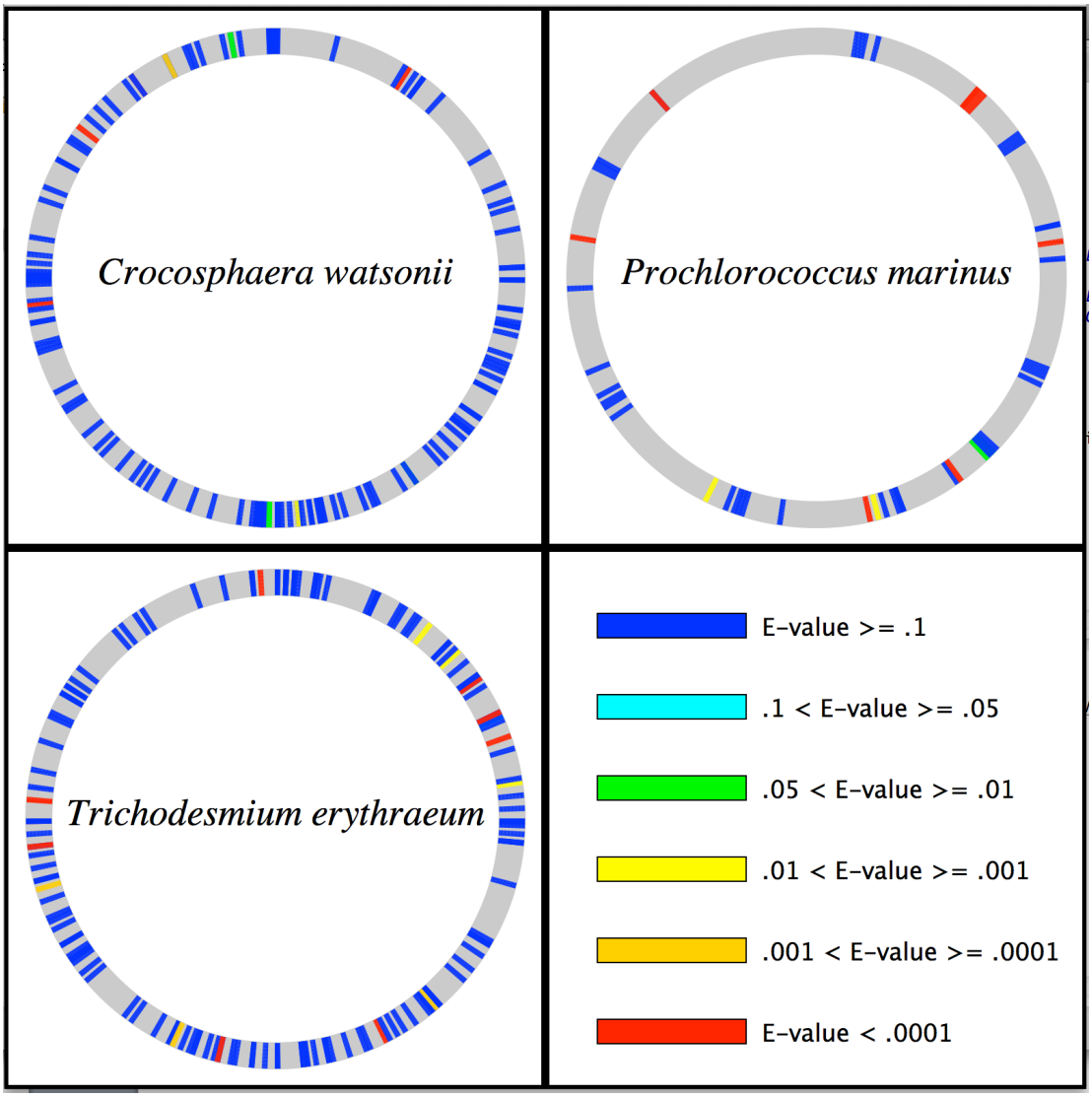


Figure 8. Relative positions on chromosome of merge candidates, colored by E-value. Hotter colors represent lower E-values. Gene extent has been exaggerated for visibility.

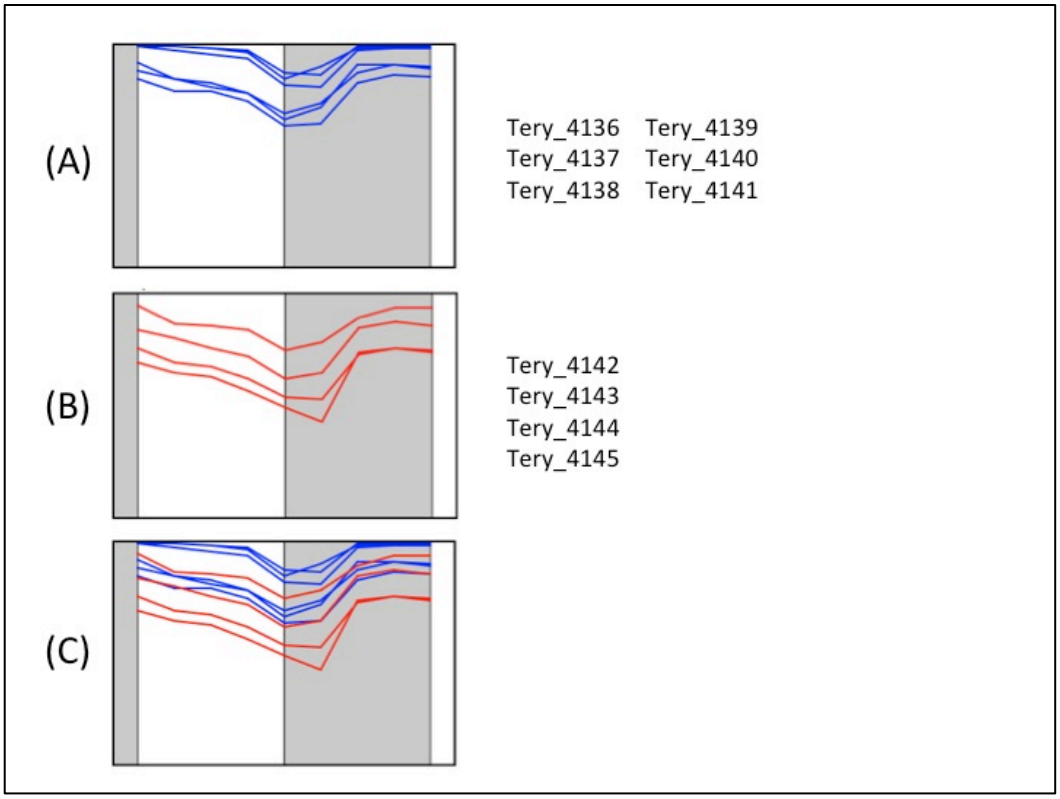
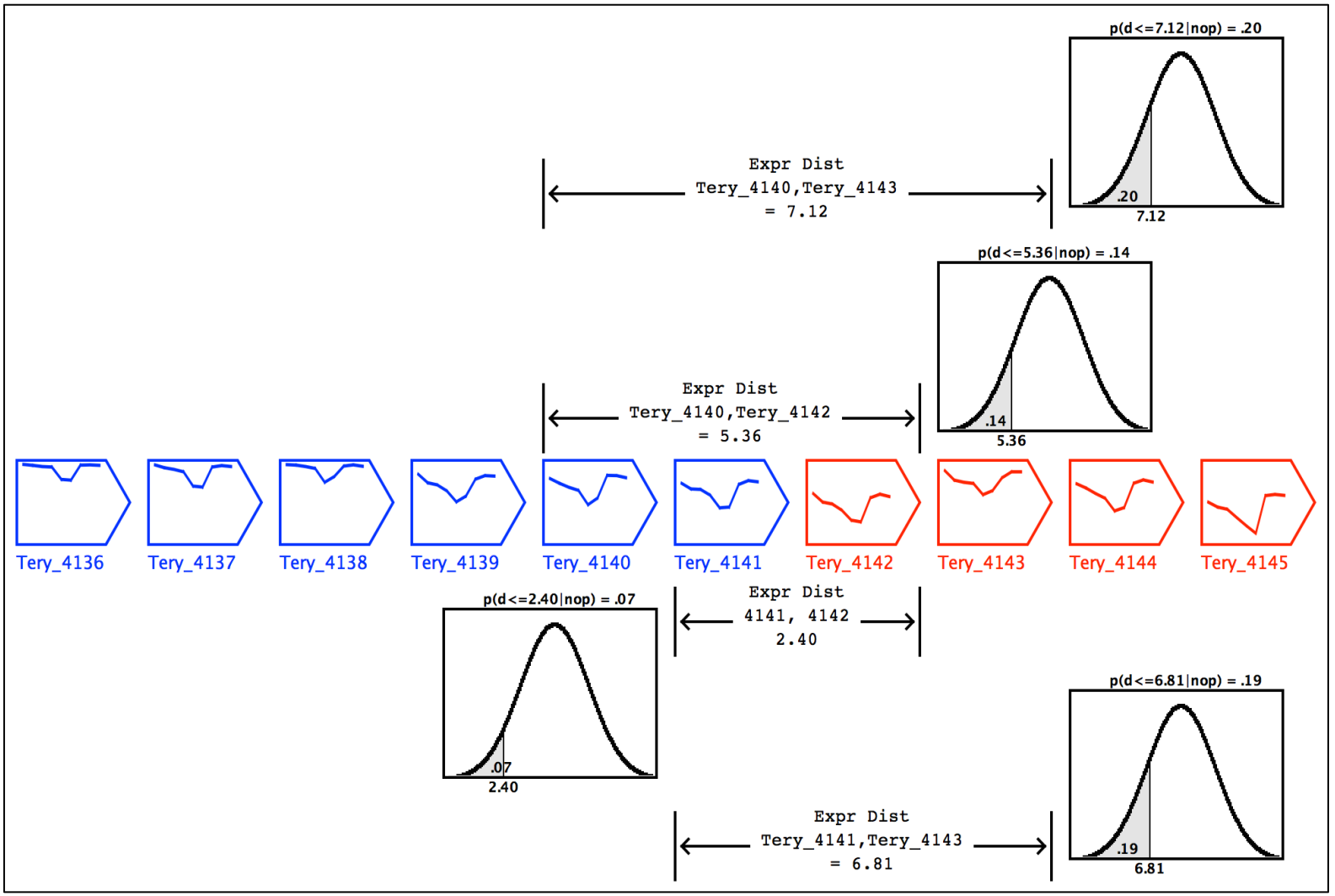


Figure 9. Expression signatures of members of two adjacent prior predicted operons in *Trichodesmium erythraeum*. A) Upstream operon (6 genes). (B) Downstream operon (4 genes). (C) Superimposed signatures.

Figure 10. Analysis of adjacent prior predicted operons from Figure 9. Euclidean expression distances between last 2 ” genes and first 2 “red” genes are shown. Expression distances are converted to probabilities using the distribution fitted to negative training set for *T. erythraeum*. For example in top distance bar and Gaussian curve, the expression distance between Tery_4140 and Tery_4143 is 7.12. The corresponding shaded area (.20 of total area) under the Gaussian curve is the cumulative probability of observing two genes whose expression is at least as similar as the expressions of Tery_4140 and Tery_4143, when the two genes are *not* members of the same operon. The E-value for merging the two prior predicted operons is the product of the cumulative probabilities: .20 * .14 * .07 * .19 = .00037.


p(d<=7.12|nop) = .20
Expr Dist
Tery_4140,Tery_4143
= 7.12
.20
7.12
p(d<=5.36|nop) = .14
Expr Dist
Tery_4140,Tery_4142
= 5.36
.14
5.36
Tery_4136
Tery_4137
Tery_4138
Tery_4139
Tery_4140
Tery_4141
Tery_4142
Tery_4143
Tery_4144
Tery_4145
p(d<=2.40|nop) = .07
Expr Dist
4141, 4142
2.40
.07
2.40
p(d<=6.81|nop) = .19
Expr Dist
Tery_4141,Tery_4143
= 6.81
.19
6.81

Organism	Strain	Operon	Operon length (genes)	Reference
Anabaena	PCC 7120	*devBCA*	3	Fiedler et al., 1998
Anabaena	PCC 7120	*nir-nrtABC-narB*	5	Frias et al., 1997
Frymella		*apcA1B1C1E1*	4	Houmard et al., 1996
Frymella	UTEX 481	*cpcBAEF*	4	Lomax et al., 1987
Microcystis	(various)	*mcyABC*	3	Mikalsen et al., 2003
Prochlorococcus	MED 4	*mutT-phrB*	2	Osburne et al., 2010
Synechococcus	PCC 7002	*isiAB*	2	Leonhardt & Straus, 1992
Synechococcus	PCC 7002	*sufBCDS*	4	Wang et al., 2004
Synechococcus	PCC 7942	*cmpABCD*	4	Omata et al., 1999
Synechococcus	PCC 7942	*kaiABC*	3	Kutsuna et al., 2005
Synechococcus	PCC 7942	*nirA-nrtABCD-narB*	6	Suzuki et al., 1993
Synechocystis	PCC 6803	*cmpABCD*	4	Omata et al., 2001
Synechocystis	PCC 6803	*ctaCDE-orf4-ctaF*	5	Peschek 1996
Synechocystis	PCC 6803	*psaAB*	2	Muramatsu & Hihara, 2006

Table 1. Examples of experimentally validated cyanobacterial operons.

Organism	Sample size	Mean expression distance	Standard deviation of expression distance
Crocosphaera	1171	19.8	12.0
Prochlorococcus	637	16.3	10.7
Trichodesmium	1739	13.2	7.22

Table 2. Statistics for negative sample sets for each study. Gaussian distributions fitted to each negative sample set are used to compute E-values.

Organism	Op1, 1st Gene	Op2, 1st Gene	E-value	Merged Length
Crocosphaera	CwatDRAFT_6659	CwatDRAFT_6656	1.42E-2	6
	CwatDRAFT_6266	CwatDRAFT_6271	1.13E-4	7
	CwatDRAFT_6345	CwatDRAFT_6349	7.86E-7	6
	CwatDRAFT_5350	CwatDRAFT_5353	0	5
	CwatDRAFT_4990	CwatDRAFT_4993	0	5
	CwatDRAFT_4055	CwatDRAFT_4051	3.42E-2	6
	CwatDRAFT_3360	CwatDRAFT_3358	6.86E-3	4
	CwatDRAFT_2389	CwatDRAFT_2396	1.67E-2	9
	CwatDRAFT_0743	CwatDRAFT_0683	0	7
Prochlorococcus	PMM0201	PMM0203	4.09E-12	6
	PMM0392	PMM0395	2.02E-5	5
	PMM0747	PMM0753	3.25E-3	8
	PMM0928	PMM0931	5.59E-18	5
	PMM0940	PMM0943	9.19E-3	5
	PMM1053	PMM1056	5.62E-10	5
	PMM1099	PMM1102	4.42E-2	7
	PMM1333	PMM1336	1.81E-5	6
	PMM1529	PMM1533	1.72E-12	10
	PMM1533	PMM1540	1.22E-19	26
Trichodesmium	Tery_0062	Tery_0065	4.31E-18	5
	Tery_0754	Tery_0757	1.33E-3	5
	Tery_1181	Tery_1185	2.52E-5	6
	Tery_1263	Tery_1268	2.15E-8	5
	Tery_1367	Tery_1369	5.32E-7	4
	Tery_1520	Tery_1525	1.57E-4	7
	Tery_2157	Tery_2162	2.97E-4	7
	Tery_2329	Tery_2334	3E-8	7
	Tery_2858	Tery_2861	1.57E-13	6
	Tery_3102	Tery_3106	3.74E-4	9
	Tery_3879	Tery_3882	2.66E-3	5
	Tery_4035	Tery_4040	7.72E-16	7
	Tery_4136	Tery_4142	3.72E-4	10
	Tery_4256	Tery_4258	2.83E-16	5
	Tery_4375	Tery_4379	8.29E-3	7
	Tery_4487	Tery_4489	3.51E-3	4

Table 3. Merge candidates with E-values <= .05.

Tool	Number of Studies	Graphics	Exploration	Clustering
BETR	1-2	None	None	None
Biggests	1	Time course	Limited	Biclustering
CAGED	1	Heat map with tree	None	Bayesian
DynaMiteC	2	None	None	Impulse Model
EDGE	2	Boxplots, Eigengenes	Limited	Available, method not specified
GATE	1	Hexagonal heat maps	None	None
GQL	1	Time course	None	Hidden Markov Models
LIMMA	2	None	None	None
MVQueries	1	None	None	None
PESTS	1-2	Limited	Limited	Euclidean
PRIISM	1	None	None	None
REST	Multiple	None	None	None
SAM	Multiple	Statistical graphs	None	None
STEM	1	Time course	None	Model profile similarity
TimeClust	1	Time course, self-organizing maps, dendrograms	Limited	Many methods
TRAM	1	Time course	Limited	Hidden Markov Models

Table 4. Analysis methods for time course datasets reviewed by Bar-Joseph (Bar-Joseph et al., 2012).