UNIVERSITY OF CALIFORNIA

SANTA CRUZ

COMPARATIVE GENOMICS AND NATURAL DISTRIBUTIONS OF PHENOTYPICALLY DISTINCT STRAINS OF THE NITROGEN-FIXING CYANOBACTERIUM *CROCOSPHAERA WATSONII*

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

DOCTOR OF PHILOSOPHY

in

OCEAN SCIENCES

by

Shellie R. Bench

June 2012

The Dissertation of Shellie R. Bench is approved:

Professor Jonathan P. Zehr, Chair

Professor Fitnat Yildiz

Professor Jack Meeks

Professor Eric Webb

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

Title i

Copyright ii

Table of contents iii

List of Figures and Tables vi

Abstract ix

Dedication xi

Acknowledgements xii

Chapter 1: Introduction 1

Nitrogen-fixing cyanobacteria in the global oceans 1

Physiological studies in cultivated *Crocosphaera* 2

Phenotypic variation in *Crocosphaera* isolates 3

Previous studies of genetic variation in *Crocosphaera* populations and isolates 6

Unanswered questions and focus of the dissertation 9

References 11

Chapter 2: Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features 16

Abstract 17

Introduction 18

Materials and Methods 23

WH0003 genomic DNA amplification and 454 sequencing 23

Sequence assembly and analysis and ORF identification 24

Transposase annotation 26

Genome comparisons 27

Microarray gene expression 28

Results and Discussion 29

Broad Genome Comparison 29

Shared Genome Features 30

Strain-specific Genome Features 32

Transposase Gene Comparisons 38

Conclusions 40

Acknowledgements 41

References 42

Tables 49

Figures 53

Chapter 3: Investigation of *Crocosphaera watsonii* phenotypes through whole genome comparison of six strains 58

Abstract 59

Introduction 61

Materials and Methods 63

Strain growth and genomic DNA isolation and sequencing 63

Genome assembly and annotation 65

Genome comparisons.. 67

Analysis and PCR of specific groups of functional genes 68

Results 73

Genome characteristics and sequence replication within genomes 73

Shared and unique ORFs among the six genomes 74

Phylogenetic analysis and comparison of metabolic capabilities 78

Discussion 82

Conclusions.. 91

References 93

Figures and Tables 98

Chapter 4: Natural abundances of two *Crocosphaera* types in the North and South Pacific 115

Abstract 116

Introduction 117

Methods 121

Design and testing of qPCR assays.. 121

Sample collection, DNA extraction and qPCR of cruise samples 123

Results and Discussion 126

Conclusions.. 132

References.. 133

Figures and Tables 136

Chapter 5: Conclusions and future directions 142

Appendix 1: Additional analysis of WH0003 and WH8501 transposase ORFs.. 146

Methods 146

Results and Discussion 147

References.. 151

Figures 153

Appendix 2: Chapter 2 supplemental tables 157

Appendix 3: Chapter 3 supplemental tables and figures 160

Appendix 4: Chapter 4 supplemental tables and figures 165

List of additional supplementary files 168

List of Figures and Tables

Chapter 2: Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features

Table 1. Genome assembly information and annotation summary 49

Table 2. ORFs within WH0003 strain specific genome region 50

Table 3. Transposase IS family distribution in both genomes 52

Figure 1. Nucleotide BLAST similarity of open reading frames (ORFs) and intergenic spaces (IGSs) of the two *C. watsonii* genomes to the opposite strain.. 53

Figure 2. Comparison of open reading frames and intergenic spaces between two *C. watsonii* genomes... 54

Figure 3. Taxonomic distribution of 930 ORFs in the WH0003 genome that show little or no sequence similarity to the WH8501 genome 55

Figure 4. Alignment of WH8501 and WH0003 contigs illustrating a 25kb region specific to the WH0003 genome 56

Figure 5. Expression of four IS family genes and dnaA in *C. watsonii* WH8501 57

Chapter 3: Investigation of *Crocosphaera watsonii* phenotypes through whole genome comparison of six strains

Table 1. *Crocosphaera watsonii* strain origins and phenotypes 98

Table 2. Genome sizes and gene content statistics for six *Crocosphaera watsonii* strains 99

Figure 1. Abundance of repeated sequences in each *C. watsonii* genome 100

Table 3. Counts of repeated sequences in each *C. watsonii* genome 101

Figure 2. Abundance and distribution of IS families in each of the six *C. watsonii* genomes.. 102

Figure 3. Percentage of ORFs shared between *C. watsonii* strains 103

Table 4. Percentage of ORFs shared between *C. watsonii* strains 104

Figure 4. Presence/absence of all ORFs in six *C. watsonii* genomes 105

Figure 5. Counts of ORFs found in genomes of 2, 3, 4 and 5 *C. watsonii* strains 106

Figure 6. Combined counts of ORFs found in genomes of 2, 3, 4 and 5 *C. watsonii* strains 107

Table 5. Presence/absence of 24 genes in putative EPS-critical region identified in the WH0003 genome 108

Figure 7. Phylogenetic relationship of six *C. watsonii* strains and two *Cyanothece* species, based on 25 genes 109

Table 6. Estimates of Evolutionary Divergence among six *C. watsonii* strains and two *Cyanothece* species, based on 25 genes 110

Table 7. Counts of phosphorous-related genes in each *C. watsonii* genome ... 111

Table 8. Counts of iron-related genes in each *C. watsonii* genome 112

Table 9. Counts of Photosystem I and II genes in the genome of each *C. watsonii* strain 113

Figure 8. Evolutionary relationships of *C. watsonii psbC* and *isiA* genes......... 114

Chapter 4: Natural abundances of two *Crocosphaera* types in the North and South Pacific

Table 1. Primer and probe sequences and gene information for each qPCR assay locus 136

Table 2. qPCR results for all loci with DNA from cultivated *C. watsonii* strains 137

Figure 1. North Pacific cruise sample locations 138

Figure 2. Percent of each type of *C. watsonii* found on the 0.2µm filter 139

Figure 3. Patterns of *Crocosphaera* abundance in the N. Pacific 140

Figure 4. Depth profiles of *C. watsonii* and two sub-types at three stations in the S. Pacific.. 141

Appendix 1: Additional analysis of WH0003 and WH8501 transposase ORFs

Figure S2. Distribution of transposase IS families in both genomes 153

Figure S2. Distribution of IS families in the *Crocosphaera* WH8501 genome 154

Figure S3. Schematic representation of transposase gene and conserved flanking regions 155

Figure S4. Fasta formatted ORF and flanking conserved regions for the IS5 family ISCwa22_aa1 transposase 156

Appendix 2: Chapter 2 supplemental tables

Table S1. WH8501 Re-annotated Transposase ORFs .... *see add. supp. files*

Table S2. WH8501 Transposase Isoform Counts 157

Table S3. WH8501 strain-specific ORFs .... *see add. supp. files*

Table S4. WH0003 strain-specific ORFs .... *see add. supp. files*

Table S5. WH0003 ORFs in longest strain-specific region 159

Table S1. WH8501 sequences included in transposase alignment and % ID of each to alignment consensus sequence shown in Figure S4 .... *see add. supp. files*

Appendix 3: Chapter 3 Supplemental tables and figures

Figure S1. Percent identity bins of all ORFs in each genome vs. other 5 genomes.. 160

Table S1. ORF IDs for sequences included in alignment of 25 genes for phylogenetic analysis .... *see add. supp. files*

Table S2. LysM gene PCR primer sequences and product sizes 161

Table S3. ORF IDs and lengths for the 4 LysM gene forms in 6 *C. watsonii* genomes and PCR results for 2 additional strains 162

Table S4. Strain-specific ORFs from six *C. watsonii* genomes .... *see add. supp. files*

Table S5. Counts of sequences in each category, based on genomes in which the sequences was found 163

Table S6. Phenotype-specific ORF IDs and functions from six *C. watsonii* genomes .... *see add. supp. files*

Appendix 4: Chapter 4 supplemental tables and figures

Figure S1. qPCR Standard curves of *nifH* linearized plasmids and genomic DNA from *C. watsonii* strains using *nifH* primer-probe set 165

Table S1. Flow cytometry (FCM) counts of *Crocosphaera* cells compared to *nifH* gene copies (qPCR) for samples in the North Pacific 166

Table S2. Flow cytometry (FCM) counts of *Crocosphaera* cells compared to *nifH* gene copies (qPCR) for samples in the South Pacific 167

Abstract

Shellie Bench

Comparative genomics and natural distributions of phenotypically distinct strains of the nitrogen-fixing cyanobacterium *Crocosphaera watsonii*

Crocosphaera watsonii is an ecologically important marine unicellular diazotrophic cyanobacterium. It is often abundant in oligotrophic ocean regions where it provides fixed nitrogen to nutrient-limited phytoplankton communities. Previous genetic studies have observed genetic rearrangements but very little sequence variation among natural populations or cultivated strains of *Crocosphaera*. Those strains exhibit two phenotypes (large- and small-cell) with characteristics that suggest different ecological roles and niches. Prior to this work, the genetic basis for the phenotypic differences was unknown, and molecular methods for enumerating natural *C. watsonii* could not differentiate between phenotypes. To address those unanswered scientific questions, these studies compared the genomes of six *C. watsonii* strains, three of each phenotype, which were isolated over large spatial and temporal distances. A large portion of those genome sequences were shared among all strains with nearly 100% nucleotide identity. However, there were also genes that were specific to each strain, and others were specific to each phenotype, including some which could explain phenotypic differences (e.g. EPS biosynthesis). Relative to small-cell strains, large-cell strains had larger genomes and additional genetic capabilities, including possibly increased adaptations to iron and phosphorus limitation. Clustering based on genome sequences and content showed that strains with a common phenotype were evolutionarily most closely related, regardless of their time and location of isolation. Surprisingly, the genome of the *C. watsonii* type-strain, WH8501, was quite unusual, even compared to those with the same phenotype, suggesting it may not be appropriately representative of the species. To investigate distributions of *Crocosphaera* types in the marine environment, molecular assays were developed, based on phenotype-specific genes, and applied to samples from the North and South Pacific. In those samples, small-cells dominated in the upper 75 m where abundance of both types was much greater, while large-cells dominated in samples with lower counts between 100 m and 175 m. There was also more evidence that large-cells form aggregates in the N. Pacific. Future studies will be important to determine which of the initial *C. watsonii* patterns described here can be generalized, both in genomes, and in natural distributions of the two types.

This work is dedicated to the three boys that I miss every day;

Timothy, Rebound and Mordred.

And to my grandparents, Liz and Tom;

who I wish could have been here to share this accomplishment with me.

Acknowledgements

I am fortunate to have received a great amount of support during the journey to my Ph.D., and accordingly, I have many people to thank. First and foremost, the person who deserves the most credit for helping me complete this degree is my best friend and partner in life, Chris. I cannot even begin to count all of the times and all of the ways he supported me through the many ups and downs of the last few years. It is likely that I would not have been able to accomplish this goal without him, and I could never thank him enough for his unwavering love and support.

I also owe a very great deal to my advisor, Jon Zehr. The scientific opportunities he provided and the support he gave me were absolutely key in enabling me to reach this point in my career. When I moved to California and joined Jon's lab, I essentially put my professional future in his hands, and I've never regretted that decision for a moment.

In addition, my fellow Zehr lab members, current and former, have been wonderfully supportive, personally and scientifically. Special thanks to Mary, who manages to keep a huge lab running smoothly, and whose efforts made it so much easier to complete my experiments, and who I am also glad to consider a great friend. And, of course, Nicole Pereira, I am so lucky to have found a colleague and friend with whom I share a complete personal understanding and connection, and I am so glad we became so close. People who deserve thanks for contributing to the results in this dissertation, and who I am also happy to count among my friends include Jim Trip, Irina Shilova, Martha Arciniega, Phil Heller, Ildiko Frank, and Julie Robidart. I

also want to specifically thank Jason and Kendra for the many times they provided personal and professional help. There are just too many other lab-mates to thank individually, but I hope you each know how much I appreciate your help, support and friendship.

I am also grateful to my committee members, Fitnat, Jack, and Eric for their support, understanding, and brilliant insights. Thanks, as well, to the professors and my fellow students in the UCSC Ocean Sciences department, especially my office mates in both the dungeon and D402. They are a group of wonderful people with brilliant minds, and have been an amazing resource. It has been great to be surrounded by them during my work here.

I also owe many, many thanks to Bob Feldman for putting his full trust in my abilities very early in my career, and for opening the door to the amazing world of microbial ecology. And finally, for their patience, steadfast support, and love through the years, a million thanks to my family and friends, especially Mom, Ginny, Melani and Isabelle.

The text of this dissertation includes the following previously published manuscript: Bench, S. R., Ilikchyan, I. N., Tripp, H. J. and Zehr, J. P. (2011). "Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features." *Frontiers in Aquatic Microbiology*. 2 (261).

J.P. Zehr, a co-author on that publication, directed and supervised the research, which forms the basis for the dissertation. My contribution to the work in that manuscript was: 1) Providing genomic DNA for sequencing of the WH0003 genome, 2) Bioinformatic assembly and annotation of the WH0003 genome 3) comparisons of the WH8501 and WH0003 genomes and 4) making the figures and writing the manuscript.

Chapter 1: Introduction

Nitrogen-fixing cyanobacteria in the global oceans

All marine ecosystems depend on primary production that is, for the most part, carried out by phytoplankton. In most of the global oceans, the phytoplankton community is dominated by cyanobacteria, of which the most abundant and widespread are *Prochlorococcus* and *Synechococcus* (Waterbury et al., 1986; Goericke and Welschmeyer, 1993; Liu et al., 1997; Partensky et al., 1999; Scanlan and West, 2002). However, in wide areas of the oceans, production of phytoplankton communities is limited by low available nutrients, particularly bioavailable nitrogen (N). In those areas, nitrogen (N_2)-fixing cyanobacteria, despite being much less abundant, play a key ecological role by providing fixed N to the phytoplankton community (Karl et al., 1997; Karl et al., 2002; Bonnet et al., 2009; Kitajima et al., 2009; Shiozaki et al., 2010). This new N enables more primary production thereby supporting the entire marine food web.

The major marine N_2-fixing (i.e. diazotrophic) cyanobacteria fall into three categories which are defined by morphology and life-style, as well as nitrogenase (*nifH*) gene phylogeny. One group forms symbioses with eukaryotic phytoplankton (e.g. *Richelia* and *Calothrix*). The second group includes free-living filamentous species such as *Trichodesmium*, and the third group are free-living unicellular cyanobacteria (i.e. UCYN) including *Crocosphaera*. With blooms that are easily observed in surface waters, *Trichodesmium* was historically considered the major contributor of fixed N (Capone et al., 1997). However a variety of studies have since shown that UCYN, which are distributed throughout tropical and sub-tropical oceans, could dominate global marine N_2-fixation. For example, direct counts of unicellular phycoerythrin-containing cyanobacteria (presumed to be diazotrophs) were observed at concentrations of 10^4 to 10^7 cells/liter in the North Pacific (Zehr et al., 2001; Church et al., 2005), and near 10^5 cells/liter in the Atlantic (Falcon *et al.*, 2004). A variety of studies using quantitative polymerase chain reaction (qPCR) of the *nifH* gene have found UCYN abundances between 10^3 and 10^6 gene copies/ liter in the tropical South Pacific (Moisander *et al.*, 2010), the North Pacific (Church et al., 2005; Church et al., 2008), the tropical Atlantic (Langlois *et al.*, 2008), and the South China Sea (Moisander *et al.*, 2008). Corresponding high levels of *in-situ* unicellular cyanobacterial N_2 fixation (0.02 to 4.5 nmol liter^{-1} hour^{-1}) have also been recorded (Zehr et al., 2001; Falcon et al., 2004; Montoya et al., 2004; Kitajima et al., 2009; Moisander et al., 2010). These studies have clearly demonstrated the importance of UCYN in global marine nutrient cycling.

Physiological studies in cultivated Crocosphaera

Cultured representatives of *Crocosphaera* have been isolated from the Atlantic and Pacific Oceans and from latitudes ranging from 28^{o}S to 24^{o}N. To date, all isolates are considered strains of the species *Crocosphaera watsonii*, and the established type-strain is WH8501, isolated from the S. Atlantic in 1984 (Waterbury et al., 1988). Researchers have investigated the molecular and physiological responses of *C. watsonii* (primarily in WH8501) to a variety of environmental changes. In response to phosphorus limitation, *C. watsonii* has shown reduced growth and N_2-fixation (Falcon et al., 2005) and increased expression of phosphorus scavenging genes (Dyhrman and Haley, 2006). Interestingly, *Crocosphaera* exhibits increased carbon and N_2-fixation rates under increased CO_2 and Fe replete conditions, suggesting it may be well adapted to rising atmospheric CO_2 concentrations (Fu et al., 2008). Another study showed that *Crocosphaera* may have a relatively narrow window of oxygen concentrations where it can fix N_2, which inexplicably was well below atmospheric oxygen concentrations (Compaoré and Stal, 2010). *C. watsonii* is also apparently adapted for growth and use of iron-rich nitrogenase proteins in iron limited environments through mechanisms such as regulated expression of iron response genes (Webb et al., 2001; Shi et al., 2010) and significant recycling of iron metalloproteins (Tuit et al., 2004; Saito et al., 2011). In addition to the iron genes, a number of metabolic functions are regulated on a diel basis, such as photosynthesis and N_2-fixation which are temporally separated during the diel cycle (Mohr et al., 2010; Pennebaker et al., 2010; Shi et al., 2010; Hewson et al., 2009; Saito et al., 2011). There is some evidence that diel gene expression patterns may be driven by circadian control that regulates transcription through changes in genomic DNA topology (Pennebaker et al., 2010) .

Phenotypic variation in Crocosphaera isolates

In addition to the experiments that have been carried out on single isolates, additional studies have found physiological differences among *C. watsonii* cultivated strains. Based on these differences, *Crocosphaera* strains can be divided into two broad phenotypic categories. The first phenotype is characterized primarily by larger cell diameters (over 4 µm) and production of copious amounts of extracellular polysaccharide (EPS) (Webb et al., 2009; Sohm et al., 2011). This large cell-type also has a wider temperature range for growth, higher per-cell N_2- fixation rates and higher photosynthetic efficiencies, as measured by F_v/F_m. In contrast, the small-cell (i.e. < 4 µm) phenotype, which includes the WH8501 type-strain, produces at least 10 fold less EPS and has lower F_v/F_m, and lower N_2- fixation rates (Webb et al., 2009; Sohm et al., 2011). Additionally, there is evidence of differences among isolates in their phosphorus uptake capabilities. Specifically, some strains appear to be missing a high affinity phosphate transporter gene (*pstS*) and/or an alkaline phosphatase gene (Dyhrman and Haley, 2006). There does not appear to be a different geographic distribution for the two phenotypes, as both types have been isolated from the North Pacific and South Atlantic basins (Webb et al., 2009).

In light of their observed differences, the two phenotypes may have different ecological niches and may have different influences in marine biogeochemical cycling. The different photosynthetic efficiencies of the two types suggests they may be adapted to different light-levels, and therefore may be found at different depths in the water column, similar to the paradigm of high-light and low-light adapted clades of *Prochlorococcus* (West and Scanlan, 1999; Rocap et al., 2003; Scanlan et al., 2009). In addition, the significantly different rates of per-cell N_2-fixation would result in very different amounts of new nitrogen input to the ecosystem if one type was largely dominant over the other. The most apparent difference between the types is the level of EPS production, which also has ecological relevance. The exact purpose of EPS production in *Crocosphaera* is not well understood, but EPS has been shown to have cell-protective properties in a variety of cyanobacteria (Pereira *et al.*, 2009), which could make the large-cell types more resistant to cellular degradation. Furthermore, cell aggregation has been observed in the EPS producing strains, and in EPS-associated cells in natural samples (Webb et al., 2009; Sohm et al., 2011). Such aggregates could result in a higher sinking rate for the large-cell type, thereby increasing their relative contribution of nutrient export form the photic zone. This export, analgous to marine snow, would also be enriched in carbon, because EPS has a much higher C:N ration than cellular material (Passow et al., 2001; Sohm et al., 2011). The only study to date that has investigated the two types in natural samples used microscopy to determine the relative abundance of large cells and small cells at a single depth during a research cruise in the western Equatorial and South Pacific (Webb et al., 2009). In that sample set, *Crocosphaera* were relatively low abundance (100-1400 cells/liter), the small cells slightly outnumbered the large-cells, and some aggregation of the large-cells was observed (Webb et al., 2009). However, no difference was found between the two types in response to nutrient amendments.

Previous studies of genetic variation in Crocosphaera populations and isolates

Marine cyanobacteria, as direct descendents of ancient phototrophs with deeply rooted phylogenies, typically have substantial sequence diversity, even among closely related species (Zhao and Qin, 2007; Dufresne et al., 2008). This observation has been supported by a wide variety of genome sequences and large scale environmental sequencing efforts that have focused on the most abundant taxa; *Prochlorococcus* and *Synechococcus* (Rocap et al., 2002; Ernst et al., 2003; Rocap et al., 2003; Venter et al., 2004; Rusch et al., 2007; Scanlan et al., 2009; Partensky and Garczarek, 2010). In contrast to those studies, and despite clear phenotypic differences between strains, sequences of *Crocosphaera* strains showed surprisingly little sequence variation. An examination of the 16S-23S ITS region in nine *C. watsonii* strains only found six single base positions varied out of nearly 900bp (Webb et al., 2009). That is quite different from studies that have shown up to 45% variation in the 16S-23S ITS region of *Synechococcus* and *Prochlorococcus* even when adjacent 16S rRNA sequences were over 97% identical (Rocap et al., 2002; Ernst et al., 2003; Brown and Fuhrman, 2005).

The contrasting genetic diversity patterns between *Crocosphaera* and sympatric non-N_2-fixing cyanobacteria was also observed in sequences beyond the rRNA operon. In a comparison of five functional genes within seven *Crocosphaera* strains and two large-insert environmental clones (BACs), all sequences were > 99%

identical (Zehr et al., 2007) . That level of identity in the context of the wide geographic and temporal distance between the collection of strains and BACs indicated that there is remarkably little DNA mutation accumulation among strains in this genus. This finding was supported by a metagenomic study that found similar levels of sequence identity between environmental sequences and the *C. watsonii* WH8501 genome (Hewson et al., 2009). Comparable genomic analyses in *Synechococcus* and *Prochlorococcus* showed nucleotide sequence identity of orthologous genes were often as low as 50-78% even when comparing species with > 96% 16S rRNA identity (Coleman et al., 2006; Zhao and Qin, 2007; Dufresne et al., 2008) and large scale environmental sequencing showed a similar degree of sequence variation in natural populations (Rusch et al., 2007). These results suggested the evolutionary process in *Crocosphaera* is quite different than what has been observed in *Synechococcus* and *Prochlorococcus* which appear to follow the canonical evolutionary paradigm of gaining genetic variation through mutation accumulations.

Analysis of the *C. watsonii* WH8501genome and initial sequence comparisons revealed that *Crocosphaera,* despite its apparent lack of sequence mutation accumulation, may acquire diversity through genetic mobility. An unusually high number of transposase genes was found in the *C. watsonii* WH8501 draft genome, along with evidence for positive evolutionary selection in those genes (Mes and Doeleman, 2006). Additionally, expression of *Crocosphaera* transposase genes was observed in a metatranscriptome study, demonstrating that natural *Crocosphaera* populations could be utilizing a similar process as cultured isolates (Hewson et al., 2009). This is another aspect in which *Crocosphaera* are distinguished from sympatric non-N_2-fixing marine cyanobacteria, such as *Synechococcus* and *Prochlorococcus*, which generally lack transposase genes. The most convincing evidence that *Crocosphaera* acquire variation through genetic mobility was the finding of large scale genomic rearrangements when BAC sequences were aligned to the *C. watsonii* WH8501 draft genome, (Zehr et al., 2007). The transposase genes flanking those rearrangements also suggested that such genetic mobility is likely mediated by the abundant transposase genes in the WH8501 genome (Zehr et al., 2007).

Transposons are highly abundant mobile genetic elements that mediate genome shuffling within and between a wide variety of taxa (Aziz *et al.*, 2010), including plants (Feschotte *et al.*, 2002), mammals (Lander et al., 2001; Waterston et al., 2002), bacteria (Mahillon et al., 1999; Touchon and Rocha, 2007) and archaea (Goodchild *et al.*, 2004; Filee *et al.*, 2007). Insertion sequences (IS's), a type of transposable element, have been observed in prokaryotes at frequencies from zero to over 300 per genome. Within prokaryotic genomes, IS elements can have significant metabolic costs associated with expressing the transposase genes (the protein coding portion of transposons that catalyze DNA movement), as well as increasing the probability of inactivating genes or regulatory regions and increasing homologous recombination between multi-copy IS elements in the genome (Touchon and Rocha, 2007). Genomic IS abundance has also been correlated with both genome size and the frequency of horizontal gene transfer (HGT) (Touchon and Rocha, 2007). Some of the highest numbers of transposable elements have been found in proteobacterial and cyanobacterial genomes, although a large fraction of cyanobacterial species lack any recognized transposase genes in their genomes (e.g. *Prochlorococcus*), (Kaneko et al., 2007; Touchon and Rocha, 2007; Frangeul et al., 2008; Stucken et al., 2010). Researchers have also found abundant transposases in the deep oceans, suggesting they play an important role in microbial communities in a variety of marine environments (Konstantinidis et al., 2009). Based on these observations and the evidence for the role of transposases in WH8501 genome rearrangements discussed above, these elements appear to be important in *Crocosphaera* genomes. Primarily, transposases appear to be critical in *Crocosphaera*'s unusual strategy to acquire genetic diversity in the context of very little random mutation accumulation.

Unanswered questions and focus of the dissertation

In the context of marine microorganisms, *Crocosphaera* have been relatively well studied. However, prior to the work described in this dissertation, a number of areas remained nearly unexplored. Primarily, despite increasing evidence that there are multiple phenotypes of *C. watsonii*, with ecologically important differences, the genetic basis for those differences was essentially unknown. Because of that lack of knowledge, no molecular markers were available that could differentiate between phenotypes in natural populations. As such, the distributions and relative abundances of the *C. watsonii* sub-types were also completely unknown.

The genome sequence from *Crocosphaera* WH8501, and some comparisons to that genome, provided intriguing evidence of high sequence conservation, high transposase abundance and significant genomic rearrangement. However, it was not clear if a strain with the opposite phenotype would have the same genomic signatures. The comparison of WH8501 (small-cell) genome to the WH0003 (large-cell) genome was undertaken to explore that question.

After discovering sequences and genome features unique to each of those two strains, it remained to be seen whether those features were representative of their respective phenotypes. To address that issue, four additional genomes, two of each phenotype, were sequenced and compared. Based on those six genomes, a suite of genes have been identified as characteristic of each *Crocosphaera* phenotype, including some that indicate the two types may have different adaptations to environmental changes. A subset of those genes were crucial in the development of molecular assays that can differentiate between the two types. The initial experiments, described in this dissertation, that have applied those assays to natural samples are first steps in acquiring a critical understanding of natural distributions of the two *Crocosphaera* phenotypes that have quite significant ecological differences.

References:

Aziz RK, Breitbart M, Edwards RA. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* **38**: 4207-4217.

Bonnet S, Biegala IC, Dutrieux P, Slemons LO, Capone DG. (2009). Nitrogen fixation in the western equatorial Pacific: Rates, diazotrophic cyanobacterial size class distribution, and biogeochemical significance. *Global Biogeochem Cycles* **23**: GB3012.

Brown MV, Fuhrman JA. (2005). Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquat Microb Ecol* **41**: 15-23.

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

Church MJ, Jenkins BD, Karl DM, Zehr JP. (2005). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat Microb Ecol* **38**: 3-14.

Church MJ, Bjorkman KM, Karl DM, Saito MA, Zehr JP. (2008). Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol Oceanogr* **53**: 63-77.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768-1770.

Compaoré J, Stal LJ. (2010). Oxygen and the light–dark cycle of nitrogenase activity in two unicellular cyanobacteria. *Environ Microbiol* **12**: 54-62.

Dufresne A, Ostrowski M, Scanlan D, Garczarek L, Mazard S, Palenik B *et al.* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology* **9**: R90.

Dyhrman ST, Haley ST. (2006). Phosphorus scavenging in the unicellular marine diazotroph *Crocosphaera watsonii*. *Appl Environ Microbiol* **72**: 1452-1458.

Ernst A, Becker S, Wollenzien UIA, Postius C. (2003). Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiology* **149**: 217-228.

Falcon LI, Pluvinage S, Carpenter EJ. (2005). Growth kinetics of marine unicellular N_2-fixing cyanobacterial isolates in continuous culture in relation to phosphorus and temperature. *Mar Ecol Prog Ser* **285**: 3-9.

Falcon LI, Carpenter EJ, Cipriano F, Bergman B, Capone DG. (2004). N_2 Fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: phylogeny and in situ rates. *Appl Environ Microbiol* **70**: 765-770.

Feschotte C, Jiang N, Wessler SR. (2002). Plant transposable elements: Where genetics meets genomics. *Nat Rev Genet* **3**: 329-341.

Filee J, Siguier P, Chandler M. (2007). Insertion sequence diversity in Archaea. *Microbiol Mol Biol Rev* **71**: 121-157.

Frangeul L, Quillardet P, Castets A-M, Humbert J-F, Matthijs H, Cortez D *et al.* (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9**: 274.

Fu F-X, Mulholland MR, Garcia NS, Aaron B, Bernhardt PW, Warner ME *et al.* (2008). Interactions between changing pCO_2, N_2 fixation, and Fe limitation in the marine unicellular cyanobacterium *Crocosphaera*. *Limnol Oceanogr* **53**: 2472-2484.

Goericke R, Welschmeyer NA. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Res (I Oceanogr Res Pap)* **40**: 2283-2294.

Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R. (2004). Biology of the cold adapted Archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *J Proteome Res* **3**: 1164-1176.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR *et al.* (2009). In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* **3**: 618-631.

Kaneko T, Nakajima N, Okamoto S, Suzuki I, Tanabe Y, Tamaoki M *et al.* (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res* **14**: 247-256.

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.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58**: 47-98.

Kitajima S, Furuya K, Hashihama F, Takeda S, Kanda J. (2009). Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol Oceanogr* **54**: 537-547.

Konstantinidis KT, Braff J, Karl DM, DeLong EF. (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Appl Environ Microbiol* **75**: 5345-5355.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

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.

Liu H, Nolla HA, Campbell L. (1997). *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. *Aquat Microb Ecol* **12**: 39-47.

Mahillon J, Léonard C, Chandler M. (1999). IS elements as constituents of bacterial genomes. *Res Microbiol* **150**: 675-687.

Mes THM, Doeleman M. (2006). Positive selection on transposase genes of insertion sequences in the *Crocosphaera watsonii* genome. *J Bacteriol* **188**: 7176-7185.

Mohr W, Intermaggio MP, LaRoche J. (2010). Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. *Environ Microbiol* **12**: 412-421.

Moisander PH, Beinart RA, Voss M, Zehr JP. (2008). Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME J* **2**: 954-967.

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.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. (2004). High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**: 1027-1031.

Partensky F, Hess WR, Vaulot D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63**: 106-127.

Partensky Fdr, Garczarek L. (2010). *Prochlorococcus*: advantages and limits of minimalism. *Annual Review of Marine Science* **2**: 305-331.

Passow U, Shipe RF, Murray A, Pak DK, Brzezinski MA, Alldredge AL. (2001). The origin of transparent exopolymer particles (TEP) and their role in the sedimentation of particulate matter. *Cont Shelf Res* **21**: 327-346.

Pennebaker K, Mackey KRM, Smith RM, Williams SB, Zehr JP. (2010). Diel cycling of DNA staining and *nifH* gene regulation in the unicellular cyanobacterium *Crocosphaera watsonii* strain WH 8501 (Cyanophyta). *Environ Microbiol* **12**: 1001-1010.

Pereira S, Zille A, Micheletti E, Moradas-Ferreira P, Philippis RD, Tamagnini P. (2009). Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. *FEMS Microbiol Rev* **33**: 917-941.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180-1191.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Saito MA, Bertrand EM, Dutkiewicz S, Bulygin VV, Moran DM, Monteiro FM *et al.* (2011). Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proceedings of the National Academy of Sciences*.

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

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.

Shi T, Ilikchyan I, Rabouille S, Zehr JP. (2010). Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* **4**: 621-632.

Shiozaki T, Furuya K, Kodama T, Kitajima S, Takeda S, Takemura T *et al.* (2010). New estimation of N_2 fixation in the western and central Pacific Ocean and its marginal seas. *Global Biogeochem Cycles* **24**: GB1015.

Sohm JA, Edwards BR, Wilson BG, Webb EA. (2011). Constitutive extracellular polysaccharide (EPS) production by specific isolates of *Crocosphaera watsonii*. *Front Microbio* **2**.

Stucken K, John U, Cembella A, Murillo AA, Soto-Liebe K, Fuentes-Valdes JJ *et al.* (2010). The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS ONE* **5**: e9235.

Touchon M, Rocha EPC. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* **24**: 969-981.

Tuit C, Waterbury J, Ravizza G. (2004). Diel variation of molybdenum and iron in marine diazotrophic cyanobacteria. *Limnol Oceanogr* **49**: 978-990.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Waterbury JB, Willey JM, Lester Packer and Alexander NG. (1988). Isolation and growth of marine planktonic cyanobacteria. In Methods in Enzymology: Academic Press, pp. 100-105.

Waterbury JB, Watson SW, Valois FW, Franks DG. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can Bull Fish Aquat Sci*: 71-120.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Webb EA, Moffett JW, Waterbury JB. (2001). Iron stress in open-ocean cyanobacteria (*Synechococcus*, *Trichodesmium*, and *Crocosphaera* spp.): Identification of the IdiA protein. *Appl Environ Microbiol* **67**: 5444-5452.

Webb EA, Ehrenreich IM, Brown SL, Valois FW, Waterbury JB. (2009). Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. *Environ Microbiol* **11**: 338-348.

West NJ, Scanlan DJ. (1999). Niche-partitioning of Prochlorococcus populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65**: 2585-2591.

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 U S A* **104**: 17807-17812.

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.

Zhao F, Qin S. (2007). Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. *Genetica* **129**: 291-299.

Chapter 2

Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features[1]

[1] This chapter published as: Bench, S. R., Ilikchyan, I. N., Tripp, H. J. and Zehr, J. P. (2011). "Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features." *Frontiers in Aquatic Microbiology*. 2 (261).

Abstract

Unicellular nitrogen-fixing cyanobacteria are important components of marine phytoplankton. Although non-nitrogen-fixing marine phytoplankton generally exhibit high gene sequence and genomic diversity, gene sequences of natural populations and isolated strains of *Crocosphaera watsonii*, one of the two most abundant open ocean unicellular cyanobacteria groups, have been shown to be 98-100% identical. The low sequence diversity in *Crocosphaera* is a dramatic contrast to sympatric species of *Prochlorococcus* and *Synechococcus*, and raises the question of how genome differences can explain observed phenotypic diversity among *Crocosphaera* strains. Here we show, through whole genome comparisons of two phenotypically different strains, that there are strain-specific sequences in each genome, and numerous genome rearrangements, despite exceptionally low sequence diversity in shared genomic regions. Some of the strain-specific sequences encode functions that explain observed phenotypic differences, such as exopolysaccharide biosynthesis. The pattern of strain-specific sequences distributed throughout the genomes, along with rearrangements in shared sequences is evidence of significant genetic mobility that may be attributed to the hundreds of transposase genes found in both strains. Furthermore, such genetic mobility appears to be the main mechanism of strain divergence in *Crocosphaera* which do not accumulate DNA microheterogeneity over the vast majority of their genomes. The strain-specific sequences found in this study provide tools for future physiological studies, as well as genetic markers to help determine the relative abundance of phenotypes in natural populations.

Introduction

Marine phytoplankton, which are dominated by cyanobacteria in most of the world's open oceans, are important in global marine biogeochemical cycles and account for half of global carbon fixation (Waterbury et al., 1986; Goericke and Welschmeyer, 1993; Liu et al., 1997; Partensky et al., 1999; Scanlan and West, 2002). The immense genetic diversity of phytoplankton communities has been revealed through rRNA sequences and genomic sequencing of cultivated species, as well as large scale environmental sequencing efforts (Rocap et al., 2002; Ernst et al., 2003; Rocap et al., 2003; Venter et al., 2004; Rusch et al., 2007; Partensky and Garczarek, 2010). As direct descendents of ancient phototrophs with deeply rooted phylogenies, it is not surprising that cyanobacteria typically show a large amount of genomic sequence heterogeneity, even among closely related species (Zhao and Qin, 2007; Dufresne et al., 2008). However, the genome diversity among *Crocosphaera* strains and populations is an intriguing deviation from that observed trend.

In oligotrophic regions, phytoplankton production is often limited by nutrients, especially nitrogen (N), and in those areas, nitrogen (N_2) fixation provides an important source of new N that supports primary productivity (Karl et al., 1997; Karl et al., 2002; Bonnet et al., 2009; Kitajima et al., 2009; Shiozaki et al., 2010). The major marine N_2-fixing cyanobacterial taxa can be categorized into three groups based on life-style and morphology: 1) symbiotic, including *Richelia* spp. and *Calothrix* spp. 2) free-living and filamentous, like *Trichodesmium* spp. and 3) free-living and unicellular, such as *Crocosphaera* spp. The two most abundant taxa of unicellular diazotrophs, as defined by nitrogenase gene (*nifH*) phylogeny, are the uncultivated Group A (i.e. UCYN-A) and Group B, which is represented in culture by a number of *Crocosphaera watsonii* strains. Previously, free-living unicellular diazotrophs were thought to be relatively minor contributors to total marine N_2 fixation (Capone et al., 1997). However, more recent studies have reported high abundance using qPCR and direct cell counts (Zehr et al., 2001; Falcon et al., 2004; Church et al., 2005a; Church et al., 2008; Langlois et al., 2008; Moisander et al., 2008; Moisander et al., 2010), and measured high rates of *in-situ* unicellular cyanobacterial N_2 fixation (Zehr et al., 2001; Falcon et al., 2004; Montoya et al., 2004; Kitajima et al., 2009; Moisander et al., 2010), demonstrating that unicellular diazotrophs are often significant contributors of new N in the global ocean.

Crocosphaera spp. strains have been isolated from the Atlantic and Pacific Oceans between 28°S and 24°N latitudes. All isolates are strains of the species *Crocosphaera watsonii*. These strains have important phenotypic differences with ecological implications, such as the possible absence of phosphorus scavenging genes in some strains (Dyhrman and Haley, 2006), and differences in cell size, temperature growth optima, exopolysaccharide (EPS) production, and N_2 fixation rates (Webb et al., 2009). It is likely that these differences affect the way each phenotype interacts with the surrounding environment. For example, EPS production may alter cell sinking rates, and has also been shown to have cell-protective properties in some cyanobacteria (Pereira et al., 2009). The two strains described in this study have contrasting phenotypes. *C. watsonii* WH8501, isolated from the South Atlantic in

1984, has a smaller cell size (2-4 µm), narrower temperature range, and does not produce EPS. *C. watsonii* WH0003, isolated from the North Pacific in 2000, has larger cells (4.5 -5.5 µm), produces large amounts of EPS, and has per-cell N_2-fixation rates approximately five times higher than the WH8501 strain (Webb et al., 2009).

Despite differences in phenotype, a high degree of genetic similarity has been observed among *Crocosphaera* strains. For example, when comparing a 950 bp fragment of the typically variable 16S-23S rRNA ITS region, no strain was found to vary at more than five of six total variable single base positions (Webb et al., 2009). Another study examined sequences of five functional genes in seven *Crocosphaera* strains and two large-insert environmental clones (BACs) and found that all strains and BACs shared > 99% nucleotide identity for all gene fragments, suggesting that there is remarkably little DNA mutation accumulation among strains in this genus (Zehr et al., 2007). A metagenomic study supported this finding when similar levels of sequence identity were observed between environmental sequences and the *C. watsonii* WH8501 genome (Hewson et al., 2009). Such observations of *Crocosphaera* genetic conservation are a striking contrast to non-N_2-fixing marine cyanobacteria genera (*Prochlorococcus* and *Synechococcus*) that exhibit a large degree of genomic sequence divergence (Scanlan et al., 2009; Partensky and Garczarek, 2010). Genome-wide analyses of those genera have shown nucleotide sequences of orthologous genes often differ by 20 to 50% even when comparing very closely related species (Zhao and Qin, 2007; Dufresne et al., 2008), The average nucleotide identity of orthologous genes in pair-wise whole genome comparisons of cultivated *Synechococcus* and *Prochlorococcus* species (both within and between genera) was between 50% and 78%, even in comparisons between species with > 96% 16S rRNA identity (Coleman et al., 2006; Zhao and Qin, 2007; Dufresne et al., 2008). In addition, large scale environmental sequencing showed that this degree of sequence variation is also present in natural populations (Rusch et al., 2007).

Observed phenotypic variation among *C. watsonii* strains could be explained by genomic rearrangements such as that reported from alignment of environmental BAC sequences to the WH8501 draft genome (Zehr et al., 2007). That study also observed transposase genes (the genes responsible for genetic movement in transposons) near rearrangements, hinting at a mechanism for genetic mobility. More evidence that transposase genes may be important in *Crocosphaera* spp. was provided shortly after the release of the *C. watsonii* WH8501 draft genome, when the unusually high abundance of transposase genes was recognized, and evidence was found for positive evolutionary selection in a subset of those genes (Mes and Doeleman, 2006). A more recent study showed that this was not a culture-based phenomenon by observing expression of some of those transposase genes in natural *Crocosphaera* populations (Hewson et al., 2009).

Transposons are highly abundant mobile genetic elements that mediate genome shuffling within and among all domains of life (Mahillon et al., 1999; Lander et al., 2001; Feschotte et al., 2002; Waterston et al., 2002; Goodchild et al., 2004; Filee et al., 2007; Touchon and Rocha, 2007; Aziz et al., 2010). The abundance of transposable elements in genomes has been correlated with both genome size and the frequency of horizontal gene transfer (HGT) (Touchon and Rocha, 2007), and IS elements have been observed in prokaryotes at frequencies from zero to over 300 per genome, with proteobacteria and cyanobacteria species containing some of the highest numbers (Kaneko et al., 2007; Touchon and Rocha, 2007; Frangeul et al., 2008; Stucken et al., 2010). However, many cyanobacteria species do not have any recognized transposases in their genomes (e.g. *Prochlorococcus*), and a study which examined a small number of cyanobacterial genomes found very low numbers (median = 1 per genome) even in the genomes which had transposases (Touchon and Rocha, 2007). More recently, researchers found high abundances of transposases in the deep oceans, suggesting they play an important role in microbial communities in a variety of marine environments (Konstantinidis et al., 2009).

The aim of this study was to compare the genomes of two *Crocosphaera* strains (*C. watsonii* WH8501 and *C. watsonii* WH0003, referred to hereafter as WH8501 and WH0003 respectively) in order to answer the following questions: 1) Is the lack of DNA sequence divergence found in previous studies generalized across the entire genome? and 2) are there strain-specific regions in each genome that can explain the phenotypic differences between strains?

Materials and Methods:

WH0003 genomic DNA amplification and 454 sequencing

A non-axenic culture of *Crocosphaera watsonii* WH0003 was grown in nitrogen-free SO medium (Waterbury et al., 1986; Waterbury et al., 1988) in polycarbonate tissue culture flasks with a 0.2 µm pore-size vent cap (Corning Inc., Corning, NY, USA) at 26°C under a 12:12 hour light/dark cycle. Because the cells cannot be directly separated from their EPS matrix, DNA could not be extracted from cultured cells using standard methods. Instead, an aliquot of densely grown cells was subjected to 60 seconds of bead beating on a Mini-Beadbeater-96 (Biospec Products, Bartlesville, OK) with a mixture of 0.5 mm and 0.1 mm beads to physically separate a portion of the cells from their EPS. After bead-beating, the resulting mixture of cells and EPS was passed through a 10 µm swinex filter prior to being sorted using the Influx Mariner flow cytometer and cell sorter (Cytopeia Corp, Seattle, WA). The flow rate was adjusted to allow approximately 2,000 events per second during sorting. Replicates of 5,000 cells each were sorted into 1.5 ml microcentrifuge tubes containing 150 µl of TE buffer, and stored at -80°C.

After freezing, cells were thawed and pelleted at 14,000 rpm (21,000 xg) for approximately 40 minutes and the supernatant was discarded. Cells were resuspended in 7 µl of GenomiPhi V2 sample buffer (Amersham Biosciences, Piscataway, NJ), lysed by adding 2 µL of lysis buffer (400 mM KOH + 10 mM EDTA) and incubating at 65°C for 3 min. The lysis was terminated by adding 2 µL of neutralization buffer (600 mM Tris HCl, pH 7.5, 400 mM HCl) and placing the samples on ice. The resulting whole cell lysis was used directly in a 21 μL reaction by adding 8.5 μL of reaction buffer, and 1.5 μL of GenomiPhi V2 enzyme mix (Amersham Biosciences, Piscataway, NJ). Amplification was carried out in a thermal cycler at 30^{o}C for 105 minutes, terminated at 65°C for 10 minutes, followed by temporary storage at 4°C, and long-term storage at -20°C. Prior to 454 sequencing, amplified genomic DNA was quantified using Pico Green (Invitrogen Corporation, Carlsbad, CA).

Shotgun library construction was carried out at the UCSC Genome Sequencing Center (http://biomedical.ucsc.edu/GenomeSequencing.html) using sorted cell amplified DNA and sequenced on the Genome Sequencer FLX instrument using Titanium Series protocols according to the manufacturer's specifications (454 Life Sciences, Branford, CT).

Sequence assembly and analysis and ORF identification

The ½ chip 454 sequencing run produced 540,451 reads, with an average length of 418 bp for a total of 225,977,489 bp (~37x coverage of the genome). All reads were assembled using Version 2.0.00 of the Newbler GS De Novo Assembler program (454 Life Sciences, Branford, CT). The assembly was run via command line interface using the “-nrm”, “-consed” and “-large” flags. All other parameters used were the default values, as described in the manufacturer’s publication, “Genome Sequencer Data Analysis Software Manual”.

The assembly resulted in 1390 contigs, ranging in length from 500 bp to 46,275 bp with a total length of 6,130,298 bp. Each contig was compared to the *C.*

watsonii WH8501 draft genome (GenBank GI #67858163) using nucleotide BLAST (Altschul et al., 1990), and contigs with over 100 bp sequence alignments to WH8501 were assigned to the WH0003 draft genome. There were 899 such contigs with a total length of 5,465,610 bp. The remaining 491 contigs that showed less similarity, or no similarity, to WH8501 were compared to a database of all prokaryotic proteins using BLASTx (translated nucleotide query vs protein DB), and divided according to the taxonomy of their best BLAST alignments. There were 227 contigs (totaling 424,894 bp) which were most similar to known cyanobacterial sequences. Those contigs were labeled as "probable" WH0003 genome sequence. The remaining 260 contigs (237,640 bp) showed no homology to known cyanobacteria, and were discarded from further analysis.

ORFs in the WH0003 draft genome sequence were identified and annotated using RAST (Aziz et al., 2008). The concatenated proxy genome was the input sequence, and 5,693 features were predicted. Those features were annotated as follows; 3 rRNA sequences (in a single operon), 39 tRNA sequences (71 to 87 bp each), and 5,651 ORFs. In order to correct for the fact that the genome was artificially concatenated, 618 ORFs which were predicted by RAST to read across the ends of contigs were broken at the contig ends, and manually re-annotated using BLAST results of the resulting broken ORFs. This process produced 762 non-broken ORFs (from the original 618 contig-spanning sequences), resulting in a total of 5,795 ORFs in the WH0003 draft genome (4,553,866 bp or 83.3% coding sequence). For the WH8501 genome, the existing GenBank locations and functional annotations for all 5,958 ORFS were used, except for 1,211 ORFs that were identified as transposase genes and subsequently re-annotated with their corresponding IS family assignments (see methods and Table S1).

The WH0003 genome sequences and annotations are publicly available in GenBank (http://www.ncbi.nlm.nih.gov/). The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AESD00000000. The version described in this paper is the first version, AESD01000000. The 899 contigs confidently assigned to the WH0003 have accession numbers AESD01000001-AESD01000899, and the additional 227 contigs that are “probable” WH0003 sequences have accession numbers AESD01000900- AESD01001126.

Transposase annotation*

The observation of highly repetitive ORFs in the WH8501 genome, most of which were annotated as hypothetical proteins, led to a reassessment of the functions of those ORFs. Using a nucleotide sequence identity cutoff of 98%, ORFs were placed into isoform groups, and the number of copies of each isoform was tabulated (Table S2). Amino acid sequences for a representative of each isoform were used as query sequences in a protein BLAST (BLASTp) against all prokaryotic proteins. Sequences were annotated as transposase genes using the conservative criteria of a total identity [tID = percent identity X percent of the ORF length aligned] of more than 50% to known transposases. Those ORFs, as well as any originally annotated transposase genes not represented in the isoform groups, were assigned to IS families according to sequence similarity to known families using the BLAST tool on the ISfinder website (http://www-is.biotoul.fr/is.html) with default parameters (Siguier et al., 2006). All of the re-annotated ORFs are listed in Table S1. A similar analysis was carried out on the WH0003 genome. ORFs initially annotated as hypothetical or unknown proteins were compared (BLASTp) to all prokaryotic proteins. Sequences with alignments to known transposases were annotated based on total identity (tID) as follows; >50% tID: annotated as "transposase", 35-50% tID: annotated as "similar to transposase", 10-35% tID: annotated as "possible transposase". ORFs in all of those categories, as well as those annotated as transposases by the RAST automated annotation were assigned to IS families according to sequence similarity to known families using the ISfinder BLAST tool (http://www-is.biotoul.fr/is.html) (Siguier et al., 2006).

** See appendix 1 for additional analysis of WH8501 and WH0003 genomic transposase ORFs*

Genome comparisons

To assist in visualization of genome-wide comparison between the two strains, proxy genome sequences were created by concatenating the draft genome contigs into a single sequence. For *C. watsonii* WH8501, all 323 contigs were placed in the same order in which they are listed in GenBank, roughly in order of descending contig length. The WH0003 contigs were ordered according to the location of their best BLAST alignment to the WH8501 proxy genome. The resulting two proxy genomes were aligned and visualized using the WebACT (http://www.webact.org/WebACT/home) version of the Artemis Comparison Tool (Carver et al., 2005).

Nucleotide sequences for intergenic spaces (IGSs) over 50 bp and all ORFs from each strain were used as query sequences in BLASTn comparisons against the proxy genome of the other strain. The percent identity of the best BLAST alignment for each sequence (for sequences with alignments ≥ 50 bp) were used in Figures 1 and 2. For the taxonomic analysis of the WH0003 ORFs least similar to WH8501, sequences were placed into three bins based on tID of the best BLAST alignment (35-50% tID, 20-35% tID, and <20% tID). Translated amino acid sequences for all sequences in each bin were compared to the NCBI nr protein database using BLASTp. The results of those BLAST comparisons were used to construct the taxonomic distributions (and likely origins) of the ORFs using the MEGAN program (Huson et al., 2007).

Microarray gene expression

Methods for growth, RNA extraction, and microarray design and hybridization of whole genome expression experiments were described in Shi et al. (2010). Briefly, *C. watsonii* WH8501 cultures were grown under a 12:12 hour light/dark cycle, and RNA was extracted at 8 time points (4 in dark and 4 in light). The RNA samples from each time point were hybridized to an oligonucleotide array designed from the WH8501 draft genome (NimbleGen design ID 2007-03-

14_EW_C_watsonii). A total of 320 oligonucleotide probes representing transposase genes in three IS families and one putative transposase family (average of 80 probes per family) were included on the array, which enabled the analysis described in this manuscript. For each gene, the overall mean expression for all 8 time points was calculated, and the relative expression for each time point was calculated relative to that mean.

Results and Discussion

Broad Genome Comparison

The genomes of the two *Crocosphaera* strains (WH8501 and WH0003) were similar in size, %G+C, and number of predicted ORFs. Genomic DNA from the WH0003 strain was sequenced, assembled and analyzed (see methods), resulting in a draft genome of 5.5 Mb in 899 contigs. There were an additional 227 contigs (0.4 Mb) that had no similarity to WH8501, but were identified as “probable” WH0003 genome sequence, based on similarity to other cyanobacterial sequences (Table 1). The total length of WH0003 contigs (5.9 Mb) was similar in size to the previously sequenced 6.2 Mb genome of the WH8501 strain. The WH8501 genome is composed of fewer (323) and longer contigs than that of WH0003, despite the fact that the sequence data was over 35x coverage of the WH0003 genome. This may have been due to the difference between sequencing methods, since other pyrosequencing projects have had similar difficulties assembling genomes without paired-end data (Goldberg et al., 2006; Hofreuter et al., 2006; Rothberg and Leamon, 2008; Tripp et al., 2010). The % G+C of the two genomes was very similar (37.1% for WH8501 and 37.7% for WH0003) and both genomes had just under 6,000 predicted genes (ORFs) (Table 1). Genome sequences of other cyanobacteria have much higher variability in GC content within a single genus. For example, total % G+C ranges from 31% to 51% in completed *Prochlorococcus* genomes and 52% to 66% in *Synechococcus* genomes (Partensky and Garczarek, 2010). Genome size also varies more within these groups than in the two *Crocosphaera* strains (< 6% variation), with completed genome sizes (as listed in NCBI completed genomes http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) ranging from 2.2 Mb to 3.4 Mb (up to 50% variation) in *Synechococcus* and 1.6 Mb to 2.7 Mb (up to 56% variation) in *Prochlorococcus*.

Shared Genome Features

Most coding and non-coding regions of each genome were nearly identical between the two *Crocosphaera* strains. Nucleotide BLAST comparisons of coding sequences (ORFs) and non-coding intergenic spaces (IGSs) between the genomes revealed that over 80% of each genome was >98% identical to the other strain at the DNA sequence level (Figure 1). Below the highest category of sequence identity, the number of sequences in each bin dropped rapidly, and for all BLAST alignments over 50 bp, there were very few (~ 5%) sequences between 92 and 97% identical, and none was less than 78% identical (Figure 1). The finding that most of the two

Crocosphaera strain genomes are nearly identical is consistent with previous results that reported little to no sequence variation among a number of genetic markers targeting functional genes of cultivated strains and natural populations (Zehr et al., 2007). Surveys of the *C. watsonii* nitrogenase gene *nifH* also showed very little, if any, sequence variation in either the Atlantic (Langlois et al., 2005) or the Pacific Oceans (Church et al., 2005a; Church et al., 2005b). Additionally, 16S-23S rRNA ITS sequences for 10 *Crocosphaera* strains varied at fewer than five single base positions in a 950 bp amplicon, and some strains shared 100% identity in this region that is typically variable (Webb et al., 2009). This level of genetic conservation is particularly notable in phenotypically distinct *Crocosphaera* strains that have been isolated from multiple natural populations, over several decades and from multiple ocean basins. The genome conservation observed in the two *Crocosphaera* genomes described in this study and the previously described environmental sequences (Zehr et al., 2007) could be explained using three possible mechanisms. First, the species could have a very effective DNA replication and error correction system, similar to *Deinococcus radiodurans* (Slade et al., 2009). Second, the *Crocosphaera* population could have undergone a relatively recent global selective sweep as a result of a single strain gaining a trait that conferred a significant advantage over all other strains in the species. The third possibility is that a recent bottleneck could have severely reduced the population and resulted in loss of most genetic variation followed by a global re-distribution of the species throughout tropical waters. However, with the limited sequence data sets currently available, it is not possible to establish which of these three best explains the current observations. Future studies of genomic sequence variation within and between additional strains, and among natural *Crocosphaera* populations would help answer this question.

The high degree of nucleotide sequence conservation in *Crocosphaera* strains also contrasts with sequence divergence observed in the sympatric, non-N_2-fixing cyanobacteria *Prochlorococcus* and *Synechococcus* (Rocap et al., 2002; Ernst et al., 2003; Brown and Fuhrman, 2005; Rusch et al., 2007; Partensky and Garczarek, 2010). Pair-wise, whole genome comparisons of *Synechococcus* and *Prochlorococcus* species as well as metagenomic sequencing of environmental samples showed a much higher degree of variation in cultivated strains and in natural populations (Coleman et al., 2006; Rusch et al., 2007; Zhao and Qin, 2007; Dufresne et al., 2008). However, such sequence divergence was not seen in comparisons of environmental samples to the WH8501 genome (Zehr et al., 2007; Hewson et al., 2009), nor in the comparison between the *Crocosphaera* strains described in this manuscript, which suggests that genome sequence diversity among *Crocosphaera* strains is much different than in *Synechococcus* and *Prochlorococcus* taxa.

Strain-specific Genome Features

While much of the genome of *C. watsonii* is highly conserved at the nucleotide sequence level, genetic variation between strains mostly is present as genome rearrangements, insertions and deletions. Alignments of *Prochlorococcus* genomes showed that strain-specific genetic material is often localized to large islands of variation (10 – 90 kb each) that do not have homology to other strains (Coleman et al., 2006). The sequences in each *Crocosphaera* genome that were strain-specific (10 to 15% of coding and non-coding sequences, see Figure 1) were further analyzed to determine their genomic locations and whether they were localized to similar islands of variation. Proxy genome sequences were constructed by concatenating the contigs from each strain into a single sequence (see methods). The BLAST percent identity of each ORF and IGS from one strain were plotted at the position of the best BLAST match on the proxy genome of the other strain to illustrate sequence similarity across the genomes as well as regions where there was little or no sequence identity (Figure 2). The vast majority of the nearly identical coding and non-coding sequences were spread across both proxy genomes, and regions without similarity (i.e. strain-specific regions) occurred mostly in small fragments across both genomes, rather than grouped into large islands. Because the proxy genomes were both constructed from many contigs, it is possible that some larger stretches of strain-specific sequence were not properly arranged, so their full lengths would be unknown, but that cannot be assessed without closed genomes. However, the vast majority of WH0003 contigs contained some regions that were shared with high nucleotide identity to WH8501; and yet, even on those contigs, the strain-specific regions were consistently small and numerous. This suggests that it is not an artifact of the genome status, but that a multitude of insertions, deletions and genomic rearrangements have occurred since the strains diverged.

The WH0003 ORF sequences that were least similar to the WH8501 genome were compared to public sequence databases and were generally most similar to closely related cyanobacteria. Using a combination of sequence identity and length of BLAST alignment, the total percent identity ([% ID] multiplied by [% of sequence length aligned]) to the WH8501 genome was calculated for WH0003 ORF sequences. There were 930 ORFs in theWH0003 genome with less than 50% total ID (tID) to the WH8501 genome. A combined tree of presumed taxonomy (Figure 3) showed that about 15% of these ORFs (142 of the 930 total), with all tID categories proportionately represented, could not be assigned or had no BLAST hits at the MEGAN "MinScore" value of 35 (Huson et al., 2007). Nearly all assigned ORFs (777 of 788) were most similar to known cyanobacteria. Most (34 of 49) of the ORFs taxonomically identified as *C. watsonii* WH8501 were from the two tID categories most similar to the WH8501 genome, and most of the remaining sequences (379 ORFs) were assigned to various *Cyanothece* spp. (a closely related unicellular N_2-fixing cyanobacterium). Only ORFs from the least similar tID category were assigned to any species other than *Crocosphaera* and *Cyanothece*, and most of those were assigned to other cyanobacterial genera (Figure 3). This suggested that most of the WH0003 strain specific ORFs were either horizontally transferred from cyanobacteria, or were ancestral cyanobacterial genes that have been lost from the WH8501 genome. The 11 ORFs that were taxonomically similar to non-cyanobacterial taxa could have been acquired via horizontal gene transfer (HGT), or could be genes that do not have homologues in genomes of other cyanobacteria sequenced to-date. Because there are many transposase genes in the *Crocosphaera* genomes and the abundance of insertion sequences in genomes is positively correlated with the extent of HGT (Touchon and Rocha, 2007), it is not surprising that these genomes show some evidence of HGT.

Most of the strain-specific ORFs in both genomes did not have annotated functions, were transposases, or were redundant with the functions of shared genes, leaving a relatively small number of gene functions that could be correlated with phenotypic divergence. There were 351 ORFs in WH8501 that had no BLAST similarity (i.e. no alignments >50 bp) to WH0003 (Table S3), half of which (176 ORFs) were transposases (also noted with an ** in Table S1) and ~100 more were annotated as hypothetical or unknown function. The majority of these genes showed diel expression patterns (listed in right column of Tabls S3) in a previous microarray study (Shi et al., 2010). The functions of the 71 ORFs with annotated functions (aside from transposases) are listed in the top two sections of Table S3. Most of those had an identical or nearly identical function annotated in the WH0003 genome, suggesting that the function is not missing from the WH0003 genome, but is being performed by an homologous gene. There were only nine WH8501 ORFs with functions not found in the WH0003 genome (listed at the top of Table S3). In contrast, the WH0003 genome had a larger number (609) of ORFs without BLAST similarity to the WH8501 genome (Table S4). The majority (370) of those ORFs had no assigned function, and only 24 were transposases. The functions of the remaining 215 ORFs with non-transposase functions are listed in the top section of Table S4. A significant portion (57) of those 215 ORFs were annotated with functions that had no homologues in the WH8501 genome (using annotated gene descriptions). The observation that the WH8501 genome contains a much smaller number of ORFs without functional homologues than the WH003 genome may be an indication that WH8501 has lost genetic functionality with the accumulation of the highly abundant transposase genes throughout its genome. Based on that observation, as well the larger total number of strain-specific ORFs, it seems likely that the WH0003 strain has a number of genetic capabilities that are not present in the WH8501 strain, and which may help explain the phenotypic differences between the strains.

Examination of the two longest WH0003 strain-specific regions showed that one is probably involved in DNA processing, and the other is involved in EPS biosynthesis, which is distinctive of that strain's phenotype. The largest region unique to the WH0003 genome was 28.5 kb long and was dominated by ORFs annotated as hypothetical or unknown (Table S5). Five of the seven functionally annotated ORFs were related to DNA replication or transcription, and the function of one other ORF (*bmgA*) is a mobilization protein that plays a role in horizontal gene transfer. These predicted ORF functions suggest that the region could provide an aspect of DNA processing not carried out by the WH8501 strain. The second largest (~25kb) strain-specific region of the WH0003 genome contained 23 ORFs, a number of which had annotated functions related to polysaccharide biosynthesis and export (Table 2). Alignment of the two genomes using the flanking shared sequences (6.6 kb from the beginning of the upstream contig and 2.3 kb to the end of the downstream contig) showed that the entire 25kb region has been replaced by a single transposase gene in the WH8501 genome (Figure 4). The %G+C of the 25 kb region is 37.2%, which is very close to the average for the entire WH0003 genome, and most of the highest quality blast alignments for ORFs in the region are to *Cyanothece* spp., a closely related cyanobacterial species. Thus, this region does not appear to be a horizontally transferred addition to the WH0003 genome, but rather a deletion from the WH8501 genome. It is also notable that six of the 15 functionally annotated ORFs in this region had functions that are not found in the WH8501 genome (indicated by an asterisk in Table 2), and eight of the 15 had functions (based on annotation or COG similarity) related to polysaccharide synthesis and export (shown in bold in Table 2 and with arrowheads in Figure 4). Those included three of the five genes proposed as the core of the cyanobacterial EPS pathway (Pereira et al., 2009); specifically, the *wzx* gene (CWATWH0003_3507), the *wza* gene (CWATWH0003_3516) and the *wzc* gene (CWATWH0003_3517). There are no ORFs with sequence homology, or even conserved domain similarity, to any of these three genes in the WH8501 genome. The other two genes in the EPS pathway (*wzb* and *wzy*) have homologues in both strains, which further supports the supposition that WH8501 once had the ability to produce EPS, but lost that functionality through one or more genomic deletion events. Because this region of the WH0003 genome appears to be important in the EPS production that is characteristic of its phenotype, it would be a prime target for physiological studies focused on EPS production, and also as a possible phenotypic marker in future studies of cultivated strains and natural populations.

Transposase Gene Comparisons*

In contrast to the genome similarities between *Crocosphaera* strains, the number of transposase genes per genome showed a six-fold difference with over 1,200 transposases identified in the WH8501genome, and just over 200 identified in the WH0003 genome. Some of the highest numbers of transposases previously found in cyanobacterial genomes were 362 and 469 in two *Microcystis aeruginosa* strains (Kaneko et al., 2007; Frangeul et al., 2008), and 260 in *Trichodesmium erythraeum* IMS101(Stucken et al., 2010). The number of transposase genes in the WH0003 genome (220) was similar in magnitude to those species, but the WH8501 genome contained significantly more than previously reported cyanobacterial genomes (Kaneko et al., 2007; Frangeul et al., 2008). In fact, the 1,211 genes annotated as transposases (see methods) constituted more than 20% of the predicted ORFs in the WH8501 genome, and was far higher than the average of 40 transposases per genome computed for 630 transposase-containing bacterial genomes (Aziz et al., 2010).

To further characterize the transposases in the genomes of both *Crocosphaera* strains, they were assigned to IS families based on sequence similarity using IS finder (Siguier et al., 2006). The resulting IS family distributions showed that WH8501 had many more genes in most families, and the relative proportions of families were quite different between strains (Table 3). The most numerous IS families in the WH8501 genome contained many identical copies of the same sequence, suggesting that their abundance is a result of widespread replication of those genes (see methods and Table S4). For instance, of the 294 ORFs assigned to the IS5 family 283 are isoforms of the same sequence, and for the IS1380 family, 119 of 120 ORFs are isoforms of the same sequence. Such replication may partly explain the disparity in transposase abundance between the genomes, as these genes were not highly replicated in the WH0003 genome (Table 3). However, it was not clear why transposases in the same IS families, and even with similar sequences have not undergone similar levels of replication in the WH0003 strain. Because homologous recombination can be enhanced between multi-copy IS elements (Touchon and Rocha, 2007), it is likely that WH8501 has undergone more genomic recombination compared to WH0003, but that is difficult to assess without finished genome sequences. The differing patterns of IS family abundance and replication between the genomes suggests that there are strain-distinct mechanisms of regulating IS element activity.

The large number of transposase genes in WH8501 may have resulted from genome assembly error or from the strain being maintained in culture for a relatively long time. At the time of genome sequencing, WH8501 had been continuously cultivated for 20 years, but WH0003 was in culture for less than half that time (~9 years) when its genome was sequenced. However, recent metatranscriptome data has shown that some of the transposases found in the WH8501 genome were actively transcribed in natural *Crocosphaera* populations (Hewson et al., 2009). In addition, microarray expression data showed that transposase genes in four IS families in the

WH8501 genome are up- and down-regulated on a daily cycle in culture. The similarity to the pattern observed for the *dnaA* gene (Figure 5), which encodes for a DNA replication initiation protein (Messer, 2002; Zakrzewska-Czerwinska et al., 2007) suggests that transposase expression may be coordinated with DNA replication as has been observed in other organisms (Ton-Hoang et al., 2010). While more work is required to investigate the full range and activity of transposase genes in these strains, the available data suggest that transposase genes are actively expressed in culture and natural populations, and some exhibit a diel expression pattern.

** See appendix 1 for additional analysis of WH8501 and WH0003 genomic transposase ORFs*

Conclusions

The whole-genome comparison of two *Crocosphaera* strains revealed that, although the strains have divergent phenotypes, the vast majority of the two genomes are essentially identical at the nucleotide level, and only a small fraction of ORFs in each genome are strain-specific. ORFs in one of the two largest contiguous strain-specific regions in the WH0003 genome are likely to play a role in EPS biosynthesis, and therefore likely to be important in establishing phenotypic characteristics. Many of the strain-specific ORFs did not have annotated functions, and future discovery of their functions may help further explain the physiological differences between strains. Strain-specific sequences will also be useful for studying genetic and phenotypic variability in natural populations. Both genomes contained an unusually large number of transposase genes, but the WH8501 strain harbored roughly six times the number of these genes compared to the WH0003 strain, and the IS family patterns of the strains were quite different. Overall, these observations support the conclusion that *Crocosphaera* spp. maintain an unusually high degree of genomic sequence conservation, without accumulating significant nucleotide-level mutations, and strains diverge through genomic insertions, deletions and rearrangements.

Acknowledgements

The authors would like to acknowledge funding from NSF grant EF0424599 for the Center for Microbial Oceanography: Research and Education (CMORE) and from the Gordon and Betty Moore Foundation Marine Microbiology Initiative (JPZ). We would also like to thank Brandon Carter and the MEGAMER facility for assistance with flow cytometry and cell sorting and John Waterbury for providing *Crocosphaera* isolates. Pia Moisander, Rachel Foster, Jason Hilton, Kendra Turk and Mary Hogan provided valuable scientific input and technical assistance. We are also grateful to Eric Webb for critical insight and discussion, and Sara Pereira for generous assistance in EPS pathway gene identification.

References

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215**,** 403-410.

Aziz, R., Bartels, D., Best, A., Dejongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., Mcneil, L., Paarmann, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9**,** 75.

Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38**,** 4207-4217.

Bonnet, S., Biegala, I.C., Dutrieux, P., Slemons, L.O., and Capone, D.G. (2009). Nitrogen fixation in the western equatorial Pacific: Rates, diazotrophic cyanobacterial size class distribution, and biogeochemical significance. *Global Biogeochem. Cycles* 23**,** GB3012.

Brown, M.V., and Fuhrman, J.A. (2005). Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquat. Microb. Ecol.* 41**,** 15-23.

Capone, D.G., Zehr, J.P., Paerl, H.W., Bergman, B., and Carpenter, E.J. (1997). *Trichodesmium*: a globally significant marine cyanobacterium. *Science* 276**,** 1221-1229.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.-A., Barrell, B.G., and Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics* 21**,** 3422-3423.

Church, M.J., Bjorkman, K.M., Karl, D.M., Saito, M.A., and Zehr, J.P. (2008). Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol. Oceanogr.* 53**,** 63-77.

Church, M.J., Jenkins, B.D., Karl, D.M., and Zehr, J.P. (2005a). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat. Microb. Ecol.* 38**,** 3-14.

Church, M.J., Short, C.M., Jenkins, B.D., Karl, D.M., and Zehr, J.P. (2005b). Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. *Appl. Environ. Microbiol.* 71**,** 5362-5370.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311**,** 1768-1770.

Dufresne, A., Ostrowski, M., Scanlan, D., Garczarek, L., Mazard, S., Palenik, B., Paulsen, I., De Marsac, N., Wincker, P., Dossat, C., Ferriera, S., Johnson, J., Post, A., Hess, W., and Partensky, F. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology* 9**,** R90.

Dyhrman, S.T., and Haley, S.T. (2006). Phosphorus scavenging in the unicellular marine diazotroph *Crocosphaera watsonii*. *Appl. Environ. Microbiol.* 72**,** 1452-1458.

Ernst, A., Becker, S., Wollenzien, U.I.A., and Postius, C. (2003). Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiology* 149**,** 217-228.

Falcon, L.I., Carpenter, E.J., Cipriano, F., Bergman, B., and Capone, D.G. (2004). N_2 Fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: phylogeny and in situ rates. *Appl. Environ. Microbiol.* 70**,** 765-770.

Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: Where genetics meets genomics. *Nat. Rev. Genet.* 3**,** 329-341.

Filee, J., Siguier, P., and Chandler, M. (2007). Insertion sequence diversity in Archaea. *Microbiol. Mol. Biol. Rev.* 71**,** 121-157.

Frangeul, L., Quillardet, P., Castets, A.-M., Humbert, J.-F., Matthijs, H., Cortez, D., Tolonen, A., Zhang, C.-C., Gribaldo, S., Kehr, J.-C., Zilliges, Y., Ziemert, N., Becker, S., Talla, E., Latifi, A., Billault, A., Lepelletier, A., Dittmann, E., Bouchier, C., and Tandeau De Marsac, N. (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9**,** 274.

Goericke, R., and Welschmeyer, N.A. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Res. (I Oceanogr. Res. Pap.)* 40**,** 2283-2294.

Goldberg, S.M.D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., Li, K., Rogers, Y.-H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M., and Venter, J.C. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 103**,** 11240-11245.

Goodchild, A., Raftery, M., Saunders, N.F.W., Guilhaus, M., and Cavicchioli, R. (2004). Biology of the cold adapted Archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* 3**,** 1164-1176.

Hewson, I., Poretsky, R.S., Beinart, R.A., White, A.E., Shi, T., Bench, S.R., Moisander, P.H., Paerl, R.W., Tripp, H.J., Montoya, J.P., Moran, M.A., and Zehr, J.P. (2009). In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* 3**,** 618-631.

Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L., and Galan, J.E. (2006). Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect. Immun.* 74**,** 4694-4707.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17**,** 377-386.

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A., Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Katoh, M., Nakazaki, N., Nakayama, S., Yamada, M., Tabata, S., and Watanabe, M.M. (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res.* 14**,** 247-256.

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

Karl, D., Michaels, A., Bergman, B., Capone, D., Carpenter, E., Letelier, R., Lipschultz, F., Paerl, H., Sigman, D., and Stal, L. (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* 57/58**,** 47-98.

Kitajima, S., Furuya, K., Hashihama, F., Takeda, S., and Kanda, J. (2009). Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol. Oceanogr.* 54**,** 537-547.

Konstantinidis, K.T., Braff, J., Karl, D.M., and Delong, E.F. (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Appl. Environ. Microbiol.* 75**,** 5345-5355.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., Mcpherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409**,** 860-921.

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

Langlois, R.J., Laroche, J., and Raab, P.A. (2005). Diazotrophic diversity and distribution in the tropical and subtropical Atlantic ocean. *Appl. Environ. Microbiol.* 71**,** 7910-7919.

Liu, H., Nolla, H.A., and Campbell, L. (1997). *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. *Aquat. Microb. Ecol.* 12**,** 39-47.

Mahillon, J., Léonard, C., and Chandler, M. (1999). IS elements as constituents of bacterial genomes. *Res. Microbiol.* 150**,** 675-687.

Mes, T.H.M., and Doeleman, M. (2006). Positive selection on transposase genes of insertion sequences in the *Crocosphaera watsonii* genome. *J. Bacteriol.* 188**,** 7176-7185.

Messer, W. (2002). The bacterial replication initiator DnaA. DnaA and *oriC*, the bacterial mode to initiate DNA replication. *FEMS Microbiol. Rev.* 26**,** 355-374.

Moisander, P.H., Beinart, R.A., Hewson, I., White, A.E., Johnson, K.S., Carlson, C.A., Montoya, J.P., and Zehr, J.P. (2010). Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* 327**,** 1512-1514.

Moisander, P.H., Beinart, R.A., Voss, M., and Zehr, J.P. (2008). Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME J* 2**,** 954-967.

Montoya, J.P., Holl, C.M., Zehr, J.P., Hansen, A., Villareal, T.A., and Capone, D.G. (2004). High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* 430**,** 1027-1031.

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* 63**,** 106-127.

Partensky, F.D.R., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. *Annual Review of Marine Science* 2**,** 305-331.

Pereira, S., Zille, A., Micheletti, E., Moradas-Ferreira, P., Philippis, R.D., and Tamagnini, P. (2009). Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. *FEMS Microbiol. Rev.* 33**,** 917-941.

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* 68**,** 1180-1191.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., Johnson, Z.I., Land, M., Lindell, D., Post, A.F., Regala, W., Shah, M., Shaw, S.L., Steglich, C., Sullivan, M.B., Ting, C.S., Tolonen, A., Webb, E.A., Zinser, E.R., and Chisholm, S.W. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424**,** 1042-1047.

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.* 26**,** 1117-1124.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.-H., Falcã³n, L.I., Souza, V., Bonilla-Rosso, G.N., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., and Venter, J.C. (2007). The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 5**,** e77.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I., and Partensky, F. (2009). Ecological Genomics of Marine Picocyanobacteria. *Microbiol. Mol. Biol. Rev.* 73**,** 249-299.

Scanlan, D.J., and West, N.J. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol. Ecol.* 40**,** 1-12.

Shi, T., Ilikchyan, I., Rabouille, S., and Zehr, J.P. (2010). Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* 4**,** 621-632.

Shiozaki, T., Furuya, K., Kodama, T., Kitajima, S., Takeda, S., Takemura, T., and Kanda, J. (2010). New estimation of N_2 fixation in the western and central Pacific Ocean and its marginal seas. *Global Biogeochem. Cycles* 24**,** GB1015.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34**,** D32-36.

Slade, D., Lindner, A.B., Paul, G., and Radman, M. (2009). Recombination and Replication in DNA Repair of Heavily Irradiated Deinococcus radiodurans. *Cell* 136**,** 1044-1055.

Stucken, K., John, U., Cembella, A., Murillo, A.A., Soto-Liebe, K., Fuentes-Valdes, J.J., Friedel, M., Plominsky, A.M., Vasquez, M., and Glockner, G. (2010). The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS ONE* 5**,** e9235.

Ton-Hoang, B., Pasternak, C., Siguier, P., Guynet, C., Hickman, A.B., Dyda, F., Sommer, S., and Chandler, M. (2010). Single-Stranded DNA Transposition Is Coupled to Host Replication. *Cell* 142**,** 398-408.

Touchon, M., and Rocha, E.P.C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24**,** 969-981.

Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R.A., Desany, B.A., Niazi, F., Affourtit, J.P., and Zehr, J.P. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464**,** 90-94.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H.O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304**,** 66-74.

Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.***,** 71-120.

Waterbury, J.B., Willey, J.M., and Lester Packer and Alexander, N.G. (1988). "Isolation and growth of marine planktonic cyanobacteria," in *Methods Enzymol.*: Academic Press), 100-105.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigó, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420**,** 520-562.

Webb, E.A., Ehrenreich, I.M., Brown, S.L., Valois, F.W., and Waterbury, J.B. (2009). Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. *Environ. Microbiol.* 11**,** 338-348.

Zakrzewska-Czerwinska, J., Jakimowicz, D., Zawilak-Pawlik, A., and Messer, W. (2007). Regulation of the initiation of chromosomal replication in bacteria. *FEMS Microbiol. Rev.* 31**,** 378-387.

Zehr, J.P., Bench, S.R., Mondragon, E.A., Mccarren, J., and Delong, E.F. (2007). Low genomic diversity in tropical oceanic N_2-fixing cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* 104**,** 17807-17812.

Zehr, J.P., Waterbury, J.B., Turner, P.J., Montoya, J.P., Omoregie, E., Steward, G.F., Hansen, A., and Karl, D.M. (2001). Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* 412**,** 635-638.

Zhao, F., and Qin, S. (2007). Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. *Genetica* 129**,** 291-299.

Table 1. Genome assembly information and annotation summary.

Strain (NCBI ID)	**total genome length (bp)**	**# of contigs**	**longest contig**	**Average contig length**	**genome % G+C**	**# of ORFs**	**# of transposases**
WH8501 (GI #67858163)	6,238,156	323	720,107	19,313	37.1	5,958	1,211
WH0003 (AESD01000001-899)	5,465,610	899	46,275	6,079	37.7	5,795	220
probable WH0003 (AESD01000900-1126)	424,894	227	15,256	1,872	37.3	350	9

Table 2. ORFs within WH0003 strain specific genome region. (Functions related to polysaccharide synthesis and export are in bold.)

ORF Locus Tag	ORF start	ORF stop	RAST Annotated function	most similar COG	COG description
ContigC00178 (NCBI accession # AESD01000522)					
CWATWH0003_3496	7204	6563	hypothetical protein	**COG0463**	**Glycosyltransferases involved in cell wall biogenesis**
CWATWH0003_3497	7824	7303	short-chain dehydrogenase/ reductase SDR	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
CWATWH0003_3498	7971	7840	hypothetical protein	COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
*CWATWH0003_3499	8768	8040	**Sugar transferase involved in lipopolysaccharide synthesis**	**COG2148**	**Sugar transferases involved in lipopolysaccharide synthesis**
CWATWH0003_3500	9766	8780	Pyruvate dehydrogenase (lipoamide)	COG0022	Thiamine pyrophosphate-dependent dehydrogenases, E1 component beta subunit
CWATWH0003_3501	10854	9811	Pyruvate dehydrogenase (lipoamide)	COG1071	Thiamine pyrophosphate-dependent dehydrogenases, E1 component alpha subunit
CWATWH0003_3502	11877	10945	putative aldo/keto reductase	COG0667	Predicted oxidoreductase (related to aryl-alcohol dehydrogenases)
*CWATWH0003_3503	12620	11937	macrocin-O-methyltransferase	none	
CWATWH0003_3504	14047	12869	**glycosyl transferase, group 1**	**COG0438**	**Predicted glycosyltransferases**

Contig02285 (NCBI accession # AESD01000523)

*CWATWH0003_3505	1039	26	**WblG protein**	**COG0438**	**Predicted glycosyltransferases**
CWATWH0003_3506	2511	1174	hypothetical protein	none	
*CWATWH0003_3507	3946	2633	**O-antigen translocase**	**COG2244**	**Membrane protein involved in the export of O-antigen and teichoic acid (*wzx*-like)**
CWATWH0003_3508	4998	3946	DegT/DnrJ/EryC1/StrS aminotransferase family protein	COG0399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis
CWATWH0003_3509	5645	5301	hypothetical protein	none	
CWATWH0003_3510	5983	5657	hypothetical protein	none	
CWATWH0003_3511	6492	6325	hypothetical protein	none	
CWATWH0003_3512	7097	6507	acetyltransferase, putative	COG0110	Acetyltransferases (the isoleucine patch superfamily)
CWATWH0003_3513	8034	7090	oxidoreductase domain protein	COG0673	Predicted dehydrogenases and related proteins
*CWATWH0003_3514	9356	8031	UDP-N-acetyl-D-mannosamine 6-dehydrogenase, putative	COG0677	UDP-N-acetyl-D-mannosaminuronate dehydrogenase
CWATWH0003_3515	11421	9538	**polysaccharide biosynthesis protein CapD**	**COG1086**	**Predicted nucleoside-diphosphate sugar epimerases**
*CWATWH0003_3516	12854	11541	**polysaccharide export protein**	**COG1596**	**Periplasmic protein involved in polysaccharide export (*wza*-like)**
CWATWH0003_3517	15181	12926	hypothetical protein	**COG3206**	**Uncharacterized protein involved in exopolysaccharide biosynthesis (*wzc*-like)**
CWATWH0003_3518	15424	17586	hypothetical protein	none	

* - genes without homologous functions in the WH8501 genome

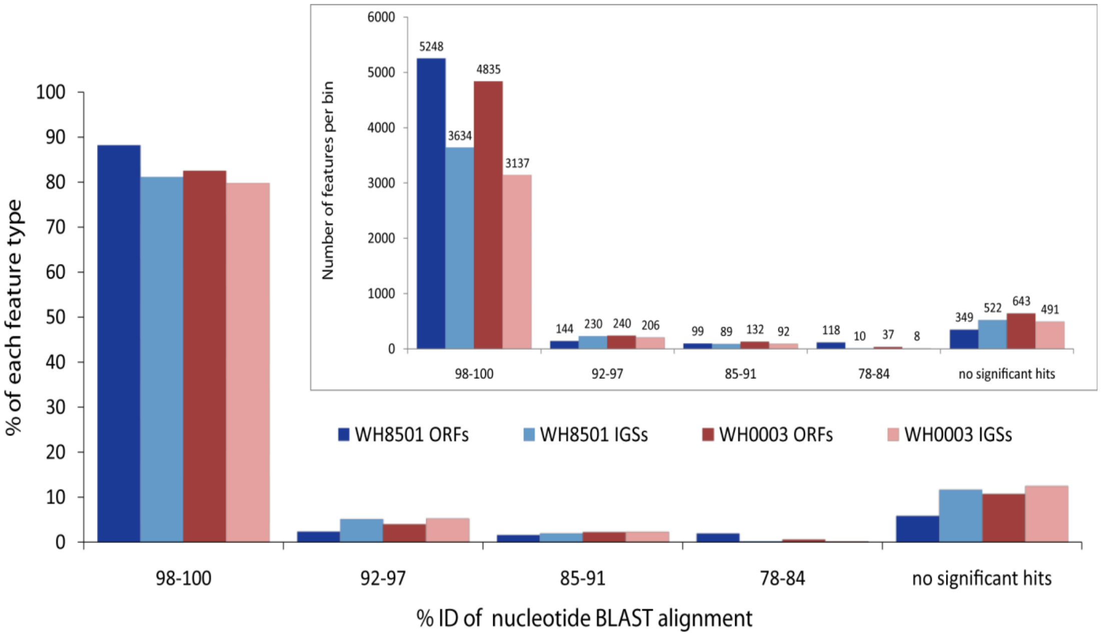
Table 3. Transposase IS family distribution in both genomes

IS family	WH8501	WH0003
IS630	306	5
IS5	294	9
IS1634	152	9
IS1380	120	14
IS200/IS605	83	115
IS66	77	1
ISAzo13	49	2
IS3	41	0
IS4	38	6
IS701	32	1
IS607	14	22
ISAs1	3	3
Tn3	1	6
other	1	4
unknown[a]	210	18
total	1421	215

a – These ORFs were not included in those re-annotated in the WH8501 genome because they could not be assigned with confidence to an IS family.

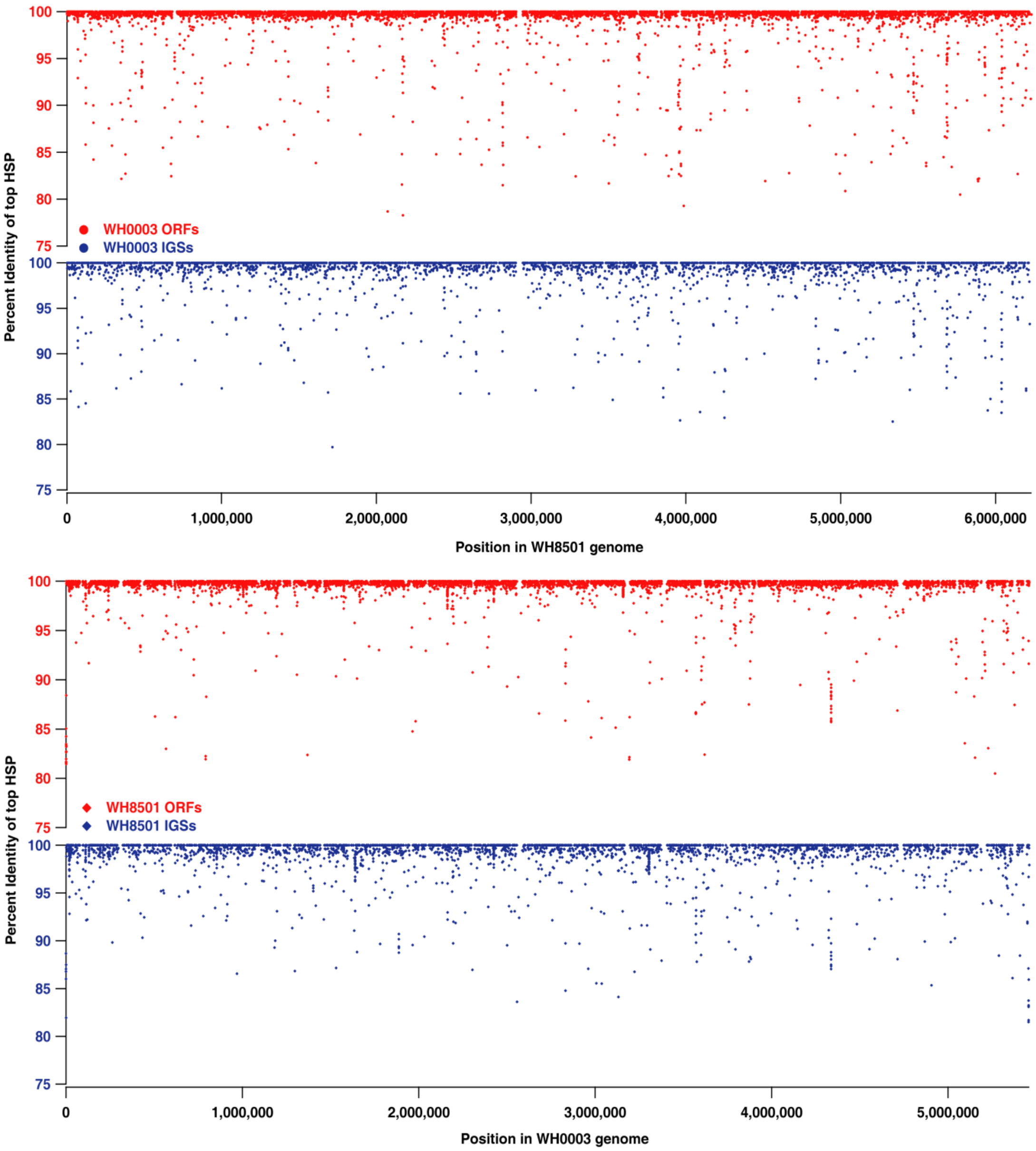
Figures

Figure 1. Nucleotide BLAST similarity of open reading frames (ORFs) and intergenic spaces (IGSs) of the two *C. watsonii* genomes to the opposite strain.



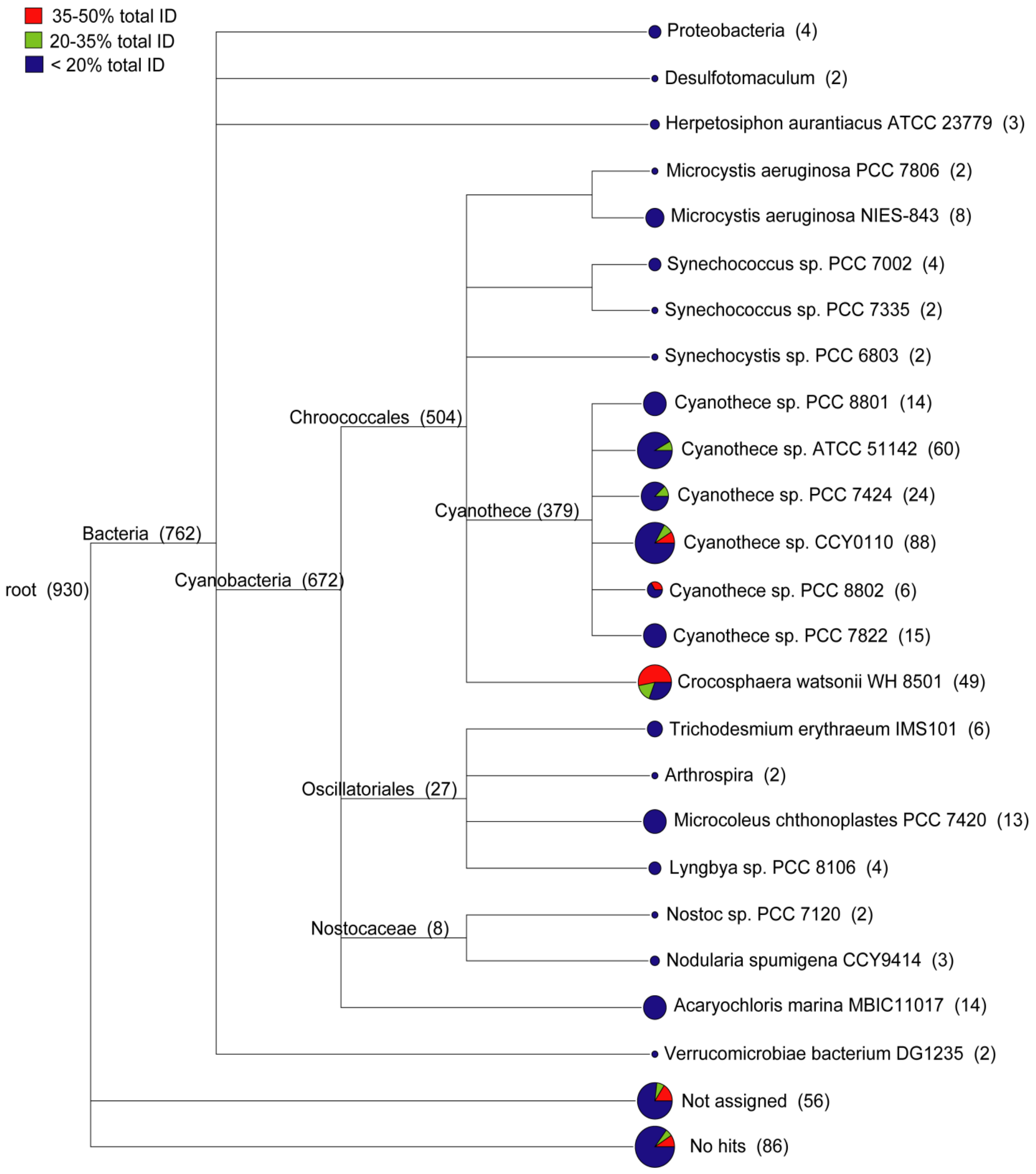
Each feature was binned according to the percent identity of the top BLAST alignment. E-values above 0.001 and alignments shorter than 50 bp were considered not significant. Bars in inset figure show total number of features in each bin, and the main figure bars represent the percent of the total number each feature type.

Figure 2. Comparison of open reading frames and intergenic spaces between two *C. watsonii* genomes.



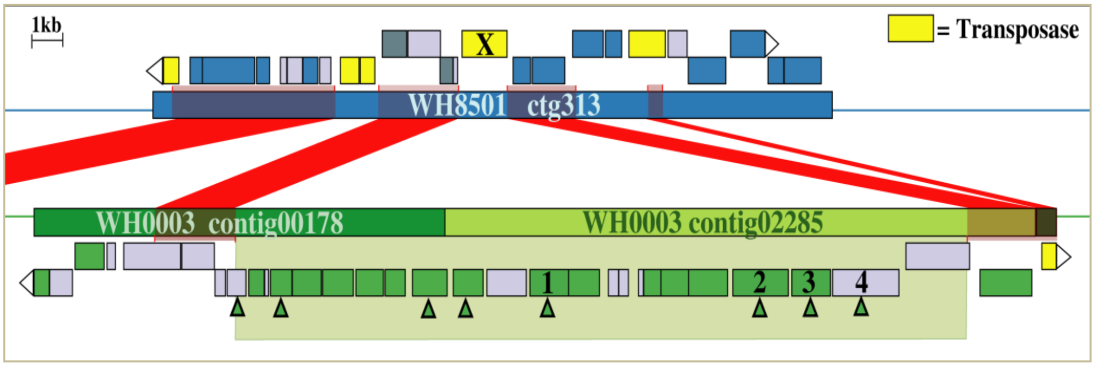
Nucleotide sequence identity of open reading frames (ORFs, in red) and intergenic spaces (IGSs, in blue) between *C. watsonii* genomes. Each point represents a single sequence with the x-coordinate as the subject position of top BLAST hit (i.e. highest scoring pair or "HSP") in the proxy genome sequence, and the y-coordinate as the percent identity of the top BLAST alignment when compared to alternate strain genome.

Figure 3. Taxonomic distribution of 930 ORFs in the WH0003 genome that show little or no sequence similarity to the WH8501 genome.



WH0003 sequences were binned according to the level of sequence similarity to the WH8501 genome across the entire ORF. The number of sequences in each bin were as follows: 751 in the <20% bin; 75 in the 20-35% bin; and 104 in the 35- 50% bin. Pie charts show the relative contributions of the three bins to each branch.

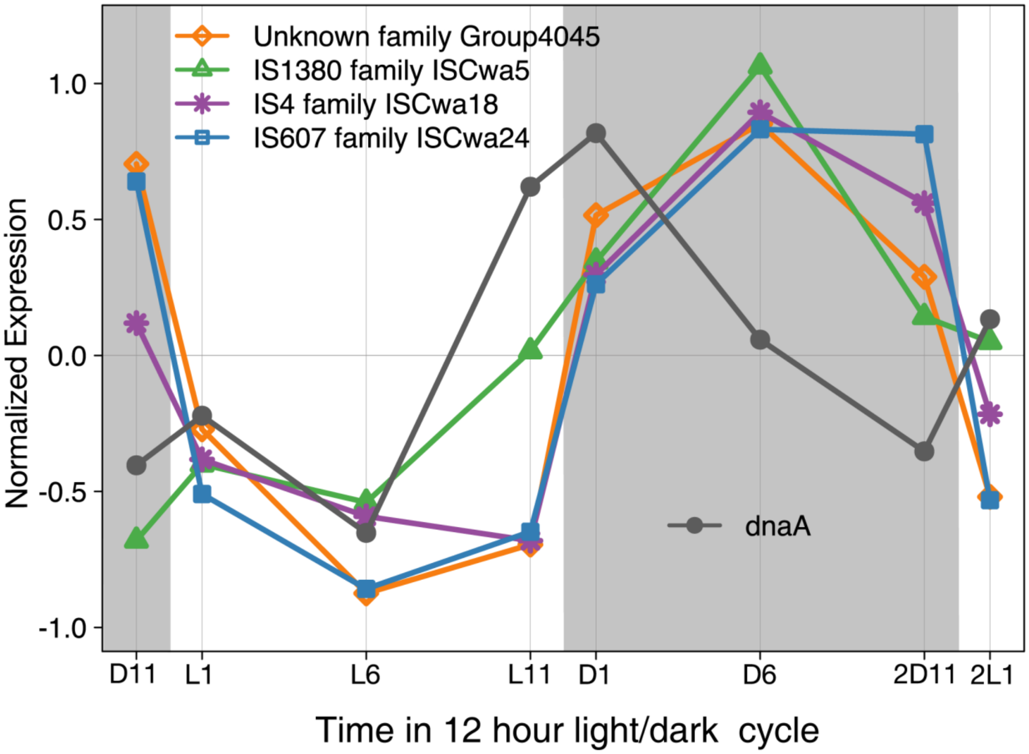
Figure 4. Alignment of WH8501 and WH0003 contigs illustrating a 25kb region specific to the WH0003 genome.



WH8501 contig (top) was aligned with WH0003 contigs (bottom), showing a 25kb region of the WH0003 genome (genes within large green shaded box) that has been replaced by a single transposase gene in the WH8501 genome (marked with an X). Red connecting bars and shading indicate regions of sequence homology. Hypothetical genes are in light gray, transposase genes are yellow and ORFs with other annotated functions are in blue (WH8501) or green (WH0003). ORFs with functions related to polysaccharide synthesis or export are marked with green arrowheads. Descriptions of the numbered genes are listed below. See Table 2 for annotated functions and COG similarities of the 25 contiguous, WH0003-specific ORFs.

1) CWATWH0003_3507: “O-antigen translocase”, similar to *wzx*
2) CWATWH0003_3515: “polysaccharide biosynthesis protein CapD”
3) CWATWH0003_3516: “polysaccharide export protein”, similar to *wza*
4) CWATWH0003_3517: “uncharacterized protein involved in exopolysaccharide biosynthesis”, similar to *wzc*.

Figure 5. Expression of four IS family genes and *dnaA* in *C. watsonii* WH8501.



Expression of four IS family genes and *dnaA* over 26 hour time period with a 12 hour light (L in white)/dark (D in grey) cycle. Expression values for each gene were normalized to average expression for that gene over the entire 26 hour time course, with negative expression values indicating down regulation, and positive values indicating up regulation.

Chapter 3

Investigation of *Crocosphaera watsonii* phenotypes through whole genome comparison of six strains

Abstract

Crocosphaera watsonii, a unicellular nitrogen-fixing cyanobacteria found in oligotrophic oceans, can significantly influence marine biogeochemical cycles, particularly carbon and nitrogen. Isolates of *C. watsonii* can be separated into two phenotypes with environmentally important differences, suggesting that the two types may have different ecological roles and niches. To better understand the evolutionary history and differences in metabolic capabilities among strains and phenotypes, this study compared the genomes of six *C. watsonii* strains, three in each phenotypic group, which had been isolated over decades from a variety of ocean basins. While a substantial portion of each genome was nearly identical to sequences found in the other strains, sequences were identified that were specific to each strain or phenotype, and some of those seemed to account for the phenotypic divergence of the strains. While the small-cell type strains had smaller genomes and a relative loss in genetic capabilities, the large-cell type strains were characterized by larger genomes, some genetic redundancy, and potentially increased adaptations to iron and phosphorus limitation. Consequently, strains with a shared phenotype were evolutionarily more closely related to each other than to those with the opposite phenotype, regardless of where or when they were isolated. An unexpected finding was that the *C. watsonii* type-strain, WH8501, was quite unusual among these genomes, even those with a shared phenotype, indicating it may not be the ideal choice to represent the species. The genome sequences and analyses reported in this study will be important in future studies to test the proposed difference in adaptation of the two phenotypes to nutrient

limitation, and to identify phenotype-specific distributions within natural *Crocosphaera* populations.

Introduction

Crocosphaera watsonii is a unicellular nitrogen (N_2)-fixing cyanobacterium that is widely distributed throughout tropical and sub-tropical oligotrophic oceans. In those regions, the low level of bioavailable nitrogen (N) often limits primary production, and N_2-fixing (i.e. diazotrophic) phytoplankton can be an important source of N for the phytoplankton community. A variety of studies have demonstrated that unicellular diazotrophic cyanobacteria, especially *Crocosphaera* and UCYN-A, are abundant and contribute significant amounts of N in many oligotrophic regions (Church *et al.*, 2005, Church *et al.*, 2008, Langlois *et al.*, 2008, Moisander *et al.*, 2010, Zehr *et al.*, 2001, Falcon *et al.*, 2004, Montoya *et al.*, 2004, Kitajima *et al.*, 2009). *Crocosphaera* strains, all of which are the species *Crocosphaera watsonii*, have been successfully isolated from multiple ocean basins and maintained in culture for many years. Although these strains exhibit phenotypic differences, genetic comparisons have found the vast majority of sequences to be essentially identical at the nucleotide level among cultivated strains and environmental sequences (Zehr *et al.*, 2007, Bench *et al.*, 2011). In the context of such high levels of sequence conservation, *C. watsonii* strains appear to diverge and maintain genetic diversity through genetic rearrangements and by incorporating strain-specific sequences (Zehr et al., 2007, Bench et al., 2011). Large numbers of mobile genetic elements (i.e. transposase genes) in the *C. watsonii* WH8501 genome provide a possible mechanism for such genetic insertions, deletions, and rearrangements (Bench et al., 2011). *Crocosphaera* are distinguished by these characteristics from sympatric non-N_2-fixing marine cyanobacteria, such as *Synechococcus* and *Prochlorococcus*, which generally lack transposase genes and have a high degree of genomic sequence diversity in cultured strains and environmental sequences (Scanlan *et al.*, 2009, Partensky & Garczarek, 2010, Zhao & Qin, 2007, Dufresne *et al.*, 2008, Coleman *et al.*, 2006, Rusch *et al.*, 2007)

Physiological studies of cultivated and natural populations of *C. watsonii* have identified a number of genetic strategies in the species which appear to be adaptions to the oligotrophic environment in which they are found. These include regulation of gene expression, nitrogen fixation rates, and cellular protein content in response to changes in nutrient (e.g. iron and phosphorus) levels and other environmental variables (Hewson *et al.*, 2009, Shi *et al.*, 2010, Saito *et al.*, 2011, Falcon *et al.*, 2005, Dyhrman & Haley, 2006, Webb *et al.*, 2001, Fu *et al.*, 2008, Compaoré & Stal, 2010, Tuit *et al.*, 2004). Currently cultivated *C. watsonii* strains can be divided into two broad phenotypic categories; 1) those that produce large amounts of exopolysaccharide (EPS) and have cell diameters over 4 µm and 2) those that do not produce noticeable EPS, and have cell diameters less than 4 µm (Webb *et al.*, 2009, Sohm *et al.*, 2011). The most striking difference between the two phenotypes in culture is that the large-cell strains produce over 10 times the amount of EPS as the small size strains (Sohm et al., 2011). While the exact reason EPS is produced in *Crocosphaera* sp. is not well understood, EPS production is known to have cell protective properties (Pereira *et al.*, 2009), and can also enhance carbon export from surface waters in the form of marine snow (Passow *et al.*, 2001, Sohm et al., 2011).

A recent genomic comparison of two *Crocosphaera* strains, one of each of the two phenotypes, identified a region in the large cell-type genome that is likely to play an important role in EPS production (Bench et al., 2011). This region contained 25 genes, many of which encoded functions related to EPS biosynthesis, and all of which were absent from the small-cell type genome. The two phenotype groups have additional, ecologically relevant differences in phosphorus scavenging gene content, growth temperature optima, per-cell nitrogen fixation rates, and photosynthetic efficiency (Webb et al., 2009, Sohm et al., 2011, Dyhrman & Haley, 2006).

To better understand the genetic basis of the *C. watsonii* phenotypes, this study compared the genomes of six strains, three in each phenotypic group, isolated over wide spatial and temporal distances. This comparison included examining the evolutionary relationships among strains and identifying genomic features and metabolic capabilities that are unique to strains and phenotypes.

Materials and Methods

Strain growth and genomic DNA isolation and sequencing

The phenotypes, isolation location and genome GenBank accession numbers for *C. watsonii* strains described in this study are listed in Table 1. All strains were grown in nitrogen-free SO medium (Waterbury *et al.*, 1986, Waterbury *et al.*, 1988) in polycarbonate tissue culture flasks with a 0.2 μm pore-size vent cap (Corning Inc., Corning, NY, USA) at 26°C under a 12:12 hour light/dark cycle. The genome of the WH8501 strain was sequenced by the Joint Genome Institute (JGI) and the resulting publicly available sequence (accession number in Table 1) was used for comparisons in this study. The WH0003 strain's genome was sequenced prior to this study by the authors of this study, with detailed methods described in (Bench et al., 2011). Briefly, a non-axenic culture was subjected to bead-beating to detach cells in from their extracellular matrix (ECM), and subsequently, cells were sorted using florescence activated cell sorting (FACS). The genomic DNA from the sorted cells was amplified using the GenomiPhi V2 amplification kit (Amersham Biosciences, Piscataway, NJ).

Genomic DNA for the four additional strains described in this study was obtained using the same methods as the WH0003 strain ((Bench et al., 2011)), with the following modifications: the WH8502 and WH0401 strains do not produce large amounts of ECM, so they were sorted without bead beating, and instead of GenomiPhi, the amplification kit used for all four strains was the REPLI-g Midi kit (Qiagen, Valencia, CA). The REPLI-g amplification method was based on the protocol provided by Qiagen for "small numbers of cells or single cells" (http://www.qiagen.com/products/genomicdnastabilizationpurification/replig/repligminimidikits.aspx#Tabs=t2). Specifically, sorted cells were spun at 14,000 rpm (20,800xg) for 40 minutes and the supernatant was discarded. Pelleted cells were resuspended in 2.5 µl of PBS followed by 3.5 µl of Buffer D2 (see Qiagen protocol above), and lysed in a 65C water bath for 5 minutes. The lysed cells were placed on ice, and lysis was terminated by adding 3.5 µl of Stop Solution. Amplification was immediately carried out in 50µl reactions, which contained the cell-lysis mix plus 1µl of REPLI-g Midi DNA Polymerase, 29µl of REPLI-g Midi reaction buffer, and 10 µl of RT-PCR grade H_20.

Prior to 454 sequencing, amplified genomic DNA was quantified using Pico Green (Invitrogen Corporation, Carlsbad, CA). Using sorted cell amplified DNA, shotgun libraries for each strain were constructed and sequenced by the UCSC Genome Sequencing Center (http://biomedical.ucsc.edu/GenomeSequencing.html) on the Genome Sequencer FLX instrument using Titanium Series protocols according to the manufacturer's specifications (454 Life Sciences, Branford, CT).

Genome assembly and annotation

For the four strains sequenced as part of this study, an average of 363,200 reads were generated per genome, with an average read length of 374 bp, there were approximately 135,900 kb of sequence data for each genome (~23x to 30x coverage of each genome, depending on genome size). The reads for each strain were assembled separately using Version 2.0.00 of the Newbler GS De Novo Assembler program (454 Life Sciences, Branford, CT). The assembly was run via command line interface using the "-nrm", "-consed" and "-large" flags. All other parameters used were the default values, as described in the manufacturer's publication, "Genome Sequencer Data Analysis Software Manual".

ORFs in each of the contig sequences for all four draft genomes were identified and annotated using RAST (Aziz *et al.*, 2008). A small number of contigs [2 contigs, ~1.3 kb, from the WH0401 genome, and 18 contigs (~14kb) from the

WH0005 genome] were removed from further analysis based on a lack of recognizable coding sequence and/or their lack of any homology to known cyanobacterial sequences. In addition to the annotated ORFs, each genome contained a single rRNA operon and 39 tRNAs. The final number of bases and contigs in each genome, as well as the %G+C and number of annotated ORFs are listed in Table 2. The genome sequences and annotations deposited at DDBJ/EMBL/GenBank are publicly available at http://www.ncbi.nlm.nih.gov/ using the accession numbers listed in the Table 1.

Transposase genes were identified and assigned to IS families using the BLAST tool on the ISfinder website (http://www-is.biotoul.fr/is.html) with default parameters (Siguier *et al.*, 2006). ORFs with protein BLAST (BLASTp) E-values of $<10^{-3}$ were annotated as transposases, and assigned to the IS family of the top BLAST hit. Also, a small number of ORFs (10 to 18 per genome) were annotated with the transposase function by RAST, but did not have qualifying BLAST hits to the ISfinder database. These were included in all of the transposase analyses, such as genome counts, and IS family tallies. ORFs without a qualifying ISfinder hit and RAST annotation lacking IS family information were listed as "unknown" in the IS family counts. A very similar process was used to identify transposase genes in the WH0003 and WH8501, with one additional pre-analysis step for the WH8501 genome which involved identification and grouping of highly replicated sequences genomes (see methods in (Bench et al., 2011)).

Genome comparisons

Comparisons between all six genomes were based on nucleotide BLAST of ORFs. The ORFs from each genome were used as queries in BLAST comparisons against the other five genomes, and the criteria used to classify an ORF as shared between genomes was > 95% nucleotide identity over at least 70% of the ORF length. These criteria were based on the observation that the nucleotide sequences for shared ORFs were generally >99% identical, and fell off rapidly below 98% (Fig. S1). The same criteria were used to cluster ORFs within a single genome into repeat groups using the cd-hit web server (Huang *et al.*, 2010, Li & Godzik, 2006). To assess similarities across all six genomes, six tables of BLAST results (one for each single genome versus the other five genomes) were merged using custom software according to sequence similarity and binned by the genomes in which the sequence was present. From the original 34,455 ORFs in the six genomes, this process produced a non-redundant set of 11,635 sequences that could represent all ORFs in all six genomes. Each of the 11,635 sequences were grouped based on the pattern of presence or absence of that ORF in the six genomes, and the total number of sequences was calculated for each of the 63 possible pattern-groups.

The counts of sequences within each pattern-group (with the same total number of genomes) were analyzed for non-random distribution by comparing the observed value to the value expected if all groups were equal. For example, there were 1,727 sequences within 15 possible pattern-groups for sequences found in exactly two genomes, and if all groups were equally distributed, the expected value for each category would be 1727/15 or 115.1 sequences per pattern-group. The statistical significance of differences between observed and expected values was assessed using the χ^2 goodness-of-fit test in Minitab (Minitab Inc., State College, PA). In addition, pattern-groups for 2, 3, 4, and 5 genomes were summed into larger groupings based on the strain phenotypes (e.g. sequences found only in large-cell genomes were considered a single group) and the resulting distributions were similarly assessed for statistically significant deviation of observed values from expected values.

Analysis and PCR of specific groups of functional genes

In order to investigate the phylogenetic relationships of the six *C. watsonii* strains, nucleotide sequences from 25 ORFs were concatenated and aligned and used to construct a phylogenetic tree and distance matrix. The 25 genes were chosen using the following criteria: 1) They were present in all 6 strains, 2) they had some variation between strains (i.e. the vast majority of 100% identical ORF sequences could not be used), and 3) they had homologues in the two *Cyanothece* species used as the outgroup (sp. 51142 and CCY0110, which are the two most closely related genomes available, based in 16S rRNA similarity). Because the *C. watsonii* strains are very closely related, nucleotide sequences were compared, rather than translated amino acids. This allowed the analysis to take into account all possible sequence variation, including synonymous third position changes. The sequence IDs for the original 200 sequences (25 ORFs from each of 8 genomes) that were concatenated are listed in the

Table S1. Eight of the 150 *Crocosphaera* ORFs were split into two sequences, either because they were on two contigs, or by an internal stop codon, which probably arose from sequencing error. For these sequences, the two sequences are listed together in a single cell, in the order in which they were aligned. The sequences were manually concatenated into a single sequence for each genome, and the resulting 8 sequences were initially aligned using ClustalX v2.0.11 (Larkin *et al.*, 2007, Thompson *et al.*, 1994, Thompson *et al.*, 1997), followed by some manual curation and phylogenetic tree construction in MEGA4 (Tamura *et al.*, 2007). The Neighbor-Joining method, with 1000 bootstrap replicates, was used to construct the phylogeny (Saitou & Nei, 1987, Felsenstein, 1985). Evolutionary distances for the tree and distance matrix were calculated based on the same alignment, using the Jukes-Cantor method in MEGA4 (Jukes & Cantor, 1969, Tamura et al., 2007). For both the phylogeny and distance matrix, all codon positions were included (1st, 2nd, 3rd, and noncoding), and positions containing alignment gaps and missing data were eliminated only in pair wise sequence comparisons (Pair wise deletion option). There were 22,611 positions in the final dataset.

Because of prior observations of differences between strains in photosynthetic efficiency (Sohm et al., 2011), per-cell N_2-fixation rates (Webb et al., 2009), and phosphorus scavenging genes (Dyhrman & Haley, 2006), ORFs with functions related to these processes were compared. All ORFs in the each of the six *C. watsonii* genomes were compared to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and given KEGG orthology (KO) assignments using the single-directional best hit method to assign orthologs via the web interface (http://www.genome.jp/tools/kaas/) of the KEGG Automatic Annotation Server (KAAS) (Moriya *et al.*, 2007). Using the KO description and/or the RAST annotated function, genes with roles in the structure or function of the two photosystems, N_2-fixation, and phosphorous transport/metabolism were identified. The number of forms of each gene was totaled for each *C. watsonii* genome using the same criteria used to create the table of 11,635 sequences described above (>95% ID over >70% of the ORF).

The observation of at least one phenotype-specific form of the *isiA* gene led to a more detailed analysis of those ORFs. The *C. watsonii psbC* and *isiA* genes were used as protein BLAST query sequences against the NCBI nr protein database, and the ten most similar sequences to each were retrieved, followed by removing redundant sequences. The resulting set of protein sequences were aligned along with the *C. watsonii* ORFs (58 sequences total) using the online Multiple Sequence Comparison by Log- Expectation (MUSCLE) tool with default parameters (http://www.ebi.ac.uk/Tools/msa/muscle/). A phylogenetic tree was generated from the resulting alignment using the UPGMA method with 500 bootstrap replicates in MEGA5 (Sneath & Sokal, 1973, Felsenstein, 1985, Tamura *et al.*, 2011). Evolutionary distances, as the number of amino acid substitutions per site, were computed using the Poisson correction method (Zuckerkandl & Pauling, 1965). All ambiguous positions were removed for each sequence pair, and there were a total of 776 positions in the final dataset. In the WH0402 genome, two of the *isiA* genes were split into two adjacent ORFs by a stop codon, with adjacent ORFs homologous to adjacent regions of full length *isiA* sequences, suggesting that the stop codons may have arisen from sequencing errors. In both cases the adjacent ORFs were merged prior to alignment, and are so noted on the phylogenetic tree in Fig. 8. The *psbC* clade was identified by examining the sequence alignment for the ~114 amino acid region (between the 5th and 6th transmembrane domains) of the protein that is known to be absent from IsiA proteins (Laudenbach & Straus, 1988, Bricker, 1990).

Prior to the sequencing of any *Crocosphaera* genome other than WH8501, fosmid libraries were constructed for multiple *C. watsonii* strains. The whole genome sequences, which followed shortly afterward, made detailed analysis of those libraries redundant, so they are not included in this study. However, initial analyses of fosmid end sequences identified a gene in the library from a large-cell strain that was not present in the WH8501 genome. Three different forms of this gene, peptidoglycan-binding LysM:Peptidase M23B (referred to hereafter as *lysM*), were present in the WH8501 genome, and the fosmid end sequence was a fourth form with <90% similarity to the other three forms. Based on that finding, a PCR assay was developed which provided support for the genomic comparisons. One reverse primer PCR primer was designed to a conserved region of the *lysM* gene (complementary to all four gene forms), and individual forward primers were designed for each form using Primer 3 (Rozen & Skaletsky, 1999). The sequences for all primers and PCR product sizes are given in Table S2. PCR reactions were carried out in 50 µl reactions, containing 2 µl of template DNA from 8 cultures (4 large-cell and 4 small cell types).

Final reaction concentrations of reagents were as follows: 1x PCR buffer; 2% DMSO, 0.2mM each of dNTPs; 0.4 µM of each primer, and 2 units of Platinum taq (Invitrogen, Life Technologies, Grand Island, NY). Reactions underwent an initial heating step of 94°C for 90 seconds, then 30 cycles of: 94°C for 30 seconds, 56°C for 60 seconds, 72°C for 150 seconds, followed by a final extension step of 70°C for 5 minutes, and holding at 4°C. PCR reactions with bands of the expected size on an agarose gel were verified by direct sequencing following reaction clean-up with QIAquick PCR Purification kits (Qiagen, Valencia, CA). Sanger sequencing reactions and electrophoresis were completed at the UC Berkeley sequencing center using an Applied Biosystems 3730 DNA Analyzer according to manufacturer's protocols. The PCR results are summarized in Table S3, and for the six strains which now have genome sequence data, corresponding ORF IDs have been added where there was a positive PCR result.

Results

Genome characteristics and sequence replication within genomes

Genome sequence statistics were tabulated for all six strains, and are summarized in Table 2. The large-cell strains had genome sizes of nearly 6 Mb, while two of the small-cell strains had genomes closer to 4.5 Mb. WH8501 was the exception to this pattern, with the largest genome (6.2 Mb) of the six strains. The total number of ORFs per genome correlated closely with genome size, indicating similar average ORF sizes and similar coding percentages for all strains. The %G+C was nearly identical (37.6 – 37.7%) for five of the strains, and the sixth strain was very similar (37.1%). All six genomes contained a single rRNA operon, which were nearly identical among the strains. Within the 869 bp region examined in a previous study (Webb et al., 2009), there were four positions with single nucleotide differences among the six genomes. Three of these were identified by Webb et al. (2009) at alignment positions 179, 324, and 794, and the fourth was at alignment position 514. The difference observed at position 514 was a change from an adenine in five of the genomes, to a guanine in strain WH8502 (a strain which was not included in the Webb (2009) study). The WH8501 genome contained a much larger number of transposase genes than any of the other 6 strains. Aside from WH8501, the large cell strains had slightly higher genomic abundance of transposase genes than the small-cell strains. Two of the small-cell strains (WH8502 and WH0401) had the fewest (~ 100) strain specific ORFs, while WH0402 had the most (over 300), and the remaining three strains (WH8501 and two large-cell strains) had ~200 strain specific ORFs each.

The level of ORF duplication within genomes was assessed by clustering identical sequences into groups of repeated genes. For all genomes, aside from that of WH8501, highly repeated sequences were not found, and most ORFs grouped with only 1 or 2 other sequences (Fig. 1 and Table 3). WH8501 was the only strain with repeat-groups of greater than 10 sequences, which ranged from 14 to 277. Six of these were very large groups of over 100 copies of a sequence in the genome. In addition to more transposase genes and much more genomic sequence replication, the transposase genes in the WH8501 genome also had a very different composition when assigned to IS families (Fig. 2). Transposase genes in the other 5 genomes mostly fell into the IS200/605 and IS607 families, with smaller numbers observed in the IS4 and IS91 families. In contrast, the four most abundant IS families in the WH8501 genome were IS5, IS66, IS630, IS1380, and IS1634. The other 5 strains contained sequences in these families as well, but in very small numbers.

Shared and unique ORFs among the six genomes

The open reading frames (ORFs) of each strain were compared to the other five strains using nucleotide BLAST. If a query sequence was >95% identical over at least 70% of the length of the ORF, it was considered to be present in the reference strain. A high percentage (78 to 89%) of ORFs in each large-cell strain was found in the genomes of the other two large-cell strains, while a smaller percentage (62 to

70%) was found in the genomes of the three small-cell strains (Fig. 3 and Table 4). The WH8501 genome shared the most ORFs (79%) with the WH0401 genome, and much less (67 to 73%) to the other four strains. In contrast, a large percentage (78 to 87%) of ORFs in the genomes of the other two small-cell strains (WH8502 and WH0401) was shared with all five other strains. Reciprocal genome comparisons did not yield the same percentages because of differences in the numbers of ORFs in each genome (Table 4). For example, when the WH8502 ORFs were queried against the WH0003 genome, 80% of them were found. However, only 64% of the WH0003 ORFs were found in the WH8502 genome. This is not surprising given that the WH0003 genome contains 6,145 ORFs, while the WH8502 genome contains only 4,965 ORFs (Table 2).

Based on the BLAST results of each genome against the other five, a single set of sequences was established that represented all ORFs present in all six genomes. Using the same criteria described above (95% identity over at least 70% of the length of the ORF), the 34,455 ORFs from all six genomes were grouped into 11,635 sequences. The presence or absence of each of those sequences in the six genomes is shown in Fig. 4. A total of 3,825 sequences were present in all six genomes, which represented approximately 60% of ORFs in the largest genomes, and up to 80% of the smallest genomes. The nucleotide percent identity for these 3,825 sequences averaged 99.8%. Each genome also contained between ~100 and ~300 sequences that were strain-specific (i.e. absent from all other strains). ORF IDs and functions of those are listed in Table S4. For sequences found in exactly three strains, the largest category contained sequences that were present in only the three large-cell type strains (781 sequences), and the second largest was sequences found in the three small-cell strains (153 sequences) (Fig. 4B). The remaining 909 sequences present in 3 strains fell into 18 different categories with between 10 and 93 sequences in each (see Table S5). The categories with sequences present in exactly two strains showed a similarly skewed pattern, with the largest numbers of sequences in categories segregated by phenotype (Fig. 4C and Table S5). There were 15 two-strain categories containing a total of 1727 sequences. Of those, the six phenotype specific categories contained over 1,250 sequences (751 in only large-cell strains, and 502 in only small-cell strains, or an average of over 200 sequences per category), with only 474 sequences total in the remaining 9 categories (average of 53 sequences per category). Overall, 3825 sequences were shared among all six strains, 2237 sequences were found exclusively in large-cell strains (in 1, 2 or 3 genomes), and 1121 were found only in small-cell strains, with the remaining 4452 present in at least one strain of each phenotype.

The observed tendency of sequences to segregate by phenotype was tested for statistical significance using the χ^2 goodness-of-fit test. Sequences that were shared between at least two genomes, but not present in all genomes were included in the analysis. After binning by the number of genomes in which the sequence was found, the observed number of sequences in each pattern-group (i.e. category) was compared to the expected value if all categories were equal, and the differences between the observed and expected values were used in the statistical tests (Fig. 5). In all cases, the deviation from expected was statistically significant with a P-value of 0.000. The largest difference in observed values above expected values were generally in categories of a single phenotype, particularly in sequences found in the three large-cell genomes. Among sequences present in four genomes, there were four categories with many more sequences than the expected values (Fig. 5). Two of these were categories where a sequence was present in all of the small-cell genomes, plus one large cell genome, and two categories were equally split between phenotypes (i.e. two large-cell and two small-cell). The categories in the five-genome-bin showed the least amount of deviation from expected values. Among those, the smallest category was sequences present in the five genomes not including WH0003, which had an observed value well below the expected value. Overall, in categories where the observed values were far below the expected values, there was no consistent pattern of genomes or phenotypes that were included or excluded (Fig. 5).

A similar statistical analysis was conducted on counts of sequences in categories that were further binned based on strain phenotypes. In that analysis, the expected values were the product of the expected value for a single category (as described above) multiplied by how many categories were binned. For sequences found in 2 or 3 genomes, all bins that included only a single phenotype had observed values much above the expected values, and in mixed phenotype bins the observed values were well below expected values (Fig. 6). In the four-genome-bins, observed values were closer to expected, and the bin that included categories split evenly between phenotypes (i.e. sequences found in 2 genomes of each type) was the only group with a lower observed value than expected. Sequences found in five genomes could be binned into two groups; 3 large-cell and 2 small-cell, or vice versa. Both of those bins had observed values relatively close to the expected values, with one (3-large-cell genomes) slightly above expected values, and one below. As would be expected from differences in observed and expected values, the highest contribution to the χ^2 statistic was from the bin of sequences found only in the three large-cell genomes (5149 of 6509), followed by sequences found on only two large-cell genomes (476 of 6509).

In a previous study of the WH0003 genome, a 25kb region was identified as likely to be critical to EPS production because it contained a number of EPS-biosynthesis genes and was present in WH0003 genome, but absent from the WH8501 genome (Bench et al., 2011). Not surprisingly, most of the ORFs in this region were also absent from the other two small-cell strains, and were present in all large-cell strains (Table 5). Furthermore, in the large-cell strains, 23 of the 24 genes were 100% identical at the nucleotide level over the full lengths of the ORFs.

Phylogenetic analysis and comparison of metabolic capabilities

To assess the evolutionary relationships of the six *C. watsonii* strains, a set of 25 functionally unrelated ORFs were used for sequence alignment and phylogenetic tree construction. The nucleotide sequences for the 25 ORFs from each genome and corresponding ORFs from the two most closely related *Cyanothece* species were concatenated and aligned to generate a phylogenetic tree and distance matrix. As expected, the two *Cyanothece* species clustered together as an outgroup to the six *Crocosphaera* strains (Fig. 7). The six *Crocosphaera* strains clustered into two sub-clades with the three large-cell strains in one clade, and the three small-cell strains in the second clade. Over the entire 22kb alignment, the distances between the *Crocosphaera* strains was very small (Table 6). Within each of the phenotype sub-clades, distances ranged from 0 to 0.009 substitutions per site, and between the two clades distances were between 0.024 and 0.028. The distance between *Crocosphaera* strains and *Cyanothece* sp. was approximately 0.16 substitutions per site.

In addition to the genes coding for EPS-biosynthesis described above, the six *C. watsonii* genomes were explored for the presence and forms of genes involved in N_2-fixation, iron and phosphorus scavenging and metabolism and photosynthesis. All genes related to N_2-fixation that were examined (*nifB, D, E, H, K, N, T, U, V, V, W, X*, and *Z*, and *glnB*) were present in a single copy in all six genomes. Phosphorus scavenging and metabolism genes were less uniformly found in the genomes (Table 7). The *pst* genes were often present in multiple copies, with a range of copy numbers per genome that did not appear to directly correlate with phenotype, except for the presence of more *pstS* copies in the large-cell strains. There was a distinct difference between the phenotypes in the various forms of alkaline phosphatase, with *phoD* present in only large-cell strains, and *phoA* found exclusively in small-cell strains (Table 7). A single gene copy was found in all genomes for most of the other phosphorus-related genes, except for *phnD* where two copies were found in five of the six genomes. Finally, the total number of phosphorus-related genes examined here was higher (30-32) in the large-cell strains than the small-cell strains (19-25).

Among the iron-related genes examined, many did not vary in copy number among genomes (e.g. *fur*, *tonB*, and *exbB/D*), while others showed different patterns of copy numbers among strains (Table 8). Some were variable, but copy numbers did not correlate with phenotype or strain origins, such as *feoA* and *dps*. Others had higher copy numbers in the large-cell strains, including *isiA*, *feoB*, and an iron-binding ABC transporter, and only one gene (*idiB*) appeared to have higher copy numbers in the small-cell strains (Table 8). Similar to the phosphorus-related genes, but with a smaller difference, the total number of iron-related genes was higher in large-cell strains (29-33) than in small-cell strains (25-28). Overall, photosystem genes showed less variability among strains than the phosphorus and iron related genes. Photosystem I (PSI) genes were present as a single copy in all six genomes, with the exception of a few ORFs which were split by a stop codon (likely sequencing error), or onto two contigs (Table 9). In addition, *psaF* and *psaK* were not found in the WH8502 genome, and *psaL* was not found in the WH0005 genome. While most photosystem II (PSII) genes were also present in a single copy in the six genomes, there were some exceptions. These included: *psb28,* which was present in two forms in all six genomes; *psbD*, for which a second form was found in two genomes; multiple genomes missing *psbM* and *psbZ*; and a large range among genomes in the number of forms of *psbA* (Table 9).

Based on the striking pattern of *isiA* gene copy numbers observed in the six genomes, it was singled out for phylogenetic analysis. Because *isiA* shares substantial sequence similarity with *psbC*, and it can often be difficult to differentiate between the two genes based on gene annotation in public databases, both genes were included in the analysis. The *C. watsonii isiA and psbC* genes were compared to public databases, and the amino acid sequences most similar were downloaded and aligned with the *C. watsonii* sequences from all strains for phylogenetic tree construction. The resulting phylogeny showed four clades that corresponded to the four forms of *isiA* genes in the *C. watsonii* genomes, and one clade for the *psbC* genes (Fig. 8). The Clade 1 form of *isiA* has homologues in closely related cyanobacteria (e.g. *Cyanothece*), as is seen with the *psbC* clade, although the *psbC* clade has much shorter branch lengths. In the other three *isiA* clades (3, 4, and 5), the most closely related sequences to *C. watsonii* were from *Trichodesmium erythraeum*. In two of those clades (4 and 5), the only sequences were from *Crocosphaera* and *Trichodesmium*, (i.e. all sequences that were identified in public databases as most similar to the *C. watsonii* genes were more similar to the other forms, so they were placed into other clades). The three *Trichodesmium*-like *isiA* forms were also found in mostly large-cell strains, with only one copy found in small-cell strain for any of the forms. For all clades, there was very high (93 to 100%) bootstrap support for the clustering of *C. watsonii* sequences with *Cyanothece* or *Trichodesmium* sequences.

Discussion

The *C. watsonii* WH8501 genome appears to be unusual among this group of *Crocosphaera* genomes in a number of respects, including genome size and transposase abundance (Table 2), IS family distribution (Fig. 2), and genomic sequence replication (Fig. 1). There is a possibility that some differences stem from the fact that the genomic DNA preparation and sequencing methods for WH8501 were different than the other 5 strains (i.e. no cell sorting, and the use of Sanger sequencing rather than pyrosequencing). But it is not clear how such methodological differences would have led to the observed genomic differences. PCR experiments from 16 separate loci have been based on the WH8501 genomic sequence and none have found unexpected results. Four of those loci are described this study, four were developed for another project by the authors of this study (data not shown), three were described in Dyhrman and Haley (2006), and five were described in Zehr (2007). Furthermore, a whole genome microarray was designed based on the WH8501 genome sequence, and subsequent experiments using cultured cells have not shown any systematic problems that might call into question the validity of the genome sequence. Transposase content (total length of 1211 transposase ORFs = 919,337 bp) accounts for most of the extra ~1.5 Mb in the WH8501 genome compared to the other two small-cell strains. Of the six draft genomes, the WH0402 genome has the largest number of contigs and smallest average contig length and has the highest ratio of ORFs to genome size. As such, ORFs are more likely to be split between two contigs. Any strain-specific ORF that spanned two contigs would be counted twice (once for each part on the two contigs). This was seen in the genome counts of iron-related and photosystem genes (Tables 8 and 9), where WH0402 had multiple genes that were separated into 2 ORFs, both shorter than expected, and annotated with the same function. This type of ORF-splitting may partly explain why the WH0402 genome has the highest number and percentage of strain specific ORFs.

The extremely large number of transposase genes in the WH8501 genome are mostly shared with at least one other strain, with only 71 being strain-specific (Table 2). However, it is not clear why the high level of gene replication observed in WH8501 transposase genes was not observed in any of the other genomes. In addition, the relative abundances of IS families among the WH8501 strain-specific transposases are not proportional to abundances in the whole genome. For example, 38 of the 71 strain-specific transposases are in the IS1380 and IS701 families, but in the genome, these families are less abundant than a number of other families (Fig. 2). In the non-WH8501 strains the two most abundant IS families are IS200/605 and IS607, which are related to each other (Chandler & Mahillon, 2002), followed by the IS4 and IS91 families. Interestingly, three of those four families have the unusual property of lacking associated inverted repeat sequences, while the most abundant families in WH8501 (e.g. IS5, IS66, IS630, IS1380, and IS1634) are more typical insertion sequences with associated inverted repeats (Siguier et al., 2006). The lack of inverted repeats in their most common transposase families may partly explain the five genomes do not have high levels of replication. Because IS elements are known to confer adaptive advantages such as acquisition of new metabolic capabilities and increased genomic plasticity via genomic insertions, deletion and homologous recombination between multi-copy elements (Chandler & Mahillon, 2002, Lysnyansky *et al.*, 2009), it seems that the WH8501 strain should be more adaptive to environmental changes than the other strains. Future work could test this hypothesis through competitive growth experiments conducted under changing physical and/or chemical conditions.

The very high nucleotide percent identity (99.8%) for sequences shared among all six strains illustrates that there has been very little mutation accumulation since strain divergence. However, genome size and reciprocal genomic comparisons (Table 2 and Fig. 3) suggest that the larger genomes of the large-cell strains contain functions that are missing from the small-cell strains. The large-cell strains have roughly 1,000 more ORFs in their genomes than two of the small-cell strains, and the large-cell genomes show higher similarity to each other than to the small-cell genomes (Fig. 3). Because the large-cell specific ORFs do not show similarity to sequences in the small-cell strains, gene duplication cannot readily explain their larger genome sizes. Furthermore, when examined for genomic sequence replication, the large-cell strains do not show a large amount of gene duplication (Fig. 1). In contrast, sequence replication may explain why WH8501 has a much larger genome than the two strains with which it shares a phenotype. The large-cell specific functions may have been acquired after phenotypic divergence, may have been lost from the small-cell strains since divergence, or there may be a combination of both.

The presence/absence patterns of the 11,635 sequences further suggests that the large-cell strains harbor functions that are missing from the other three strains. There were nearly 800 sequences shared among all three large-cell strains, and absent from all small-cell strains (Fig. 4B, functions listed in Table S6). Of those sequences, 65% (501) were hypothetical or unknown, and only 3% (25) were transposases. The remaining 246 sequences were annotated with a wide variety of functions, including a number with functions related to DNA metabolism and modification, such as single stranded binding proteins, and DNA polymerases and primases. In addition, large-cell-specific sequences also include the EPS biosynthesis pathway genes identified in the previous 2-genome comparison (Bench et al., 2011). In fact, 17 of the 24 genes identified in a large deletion from the WH8501 genome are also absent from the other two small-cell strains, but present in all three strains characterized by abundant EPS-production (Table 5).

The dendrogram based on the presence/absence patterns of the 11,635 sequences (Fig. 4A), and the fraction of shared genes between strains both demonstrate that the *C. watsonii* strains which share a phenotype are more closely related than those that share a common origin (ocean basin, or year of isolation - see Table 1). The three large-cell strains share many more ORFs among their group, than they do with any of the small-cell strains (Fig. 3 and Table 4). In addition, the categories with sequences found exclusively in large-cell strains were the most overrepresented and contributed most to the statistical significance of differences between observed and expected counts of sequences (Fig. 5 and Fig. 6). However, these three strains do not share a common isolation history. WH0402 was isolated from the South Atlantic two years after WH0003 and WH0005 were isolated from the North Pacific. Similarly, the three small-cell strains cluster together, yet WH0401 was isolated from the North Atlantic 20 years after WH8501 and WH8502 were isolated from the South Atlantic. The two strains isolated in 2000 (WH0003 and WH0005), and the two isolated in 1984 (WH8501 and WH8502) also cluster together, but this is likely as a result of their shared phenotype. Unfortunately, we do not have genomic sequence for a strain of the opposite phenotype which was co-isolated with either of those two groups to test that theory. However, further evidence for clustering by phenotype is provided by the sequence alignment and resulting distance matrix and phylogenetic tree in Fig.7 and Table 6. With 1000 bootstrap replicates, the *C. watsonii* strains cluster by phenotype with 100% support. The evolutionary distances among the large-cell strains are nearly zero, and less than 0.01 substitutions per site among the small-cell strains, while distances between strains of opposite phenotypes are larger (~0.03).

Examination of metabolic capabilities revealed some differences among strains in genes related to iron and phosphorous metabolism, while very few were observed in photosystem genes, and no differences were observed in N_2-fixation genes. Nitrogen fixation and photosynthesis are key metabolic functions of the species, and duplication of genes required for those pathways could increase the evolutionary fitness of the strain that acquired such a duplication (by providing an alternate copy of the gene that could be used if one copy acquired a deleterious mutation). However, while there are many genes and functions that are replicated within the genomes of the six strains, no replication was observed in genes critical for N_2-fixation or most of the photosystem genes, with the exception of *psbA* and *psb28* (Table 9). In contrast, a number of iron (e.g. *isiA*, *isiB, feoA*, *feoB*, *idiB*) and phosphorus (e.g. *pstA, pstb, pstS*) related genes were present in multiple copies in the genomes, some of which were phenotype specific (Tables 7 and 8). The differences in numbers of individual genes, and the resulting differences in total number of iron- and phosphorus-related genes indicate that the large-cell strains may be better adapted to the low nutrient conditions often found in oligotrophic oceans. As such, it seems that the evolutionary divergence of the strains has resulted in phenotypes with different genetic capabilities for dealing with environmental changes in the two most critical nutrients (Fe and P) for a N_2-fixing photo-autotroph such as *C. watsonii*. Studies that have directly examined the response of *C. watsonii* (in cultures) to changes in Fe and P have observed dramatic diel recycling of iron metalloproteins (e.g. photosynthesis and N_2-fixation proteins) as well as changes in growth, gene expression , and nitrogen fixation rates (Dyhrman & Haley, 2006, Falcon et al., 2005, Compaoré & Stal, 2010, Fu et al., 2008, Shi et al., 2010, Webb et al., 2001, Tuit et al., 2004, Saito et al., 2011). Unfortunately, those studies were carried out almost exclusively on the *C. watsonii* WH8501 type strain, so future experiments with additional strains will be needed need to verify that the phenotypes are adapted differently to low or changing nutrient levels. In addition, metatranscriptomic or metaproteomic studies could be used to verify that natural populations are actively expressing the phenotype-specific genes.

The phylogenetic analysis of the *isiA* and *psbC* genes produced three distinct groups of sequences. The *psbC* sequences formed a clade with 100% bootstrap support (Fig. 8, Clade 2 in shaded in grey box), and shorter branch lengths than the *isiA* clades, indicating less sequence divergence in *psbC*. This is not surprising because PsbC, a chlorophyll binding protein, is a critical component of PSII under significant selective pressure (Chisholm & Williams, 1988). For example, a single amino acid substitution caused a significant decrease in both PSII quantum efficiency (F_v/F_m) and oxygen evolution in the cyanobacterium *Synechocystis* (Ananyev *et al.*, 2005). The iron starvation-induced chlorophyll binding protein (IsiA) is a closely related homologue of PsbC, but the two do not share the same function, as the presence of *isiA* cannot compensate for the loss of *psbC* (Singh & Sherman, 2007) . Rather, at least three different functions have been established for IsiA proteins: 1) chlorophyll-storage during iron-limited conditions, 2) dissipation of light-excitation energy, and 3) a light antennae in PSI (Singh & Sherman, 2007, Sandström *et al.*, 2001, Chauhan *et al.*, 2011). Considering these functions, it is interesting that there are striking differences in presence of *isiA* genes in the six *C. watsonii* genomes, with only one of the four *isiA* forms being found in all six strains (Clade 1 in Fig. 8). It should be noted that WH0401 is not present in Clade 1 because the ORF was truncated at the end of a relatively short (~700 bp) contig, thus the shorter sequence could not be properly aligned. As such, it is possible the full length gene is not present in the WH0401 genome, but given the location of the ORF at the end of a contig, it is more likely missing due to the multi-contig, draft status of the genome. All of the *C. watsonii (*and *Trichodesmium* IMS101) *isiA* genes in Clade 1, including the truncated WH0401 sequence, are adjacent to a downstream ORF that is identical among six strains (100% nucleotide identity), but has inconsistent annotation (labeled as "putative flavodoxin" in Table 8). In the four most recent *C. watsonii* genomes, the ORF is annotated as flavodoxin (*isiB*), which is an iron-free replacement for ferredoxin, the iron-sulfur electron transfer protein important in N_2-fixation and CO_2 fixation (Singh & Sherman, 2007). As observed in the *C. watsonii* genomes, *isiB* is commonly found in a single operon with *isiA* (Singh & Sherman, 2007, Chauhan et al., 2011). However, in the two genomes sequenced and annotated previous to this study (WH8501 and WH0003) the ORF is annotated as a hypothetical protein. In addition, BLAST sequence comparisons showed that all of the most similar sequences in available databases (KEGG and NCBI) are also annotated as hypothetical proteins. So, despite some evidence suggesting it is flavodoxin, without further work (e.g. protein modeling or biochemical studies), it is not possible to say with certainty that this ORF is indeed *isiB*.

The *isiA* genes in Clades 3, 4, and 5 appear to have a different evolutionary history than the Clade 1 genes (Fig. 8). This is illustrated by the relatively long branch length between Clade 1 and the rest of the *isiA* genes, as well as the observation that Clade 1 *isiA* genes are found in a different genomic location than the three other forms which are immediately adjacent, and on a single contig in the genomes of the three large-cell strains. Additionally, only Clade 1 forms of *isiA* are found in the species most closely related to *C. watsonii* by *nifH* and 16S rRNA phylogenies (e.g. *Cyanothece* spp.), while the other three forms are only found in *Trichodesmium* and more distantly related cyanobacteria. The co-located *isiA* genes (Fig. 8, Clades 3, 4, and 5) also have a flavodoxin gene immediately upstream, in contrast to the Clade 1 ORFs that have the putative *isiB* gene downstream. Additionally, despite amino acid sequence divergence of over 20% between the two species, *Trichodesmium* has conserved synteny for the three adjacent *isiA* ORFs, illustrating that this gene order has been maintained since they diverged from their last common ancestor.

Given the apparent ancient evolutionary origin of the adjacent *isiA* genes, it is surprising that the small-cell strains are missing most or all of these genes. However, that observation provides a concrete example of generalized genomic gene loss that was suggested by smaller genomes of the small-strains and by the significant number of genes shared among large-cell strains, but absent from small-cell strains. One mechanism for this gene loss is also found in the WH8501 genome, where the Clade 4 (Fig. 8) *isiA* gene is present, but the Clade 3 *isiA* ORF has been interrupted by a transposase gene, leaving only 195 bp of the ORF that is over 1000 bp in the genomes of the large-cell strains. Furthermore, in the WH8501 genome, the Clade 5 *isiA* gene (downstream of the Clade 3 ORF) is completely absent, as is the *isiB* gene that should be upstream of the Clade 4 gene. This is also an example of the type of gene loss that would result from genetic rearrangements, which would be expected in a genome containing abundant transposase genes. Because IsiA plays an important role in photosynthesis during iron-limited conditions (Chauhan et al., 2011), it is possible that the two to three additional copies of the *isiA* gene in the genomes of the large-cell strains could make them better able to continue photosynthesis even in low iron environments. This also may be a possible explanation for observations of higher photosynthetic efficiency (F_v/F_m) in the large-cell phenotype (Sohm et al., 2011).

Conclusions

The vast majority of genes in each of the six *Crocosphaera* genomes were shared with at least one other strain, many with multiple strains, and a large fraction were shared among all six strains at > 99% nucleotide identity, which was not surprising in light of previous studies that have found a high degree of genetic sequence conservation in the species. The genome of WH8501, which has been the type-strain for the species for decades, was quite surprisingly found to be unique within the small-cell phenotype, and among this group of isolates, in a number of respects, such as a larger genome, much more abundant transpose genes, and much higher levels of gene duplication. This certainly calls into question whether WH8501 should continue to be used as the type-strain for the species in future studies. Using a number of genetic and statistical measures, *C. watsonii* strains with the same phenotype clustered together, while similar clustering was not observed in strains with temporal or spatial proximity of isolation. Despite substantial genetic similarity among the genomes of the six strains, the strain- specific and phenotype- specific genes identified in this comparison seem to provide enough differences to result in phenotypic divergence. The resulting phenotypes are characterized by smaller genomes and more apparent gene loss in the small-cell strains, and larger genomes and more redundancy in genetic and metabolic capabilities in the large-cell strains. Finally, there is some evidence that among the redundant genes are capabilities which may make the large-cell strains better adapted to iron and phosphorus limited environments. The genome sequences analyzed in this study provide direction and important data that can be applied to future studies to test that hypothesis, and others, in isolated *Crocosphaera* strains as well as natural populations.

References

Ananyev, G., Nguyen, T., Putnam-Evans, C. & Dismukes, G. C. 2005. Mutagenesis of CP43-arginine-357 to serine reveals new evidence for (bi)carbonate functioning in the water oxidizing complex of Photosystem II. *Photochemical & Photobiological Sciences* **4**.

Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., et al. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**:75.

Bench, S. R., Ilikchyan, I. N., Tripp, H. J. & Zehr, J. P. 2011. Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features. *Front. Microbio.* **2**.

Bricker, T. 1990. The structure and function of CPa-1 and CPa-2 in Photosystem II. *Photosynthesis Res.* **24**:1-13.

Chandler, M. & Mahillon, J. 2002. Insertion sequences revisited. *In*: Craig, N. L. [Ed.] *Mobile DNA II.* 2 ed. ASM Press, pp. 305-66.

Chauhan, D., Folea, I. M., Jolley, C. C., Kouril, R., Lubner, C. E., Lin, S., Kolber, D., et al. 2011. A novel photosynthetic strategy for adaptation to low-iron aquatic environments. *Biochemistry (Mosc).* **50**:686-92.

Chisholm, D. & Williams, J. G. K. 1988. Nucleotide sequence of psbC, the gene encoding the CP-43 chlorophyll a-binding protein of Photosystem II, in the cyanobacterium Synechocystis 6803. *Plant Mol. Biol.* **10**:293-301.

Church, M. J., Bjorkman, K. M., Karl, D. M., Saito, M. A. & Zehr, J. P. 2008. Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol. Oceanogr.* **53**:63-77.

Church, M. J., Jenkins, B. D., Karl, D. M. & Zehr, J. P. 2005. Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat. Microb. Ecol.* **38**:3-14.

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F. & Chisholm, S. W. 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**:1768-70.

Compaoré, J. & Stal, L. J. 2010. Oxygen and the light–dark cycle of nitrogenase activity in two unicellular cyanobacteria. *Environ. Microbiol.* **12**:54-62.

Dufresne, A., Ostrowski, M., Scanlan, D., Garczarek, L., Mazard, S., Palenik, B., Paulsen, I., et al. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology* **9**:R90.

Dyhrman, S. T. & Haley, S. T. 2006. Phosphorus scavenging in the unicellular marine diazotroph *Crocosphaera watsonii*. *Appl. Environ. Microbiol.* **72**:1452-58.

Falcon, L. I., Carpenter, E. J., Cipriano, F., Bergman, B. & Capone, D. G. 2004. N_2 Fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: phylogeny and in situ rates. *Appl. Environ. Microbiol.* **70**:765-70.

Falcon, L. I., Pluvinage, S. & Carpenter, E. J. 2005. Growth kinetics of marine unicellular N-2-fixing cyanobacterial isolates in continuous culture in relation to phosphorus and temperature. *Mar. Ecol. Prog. Ser.* **285**:3-9.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-91.

Fu, F.-X., Mulholland, M. R., Garcia, N. S., Aaron, B., Bernhardt, P. W., Warner, M. E., SaÃ±udo-Wilhelmy, S. A., et al. 2008. Interactions between changing pCO_2, N_2 fixation, and Fe limitation in the marine unicellular cyanobacterium *Crocosphaera*. *Limnol. Oceanogr.* **53**:2472-84.

Hewson, I., Poretsky, R. S., Beinart, R. A., White, A. E., Shi, T., Bench, S. R., Moisander, P. H., et al. 2009. In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* **3**:618-31.

Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**:680-82.

Jukes, T. H. & Cantor, C. R. 1969. Evolution of protein molecules. *In*: Munro, H. N. [Ed.] *Mammalian protein metabolism III.* Academic Press, New York, pp. 21-132.

Kitajima, S., Furuya, K., Hashihama, F., Takeda, S. & Kanda, J. 2009. Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol. Oceanogr.* **54**:537-47.

Langlois, R. J., Hummer, D. & LaRoche, J. 2008. Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. *Appl. Environ. Microbiol.* **74**:1922-31.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., et al. 2007. Clustal W and clustal X version 2.0. *Bioinformatics* **23**:2947-48.

Laudenbach, D. E. & Straus, N. A. 1988. Characterization of a cyanobacterial iron stress-induced gene similar to *psbC*. *J. Bacteriol.* **170**:5018-26.

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658-59.

Lysnyansky, I., Calcutt, M. J., Ben-Barak, I., Ron, Y., Levisohn, S., Methé, B. A. & Yogev, D. 2009. Molecular characterization of newly identified IS3, IS4 and IS30 insertion sequence-like elements in *Mycoplasma bovis* and their possible roles in genome plasticity. *FEMS Microbiol. Lett.* **294**:172-82.

Moisander, P. H., Beinart, R. A., Hewson, I., White, A. E., Johnson, K. S., Carlson, C. A., Montoya, J. P., et al. 2010. Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327**:1512-14.

Montoya, J. P., Holl, C. M., Zehr, J. P., Hansen, A., Villareal, T. A. & Capone, D. G. 2004. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**:1027-31.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**:W182-W85.

Partensky, F. d. r. & Garczarek, L. 2010. *Prochlorococcus*: advantages and limits of minimalism. *Annual Review of Marine Science* **2**:305-31.

Passow, U., Shipe, R. F., Murray, A., Pak, D. K., Brzezinski, M. A. & Alldredge, A. L. 2001. The origin of transparent exopolymer particles (TEP) and their role in the sedimentation of particulate matter. *Cont. Shelf Res.* **21**:327-46.

Pereira, S., Zille, A., Micheletti, E., Moradas-Ferreira, P., Philippis, R. D. & Tamagnini, P. 2009. Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. *FEMS Microbiol. Rev.* **33**:917-41.

Rozen, S. & Skaletsky, H. 1999. Primer3 on the WWW for general users and for biologist programmers. *In*: Misener, S. & Krawetz, S. A. [Eds.] *Bioinformatics Methods and Protocols.* Humana Press, Totowa, NJ, pp. 365-86.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., et al. 2007. The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* **5**:e77.

Saito, M. A., Bertrand, E. M., Dutkiewicz, S., Bulygin, V. V., Moran, D. M., Monteiro, F. M., Follows, M. J., et al. 2011. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proceedings of the National Academy of Sciences*.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406-25.

Sandström, S., Park, Y.-I., Öquist, G. & Gustafsson, P. 2001. CP43′, the *isiA* gene product, functions as an excitation energy dissipator in the cyanobacterium *Synechococcus* sp. PCC 7942. *Photochem. Photobiol.* **74**:431-37.

Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., Post, A. F., et al. 2009. Ecological Genomics of Marine Picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**:249-99.

Shi, T., Ilikchyan, I., Rabouille, S. & Zehr, J. P. 2010. Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* **4**:621-32.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**:D32-36.

Singh, A. & Sherman, L. 2007. Reflections on the function of IsiA, a cyanobacterial stress-inducible, Chl-binding protein. *Photosynthesis Res.* **93**:17-25.

Sneath, P. H. A. & Sokal, R. R. 1973. *Numerical taxonomy: the principles and practice of numerical classification.* W. H. Freeman, 573

Sohm, J. A., Edwards, B. R., Wilson, B. G. & Webb, E. A. 2011. Constitutive extracellular polysaccharide (EPS) production by specific isolates of *Crocosphaera watsonii*. *Front. Microbio.* **2**.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.* **24**:1596-99.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.* **28**:2731-39.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**:4876-82.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994. CLUSTAL-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-80.

Tuit, C., Waterbury, J. & Ravizza, G. 2004. Diel variation of molybdenum and iron in marine diazotrophic cyanobacteria. *Limnol. Oceanogr.* **49**:978-90.

Waterbury, J. B., Watson, S. W., Valois, F. W. & Franks, D. G. 1986. Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.*:71-120.

Waterbury, J. B., Willey, J. M. & Lester Packer and Alexander, N. G. 1988. Isolation and growth of marine planktonic cyanobacteria. *Methods Enzymol.* Academic Press, pp. 100-05.

Webb, E. A., Ehrenreich, I. M., Brown, S. L., Valois, F. W. & Waterbury, J. B. 2009. Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. *Environ. Microbiol.* **11**:338-48.

Webb, E. A., Moffett, J. W. & Waterbury, J. B. 2001. Iron stress in open-ocean cyanobacteria (*Synechococcus*, *Trichodesmium*, and *Crocosphaera* spp.): Identification of the IdiA protein. *Appl. Environ. Microbiol.* **67**:5444-52.

Zehr, J. P., Bench, S. R., Mondragon, E. A., McCarren, J. & DeLong, E. F. 2007. Low genomic diversity in tropical oceanic N_2-fixing cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **104**:17807-12.

Zehr, J. P., Waterbury, J. B., Turner, P. J., Montoya, J. P., Omoregie, E., Steward, G. F., Hansen, A., et al. 2001. Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* **412**:635-38.

Zhao, F. & Qin, S. 2007. Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. *Genetica* **129**:291-99.

Zuckerkandl, E. & Pauling, L. 1965. Evolutionary divergence and convergence in proteins. *In*: Bryson, V. & Vogel, H. J. [Eds.] *Evolving Genes and Proteins.* Academic Press, New York, pp. 97-166.

Table 1. *Crocosphaera watsonii* strain origins and phenotypes

Strain	**Phenotype[a]**	**Year isolated**	**Ocean basin where isolated**	**Location were isolated**	**Genome Accession #**
WH8501	small-cell	1984	S. Atlantic	28°S, 48°W	AADV00000000.2
WH8502	small-cell	1984	S. Atlantic	26°S, 42°W	TBD
WH0003	*large-cell*	*2000*	*N. Pacific (St. ALOHA)*	*22°N, 158°W*	*AESD01000000*
WH0005	*large-cell*	*2000*	*N. Pacific (St. ALOHA)*	*22°N, 158°W*	*TBD*
WH0401	small-cell	2002	N. equatorial Atlantic	6°N, 49°W	TBD
WH0402	*large-cell*	*2002*	*S. equatorial Atlantic*	*11°S, 32°W*	*TBD*

a: large cell size strains in italics

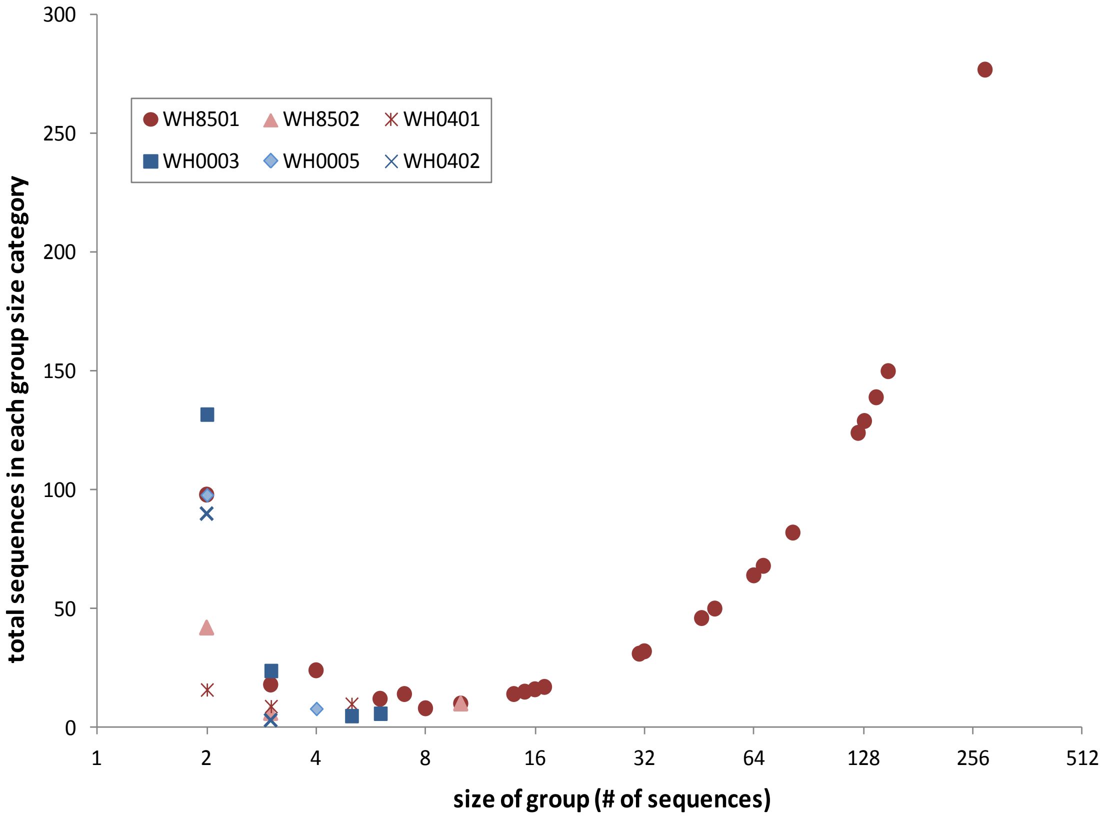
Table 2. Genome sizes and gene content statistics for six *Crocosphaera watsonii* strains

	Strain	total genome length (bp)	Average % G+C	# of ORFs	# of transposase genes	# (and %) of strain-specific ORFs[a]	# of strain-specific transposase genes	Number of contigs[b]
Small-cell strains	WH8501	6,238,156	37.1	5,958	1,211	229 (3.8%)	71	320
	WH8502	4,683,052	37.6	4,965	165	104 (2.1%)	4	869
	WH0401	4,551,017	37.7	4,997	166	132 (2.6%)	4	918
**Large-cell strains*	**WH0003*	*5,892,658*	*37.7*	*6,145*	*223*	*223 (3.6%)*	*9*	*1,130*
	**WH0005*	*5,975,524*	*37.6*	*5,919*	*204*	*167 (2.8%)*	*10*	*1,266*
	**WH0402*	*5,880,358*	*37.7*	*6,471*	*216*	*315 (4.9%)*	*19*	*1,343*

a: An ORF was considered strain-specific if it had no BLAST similarity to ORFs in any other genomes at 95% ID over 70% of the ORF length

b: The number of contigs is given as an indicator of genome completeness.

Figure 1. Abundance of repeated sequences in each *C. watsonii* genome

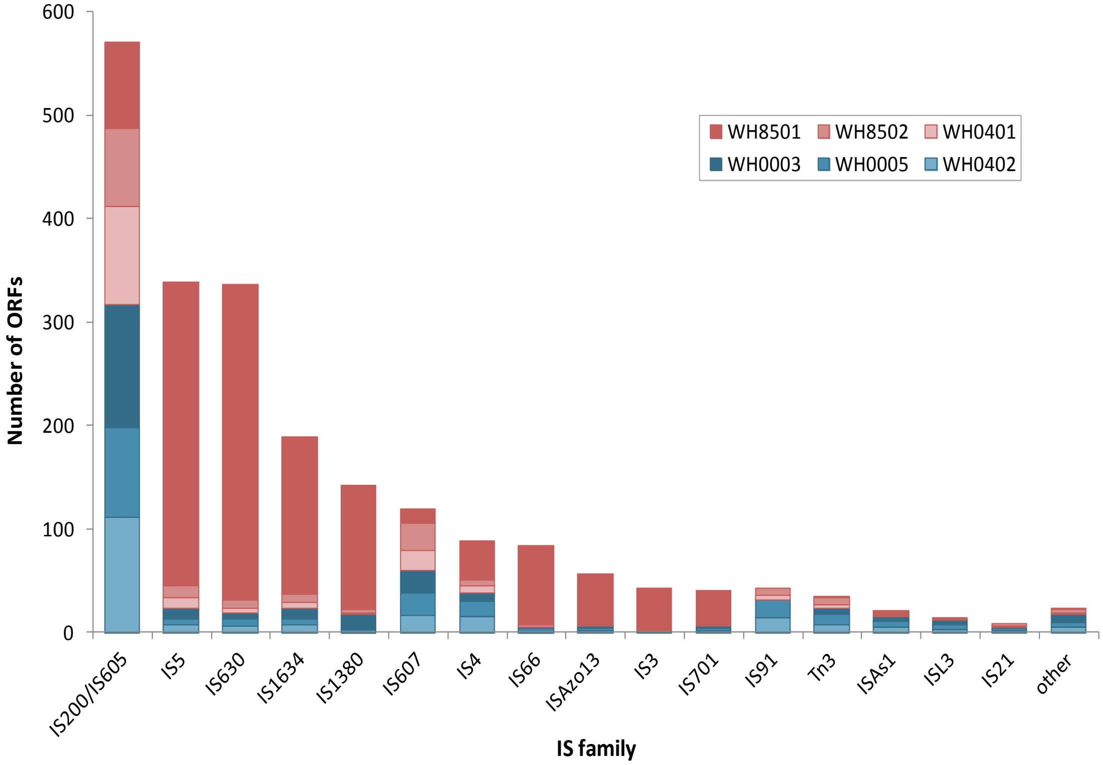


ORFs from each genome that were > 95% identical (over > 70% of the ORF length) were identified and the number of sequences that were repeated within a genome were counted. Groups were plotted as the number of sequences in one group (x-axis), versus the total number of sequences in groups of that size (e.g. 5 groups of 2 sequences each would have a total of 10 sequences, so that point would be plotted at position 2,10). For most of the strains, sequences were not highly repeated, falling into groups of 2 or 3 sequences. WH8501 was the only strain with groups of more than 10 sequences, including six groups with over 100 repeats (see Table 3).

Table 3. Counts of repeated sequences in each *C. watsonii* genome.

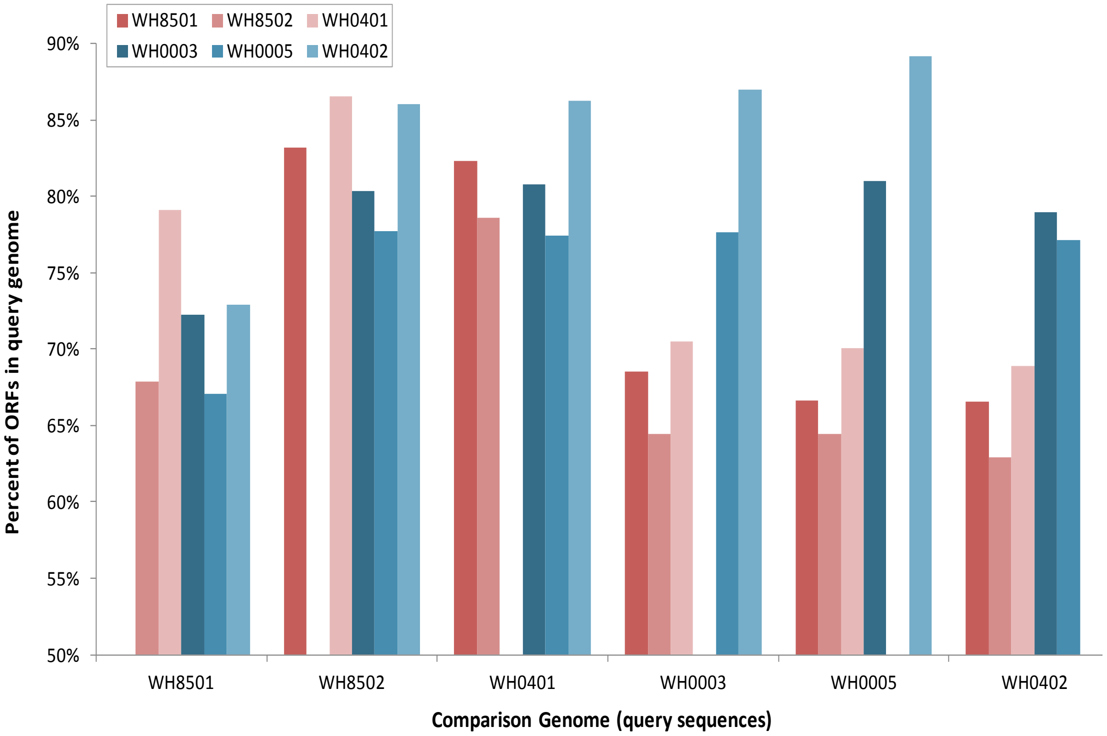
Genome	number of repeats in each group	number of groups in genome	total sequences
WH0003	6	1	6
	5	1	5
	3	8	24
	2	66	132
WH0005	4	2	8
	2	49	98
WH0401	6	1	6
	5	2	10
	3	3	9
	2	8	16
WH0402	3	1	3
	2	45	90
WH8502	10	1	10
	3	2	6
	2	21	42
WH8501	277	1	277
	150	1	150
	139	1	139
	129	1	129
	124	1	124
	82	1	82
	68	1	68
	64	1	64
	50	1	50
	46	1	46
	32	1	32
	31	1	31
	17	1	17
	16	1	16
	15	1	15
	14	1	14
	10	1	10
	8	1	8
	7	2	14
	6	2	12
	4	6	24
	3	6	18
	2	49	98

Figure 2. Abundance and distribution of IS families in each of the six *C. watsonii* genomes



The 16 most abundant IS families are shown, with remaining families combined in the "other" category. Transposase ORFs were assigned to IS families based on sequence similarity to known IS elements using ISfinder.

Figure 3. Percentage of ORFs shared between *C. watsonii* strains



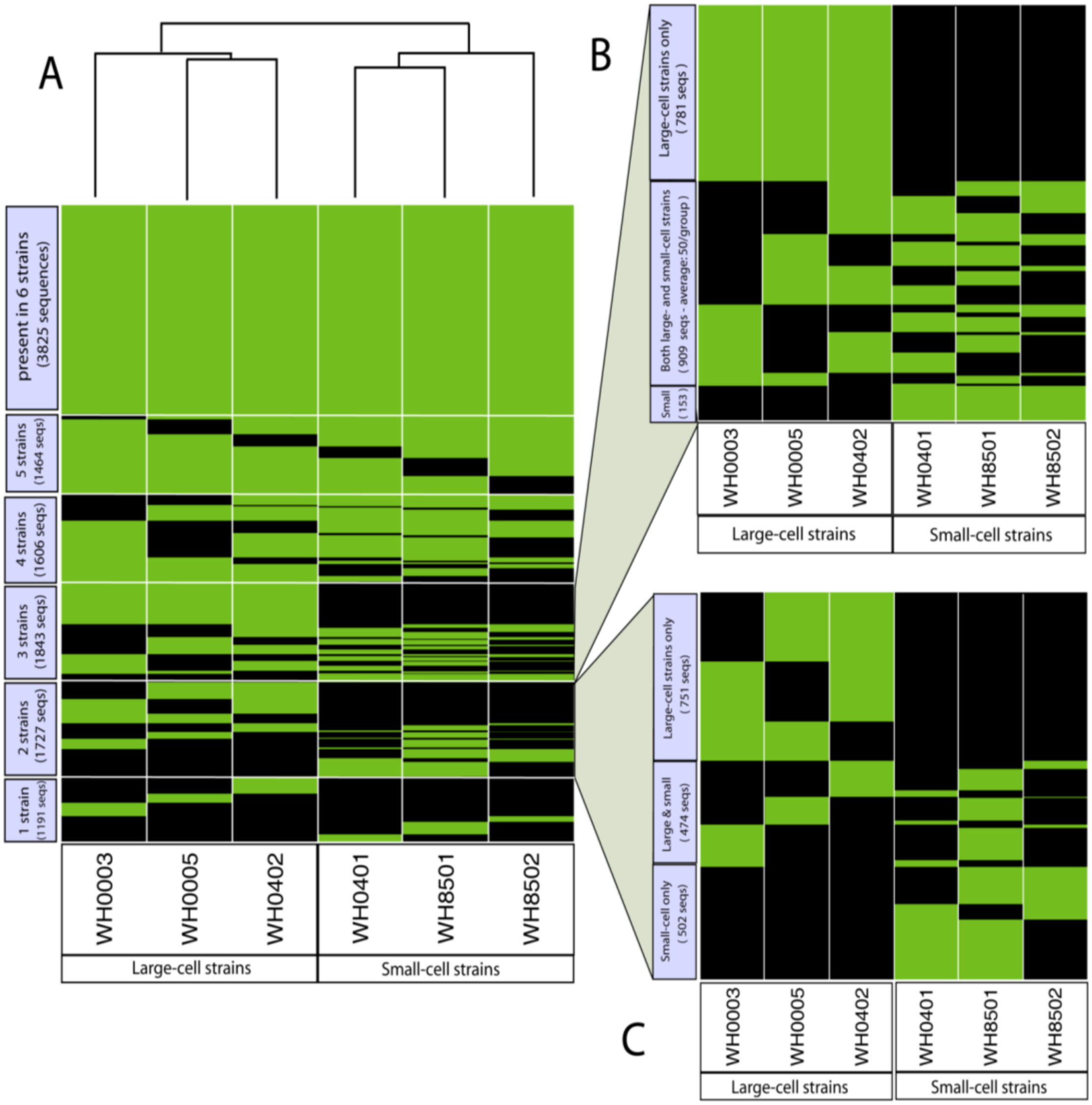
ORFs for each genome were used as query sequences against the other five genomes in nucleotide BLAST searches. Alignments >95% identity over at least 70% of the ORF were totaled and plotted as a percent of the total number of ORFs in the query genome. Small-cell strains are represented by red shades, and large-cell strains by shades of blue.

Table 4. Percentage of ORFs shared between *C. watsonii* strains

		comparison genome (query sequences)					
		WH8501	WH8502	WH0401	WH0003	WH0005	WH0402
Reference genome	WH8501		83.2%	82.3%	68.6%	66.6%	66.5%
	WH8502	67.9%		78.6%	64.4%	64.4%	62.9%
	WH0401	79.1%	86.5%		70.5%	70.1%	68.9%
	WH0003	72.2%	80.4%	80.8%		81.0%	79.0%
	WH0005	67.0%	77.7%	77.4%	77.7%		77.2%
	WH0402	72.9%	86.0%	86.3%	87.0%	89.2%	

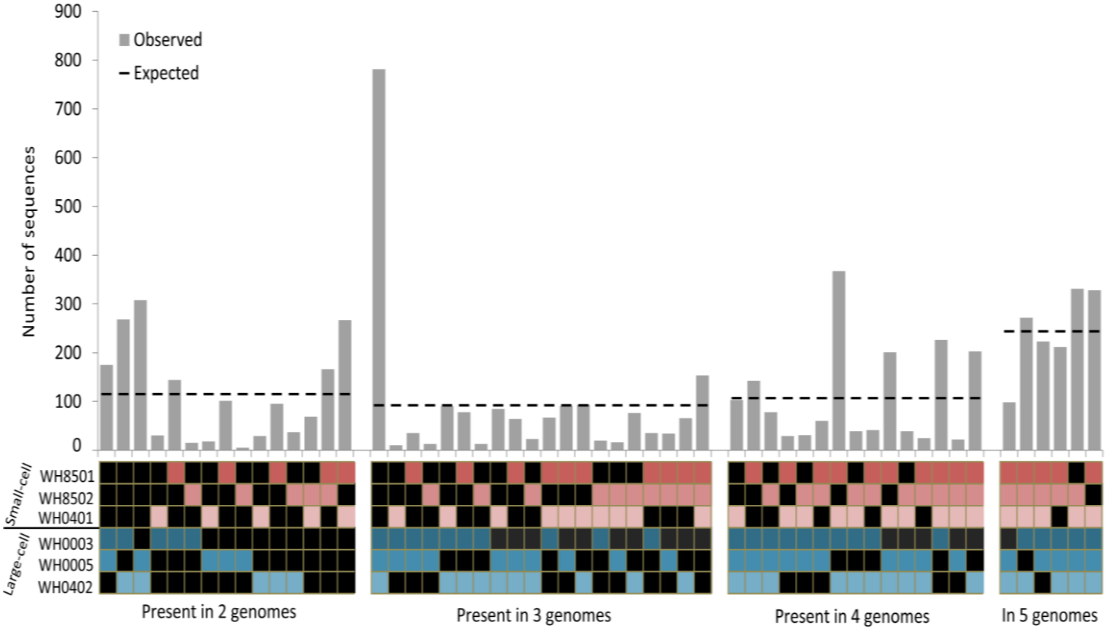
ORFs for each genome were used as query sequences against the other five genomes in nucleotide BLAST searches. Alignments >95% identity over at least 70% of the ORF were counted and are presented as a percent of the total number of ORFs in the query genome. Red shades indicate small-cell strains, and blue shades indicate large-cell strains.

Figure 4. Presence/absence of all ORFs in six *C. watsonii* genomes



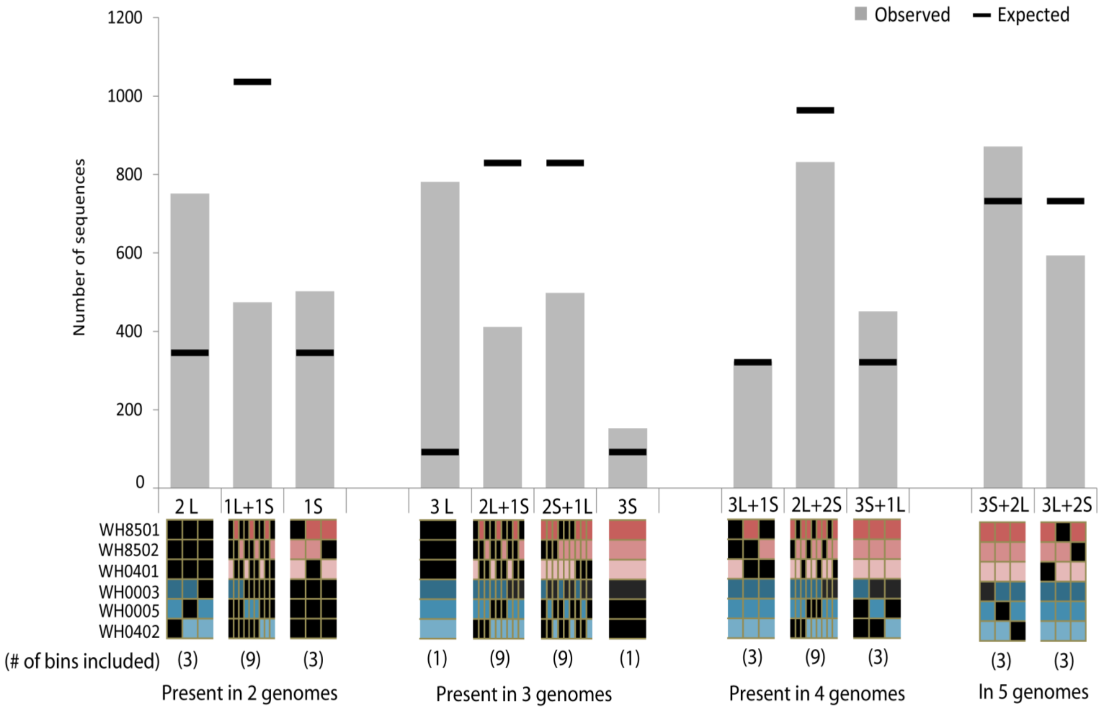
Presence (green) or absence (black) of 11,635 sequences that represent all ORFs in the genomes of six strains (A). Each strain is represented by the column above the strain names, and each row represents one sequence. Rows are grouped by the number of strains in which the sequence is found, and the total number of sequences in each category is listed on the left. The dendogram above the columns is based on the presence/absence pattern for all 11,635 rows. Zoomed in views of the sequences found in 3 strains (B) and 2 strains (C) are shown with total for sub-categories listed on the left.

Figure 5. Counts of ORFs found in genomes of 2, 3, 4 and 5 *C. watsonii* strains



Observed counts of ORFs for each category, indicated by bars, were compared to the expected count for each category, indicated by black lines. Categories were binned by the number of genomes in which a sequence was present, and expected counts were calculated by assuming all categories were equally likely in each bin. The 6-genome presence/ absence pattern for each category is indicated by the boxes below the bars (black= sequence is absent for that genome, colored = present). The $\chi 2$ goodness-of-fit test indicated statistically significant difference between observed and expected counts for all categories (n = 6640, Df = 55, $\chi 2$ = 8469, p = 0.000).

Figure 6. Combined counts of ORFs found in genomes of 2, 3, 4 and 5 *C. watsonii* strains

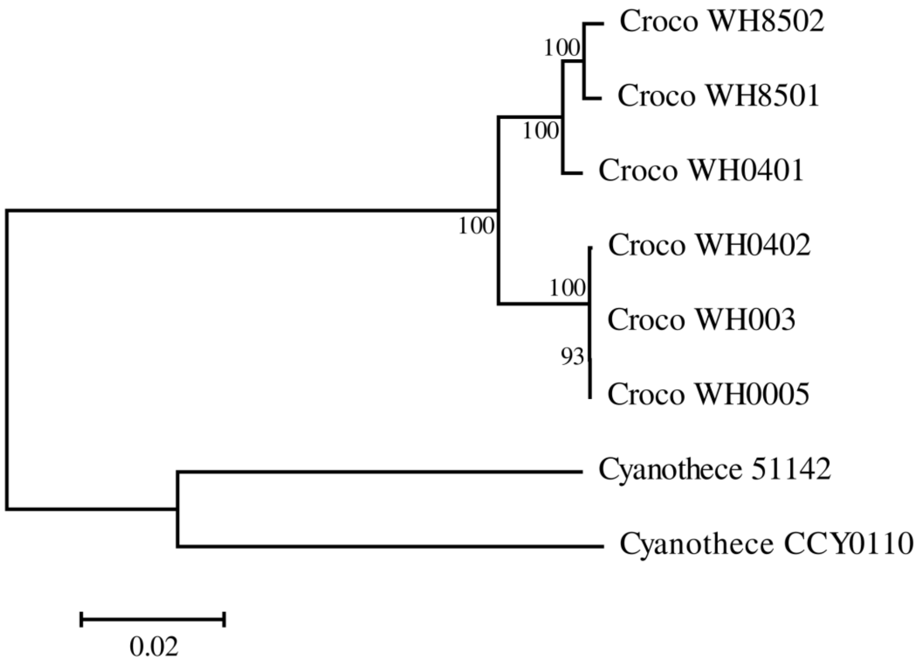


ORF sequences were binned based on the phenotypic category of the genomes in which the sequence was found (using criteria of 95% identity over >70% of the ORF length). The 6-genome presence/ absence pattern for each category is indicated by the boxes below the bars (black= sequence is absent for that genome, colored = present - with blue shades indicating large-cell strains and red shades indicating small cell strains). The total number of sequences in each bin (grey bars) was compared to expected counts (black lines) that were calculated by assuming all categories were equally likely in each bin, and summing the number of binned categories. The $\chi 2$ goodness-of-fit test indicated statistically significant differences between observed and expected counts for all categories (n = 6640 , Df =11, $\chi 2$ = 6509, p = 0.000) .

Table 5. Presence/absence of 24 genes in putative EPS-critical region identified in the WH0003 genome.

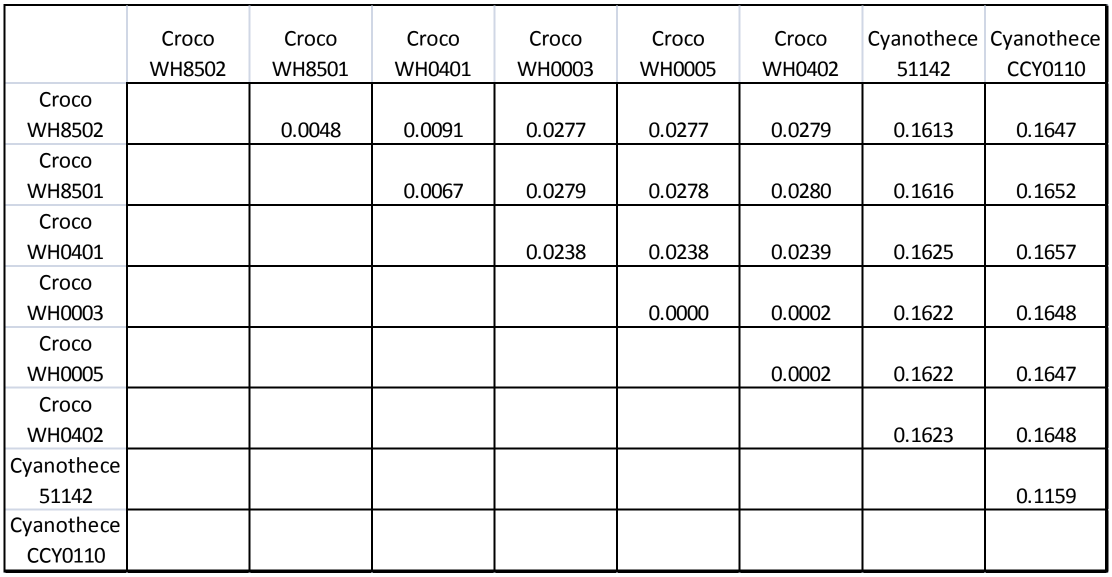
Global %ID							
WH0003	WH0005	WH0402	WH0401	WH8501	WH8502	W0003 ORF ID	Annotated Function
1	1	1	0	0.94	1	CWATWH0003_3496	hypothetical protein
1	1	1	0	0	1	CWATWH0003_3497	short-chain dehydrogenase/reductase SDR
1	1	1	0	0	0.98	CWATWH0003_3498	hypothetical protein
1	1	1	0	0	1	CWATWH0003_3499	Sugar transferase involved in lipopolysaccharide synthesis
1	1	1	0	0	0	CWATWH0003_3502	putative aldo/keto reductase
1	1	1	0	0	0	CWATWH0003_3503	macrocin-O-methyltransferase
1	1	1	0	0	0	CWATWH0003_3504	glycosyl transferase, group 1
1	1	1	0	0	0	CWATWH0003_3505	WblG protein
1	1	1	0	0	0	CWATWH0003_3506	hypothetical protein
1	1	1	0	0	0	CWATWH0003_3507	O-antigen translocase
1	1	1	0	0	0	CWATWH0003_3508	DegT/DnrJ/EryC1/StrS aminotransferase family protein
1	1	1	0	0	0	CWATWH0003_3509	hypothetical protein
1	1	1	0	0	0	CWATWH0003_3510	hypothetical protein
1	1	1	0	0	0	CWATWH0003_3511	hypothetical protein
1	1	1	0	0	0	CWATWH0003_3512	acetyltransferase, putative
1	1	1	0	0	0	CWATWH0003_3513	oxidoreductase domain protein
1	1	1	0	0	0	CWATWH0003_3514	UDP-N-acetyl-D-mannosamine 6-dehydrogenase, putative
1	1	1	0	0	0	CWATWH0003_3515	polysaccharide biosynthesis protein CapD
1	1	1	0	0	0	CWATWH0003_3516	polysaccharide export protein
1	1	1	0	0	0	CWATWH0003_3517	hypothetical protein
1	1	1	0	0	0	CWATWH0003_3518	hypothetical protein
1	1	1	1	0.99	0	CWATWH0003_3519	Animal haem peroxidase
1	1	1	0	0	0	CWATWH0003_3520a1	Transposase, IS200/IS605 family

Figure 7. Phylogenetic relationship of six *C. watsonii* strains and two *Cyanothece* species, based on 25 genes.



Evolutionary relationships were inferred based on a 25kb alignment of 25 concatenated genes using the Neighbor-Joining method and the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The optimal tree with the sum of branch length = 0.24091980 is shown drawn to scale, with branch lengths in units of base substitutions per site.

Table 6. Estimates of Evolutionary Divergence among six *C. watsonii* strains and two Cyanothece species, based on 25 genes.



	Croco WH8502	Croco WH8501	Croco WH0401	Croco WH0003	Croco WH0005	Croco WH0402	Cyanothece 51142	Cyanothece CCY0110
Croco WH8502		0.0048	0.0091	0.0277	0.0277	0.0279	0.1613	0.1647
Croco WH8501			0.0067	0.0279	0.0278	0.0280	0.1616	0.1652
Croco WH0401				0.0238	0.0238	0.0239	0.1625	0.1657
Croco WH0003					0.0000	0.0002	0.1622	0.1648
Croco WH0005						0.0002	0.1622	0.1647
Croco WH0402							0.1623	0.1648
Cyanothece 51142								0.1159
Cyanothece CCY0110								

The numbers of base substitutions per site are shown based on the pairwise analysis of 8 sequences using the Jukes-Cantor method. Each sequence includes the same 25 concatenated genes used to construct the phylogenetic tree above.

Table 7. Counts of phosphorous-related genes in each *C. watsonii* genome

Gene	**KEGG Orthology number**	**Function**	***WH0003***	***WH0005***	***WH0402***	**WH0401**	**WH8501**	**WH8502**
phnD	K02044	phosphonate binding	*2*	*2*	*2*	1	2	2
phoB	K07657	phosphate regulon transcriptional regulator	*1*	*1*	*1*	1	1	1
phoH	K06217	phosphate starvation-inducible protein	*1*	*1*	*1*	1	1	1
phoR	K07636	phosphate regulon sensor	*1*	*1*	*1*	1	1	2
phoU	K02039	phosphate transport system regulator	*1*	*1*	*1*	1	1	1
pstA	K02038	phosphate transport system permease	***3***	***3***	***3***	**1**	**3**	**1**
pstB	K02036	ATP-binding phosphate transport	*3*	*3*	*4*	3	3	3
pstC	K02037	periplasmic phosphate-binding ABC-transporter	***2***	***2***	***2***	**1**	**1**	**2**
pstS	K02040	high-affinity phosphate-binding	***6***	***5***	*7*	**3**	**3**	**4**
phoD	K01113	phosphodiesterase/ alkaline phosphatase D	***2***	***2***	***2***	**0**	**0**	**0**
phoA		alkaline phosphatase	***0***	***0***	***0***	**1**	**1**	**1**
		alkaline phosphatase (non-phoD & non-phoA)	***2***	***4***	***2***	**0**	**0**	**1**
dedA		alkaline phosphatase-like	*1*	*1*	*1*	1	1	1
pitA	K03306	inorganic phosphate transporter	*1*	*1*	*1*	1	1	1
ppa	K01507	inorganic pyrophosphatase	***2***	***1***	***2***	**1**	**1**	**2**
ppk	K00937	polyphosphate kinase	*1*	*1*	*1*	1	1	1
ppx	K01524	exopolyphosphatase	*1*	*1*	*1*	1	1	1
		Total	*30*	*30*	*32*	19	22	25

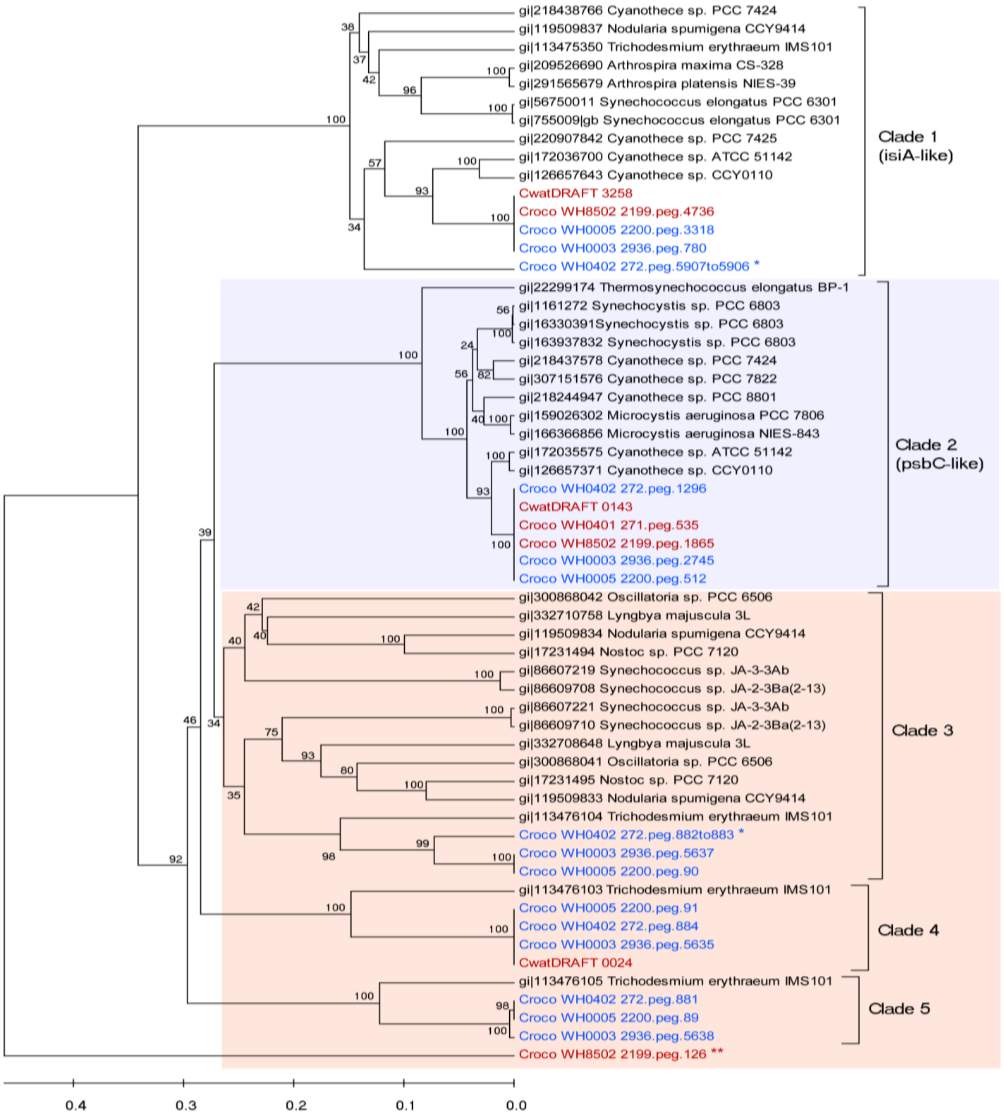
Table 8. Counts of iron-related genes in each *C. watsonii* genome (likely split ORFs are marked with *)

Gene	KEGG Orthology number	Function	*WH0003*	*WH0005*	*WH0402*	WH0401	WH8501	WH8502
isiA		iron-stress chlorophyll-binding protein (CP43′)	***4***	***4***	***4***	**1**	**3**	**2**
fldA/isiB	K03839	flavodoxin 1	*2*	*2*	*2*	2	1	2
		putative flavodoxin	*1*	*1*	*1*	1	1	1
idiA/futA/ afuA/fbpA	K02012	iron(III) transport system substrate-binding protein	*1*	*1*	*1*	1	1	1
idiB/futB/ afuB/fbpB	K02011	iron(III) transport system permease protein	***1***	***0***	***1***	**2**	**2**	**1**
idiC/futC/ afuC/fbpC	K02010	iron(III) transport system ATP-binding protein	*1*	*1*	*2**	1	1	1
feoA	K04758	ferrous iron transport protein A	***3***	***2***	***3***	**2**	**3**	**2**
feoB	K04759	ferrous iron transport protein B	***4***	***3***	***6****	**2**	**3**	**2**
fur	K03711	ferric uptake transcriptional regulator	*4*	*4*	*4*	4	4	4
tonB	K03832	ferric siderophore transport system, periplasmic binding protein	*1*	*1*	*1*	1	1	1
exbB	K03561	ferric siderophore transport system, biopolymer transport protein	*2*	*2*	*2*	2	2	2
exbD/tolR	K03559	Biopolymer transport protein	*2*	*2*	*2*	2	2	2
dps	K04047	starvation-inducible DNA-binding protein	***0***	***1***	***1***	**1**	**0**	**0**
bfr	K03594	bacterioferritin	*2*	*2*	*2*	2	2	2
	K02014	ferrichrome-iron receptor	*1*	*1*	*2*	1	1	2
	K02016	iron complex transport system substrate-binding protein (ABC transporter)	***4***	***2***	***2***	**1**	**1**	**0**
		Total (counting split ORF as 1)	*33*	*29*	*33*	26	28	25

Table 9. Counts of Photosystem I and II genes in the genome of each *C. watsonii* strain. (likely split ORFs are marked with *)

Gene	**KEGG Orthology number**	**Function**	***WH0003***	***WH0005***	***WH0402***	**WH0401**	**WH8501**	**WH8502**
Photosystem I genes								
btpA	K06971	photosystem I biogenesis protein	*1*	*1*	*2**	1	1	1
psaA	K02689	PSI P700 chlorophyll a apoprotein A1	*2**	*1*	*1*	2*	1	1
psaB	K02690	PSI P700 chlorophyll a apoprotein A2	*1*	*1*	*2**	1	1	1
psaC	K02691	photosystem I subunit VII	*1*	*1*	*1*	1	1	1
psaD	K02692	photosystem I subunit II	*1*	*1*	*1*	1	1	1
psaE	K02693	photosystem I subunit IV	*1*	*1*	*1*	1	1	1
psaF	K02694	photosystem I subunit III	*1*	*1*	*1*	1	1	0
psaK	K02698	photosystem I subunit X	*1*	*1*	*1*	1	1	0
psaL	K02699	photosystem I subunit XI	*1*	*0*	*1*	1	1	1
Photosystem II genes								
psb27	K08902	photosystem II Psb27 protein	*1*	*1*	*1*	1	1	1
psb28	K08903	photosystem II 13kDa protein	*2*	*2*	*2*	2	2	2
psbA	K02703	PSII P680 reaction center D1 protein	***7***	***3***	***1***	**3**	**2**	**3**
psbB	K02704	PSII CP47 chlorophyll apoprotein	*1*	*1*	*1*	1	1	0
psbC	K02705	PSII CP43 chlorophyll apoprotein	*1*	*1*	*1*	1	1	1
psbD	K02706	PSII P680 reaction center D2 protein	*2*	*1*	*1*	1	1	2
psbE	K02707	PSII cytochrome b559 subunit alpha	*1*	*1*	*1*	1	1	1
psbF	K02708	PSII cytochrome b559 subunit beta	*1*	*1*	*1*	1	1	1
psbH	K02709	PS II 10 kDa phosphoprotein	*1*	*1*	*1*	1	1	1
psbI	K02710	photosystem II reaction center protein I	*1*	*1*	*1*	1	1	1
psbJ	K02711	photosystem II reaction center protein J	*1*	*1*	*1*	1	1	1
psbK	K02712	photosystem II reaction center protein K	*1*	*1*	*1*	1	1	1
psbL	K02713	photosystem II reaction center protein L	*1*	*1*	*1*	1	1	1
psbM	K02714	photosystem II reaction center protein M	***1***	***0***	***1***	**1**	**0**	**0**
psbN	K02715	photosystem II reaction center protein N	*1*	*1*	*1*	1	1	1
psbO	K02716	PS II oxygen-evolving enhancer protein 1 (manganese-stabilizing protein)	*1*	*1*	*1*	1	1	1
psbP	K02717	PS II oxygen-evolving enhancer protein 2	*1*	*1*	*2*	1	1	1
psbU	K02719	photosystem II 12 kD extrinsic protein	*1*	*1*	*1*	1	1	1
psbV	K02720	photosystem II cytochrome c550	*1*	*1*	*1*	1	1	1
psbX	K02722	photosystem II protein X	*1*	*1*	*1*	1	1	1
psbZ	K02724	photosystem II reaction center protein Z	*0*	*0*	*1*	1	1	1
		Total (counting split ORF as 1)	*37*	*30*	*32*	33	31	30

Figure 8. Evolutionary relationships of *C. watsonii psbC* and *isiA* genes


gi|218438766 Cyanothece sp. PCC 7424
gi|119509837 Nodularia spumigena CCY9414
gi|113475350 Trichodesmium erythraeum IMS101
gi|209526690 Arthrospira maxima CS-328
gi|291565679 Arthrospira platensis NIES-39
gi|56750011 Synechococcus elongatus PCC 6301
gi|755009|gb Synechococcus elongatus PCC 6301
gi|220907842 Cyanothece sp. PCC 7425
gi|172036700 Cyanothece sp. ATCC 51142
gi|126657643 Cyanothece sp. CCY0110
CwatDRAFT 3258
Croco WH8502 2199.peg.4736
Croco WH0005 2200.peg.3318
Croco WH0003 2936.peg.780
Croco WH0402 272.peg.5907to5906 *
Clade 1
(isiA-like)
gi|22299174 Thermosynechococcus elongatus BP-1
gi|1161272 Synechocystis sp. PCC 6803
gi|16330391Synechocystis sp. PCC 6803
gi|163937832 Synechocystis sp. PCC 6803
gi|218437578 Cyanothece sp. PCC 7424
gi|307151576 Cyanothece sp. PCC 7822
gi|218244947 Cyanothece sp. PCC 8801
gi|159026302 Microcystis aeruginosa PCC 7806
gi|166366856 Microcystis aeruginosa NIES-843
gi|172035575 Cyanothece sp. ATCC 51142
gi|126657371 Cyanothece sp. CCY0110
Croco WH0402 272.peg.1296
CwatDRAFT 0143
Croco WH0401 271.peg.535
Croco WH8502 2199.peg.1865
Croco WH0003 2936.peg.2745
Croco WH0005 2200.peg.512
Clade 2
(psbC-like)
gi|300868042 Oscillatoria sp. PCC 6506
gi|332710758 Lyngbya majuscula 3L
gi|119509834 Nodularia spumigena CCY9414
gi|17231494 Nostoc sp. PCC 7120
gi|86607219 Synechococcus sp. JA-3-3Ab
gi|86609708 Synechococcus sp. JA-2-3Ba(2-13)
gi|86607221 Synechococcus sp. JA-3-3Ab
gi|86609710 Synechococcus sp. JA-2-3Ba(2-13)
gi|332708648 Lyngbya majuscula 3L
gi|300868041 Oscillatoria sp. PCC 6506
gi|17231495 Nostoc sp. PCC 7120
gi|119509833 Nodularia spumigena CCY9414
gi|113476104 Trichodesmium erythraeum IMS101
Croco WH0402 272.peg.882to883 *
Croco WH0003 2936.peg.5637
Croco WH0005 2200.peg.90
Clade 3
gi|113476103 Trichodesmium erythraeum IMS101
Croco WH0005 2200.peg.91
Croco WH0402 272.peg.884
Croco WH0003 2936.peg.5635
CwatDRAFT 0024
Clade 4
gi|113476105 Trichodesmium erythraeum IMS101
Croco WH0402 272.peg.881
Croco WH0005 2200.peg.89
Croco WH0003 2936.peg.5638
Clade 5
Croco WH8502 2199.peg.126 **
38
37
42
100
96
100
100
57
100
93
100
34
56
100
24
100
56
82
40
100
100
100
93
100
39
42
40
40
100
100
100
46
34
75
93
80
100
35
99
98
92
100
100
100
100
98
100
0.4
0.3
0.2
0.1
0.0

Relationships were inferred using the UPGMA method with bootstrap support, as percent of 500 replicates, shown next to the branches. The optimal tree with the sum of branch length = 5.82073916 is shown drawn to scale, with branch lengths computed using the Poisson correction method in the units of number of amino acid substitutions per site. Small-cell *Crocosphaera* strain sequences are in red text and large-cell strains are in blue text. Gray shading indicates *psbC* sequences. Orange shading indicates the three adjacent likely *isiA* ORFs in *Crocosphaera* and *Trichodesmium* genomes. ** indicates the ORF is truncated by end of contig.
* indicates two adjacent ORFs separated by a stop codon (likely sequencing error) were combined into a single sequence.

Chapter 4

Natural abundances of two *Crocosphaera* types in the North and South Pacific

Abstract

Crocosphaera watsonii is an abundant unicellular nitrogen (N_2) fixing cyanobacterium that is important in marine ecosystems, primarily because it supplies fixed nitrogen to the phytoplankton community in oligotrophic oceans. Isolates of *C. watsonii* are separated into two phenotypes (large-cell and small-cell) with genetic and metabolic differences in environmentally important processes such as exopolysaccharide production, nitrogen fixation rates, and photosynthetic efficiency, which indicates that the two types may have different impacts on nutrient cycling, and different marine niches. However, prior to this study, molecular assays for *C. watsonii* abundances in natural samples were unable to differentiate between the two phenotypes, so their distributions were essentially uncharacterized. As a first step to understanding those distributions, this study developed *C. watsonii* type-specific qPCR assays and applied them to samples from the North and South Pacific. Both samples sets showed a rapid decline in both *Crocosphaera* types between 45 and 75 m, and a dominance of the large-cell types below 100 m. In contrast, above 75 m, small-cell types typically outnumbered large-cell types by a factor of 10 or more, and this was more pronounced in the N. Pacific. In addition, despite cell-diameters of ~6 µm, a significant portion of the large-cell types were captured on 10 µm pore size filters in N. Pacific samples from above 75 m, suggesting aggregation was more common in that sub-population. Because this is the first study to report natural abundances of the two *C. watsonii* types throughout the water column, future experiments will be needed to determine which patterns observed in these samples are

typical in all ecosystems, and which vary by season or ocean basin. This study, and future work using the newly developed assays, will provide important details of *C. watsonii* populations, which are critical for assessing the role of *Crocosphaera* in global biogeochemical cycling.

Introduction

Crocosphaera watsonii is a species of unicellular nitrogen (N_2) fixing cyanobacteria (UCYN) that is important in marine primary production and biogeochemical cycling. This is especially true in oligotrophic gyres where nitrogen is often a limiting nutrient, and *C. watsonii* can provide biologically available nitrogen to the phytoplankton community by fixing nitrogen at significant rates (Zehr et al., 2001; Falcon et al., 2004; Montoya et al., 2004; Kitajima et al., 2009; Moisander et al., 2010). As such, UCYN are among the largest contributors of fixed nitrogen in marine systems and measurements of UCYN abundance, made using direct microscope counts and qPCR of the *nifH* gene, are important for estimating basin-scale N_2-fixation rates. Direct counts have recorded abundances of 10^4 to 10^7 cells/liter in the North Pacific (Zehr et al., 2001; Church et al., 2005), and near 10^5 cells/liter in the Atlantic (Falcon et al., 2004), and qPCR studies have found between 10^3 and 10^6 gene copies/ liter in multiple ocean basins (Zehr et al., 2001; Church et al., 2005; Church et al., 2008; Langlois et al., 2008; Moisander et al., 2010; Falcon et al., 2004; Moisander et al., 2008). Such measurements of natural *Crocosphaera* abundances count the species as a single population because there is no genetic variation in the *nifH* gene.

The lack of genetic variation observed in *Crocosphaera nifH* sequences was also observed in the 16S rDNA sequence, and a number of other genes that were examined in natural populations and cultivated strains (Zehr et al., 2007). Despite that apparent genetic conservation, two distinct phenotypic categories have been identified in *C. watsonii* isolates. The first (large-cell) phenotype has cell-diameters over 4 µm, produces abundant extracellular polysaccharide (EPS), has higher photosynthetic efficiencies (F_v/F_m), and higher per-cell nitrogen fixation rates (Webb et al., 2009; Sohm et al., 2011). The other (small-cell) type has cell diameters less than 4 µm, and does not produce noticeable amounts of EPS. There is also evidence that the small-cell types also grow in a narrower temperature range, and are missing some phosphorus scavenging genes that are found in the large-cell types (Dyhrman and Haley, 2006; Webb et al., 2009). More recently, genome comparisons found that the large-cell types contained a variety of genetic functions that were missing from the genomes of the small-cell types, such as EPS biosynthesis, iron stress response genes, and phosphorus metabolism genes (Bench et al., 2011; Bench et al., 2012). Because of their genetic and metabolic differences, it is likely that the two types have different impacts on biogeochemical cycling. For example, EPS is a carbon-rich compound that can protect cells and cause aggregation which increases sinking rates (Passow et al., 2001; Pereira et al., 2009; Sohm et al., 2011), which would make the large-cell types more significant contributors to carbon export from surface water than the small-cell types. The differences in iron and phosphorus related genes further suggest that the two *C. watsonii* types may be differently adapted to their chemical environment, and therefore may have different niches. However, because the standard methods used to measure natural *Crocosphaera* abundances are limited to viewing the species as a single population, very little is known about the distribution of two types in the water column or the global oceans. A single previous study carried out microscopic cell counts to quantify relative abundance of two classes of *Crocosphaera* cell size in the western South Pacific (Webb et al., 2009). That work only included a single depth, but in those samples, the smaller cells were slightly more abundant than the larger cells. Evidence of cell aggregation was also seen in the larger-cells (Webb et al., 2009), which further supports the possibility of different levels of nutrient export by the two phenotypes. In light of the ecologically relevant differences between types, is it important to better understand their distributions in order to accurately assess the impact of the species on nutrient cycling and the marine ecosystem.

The first goal of this study was to use the recently identified phenotype-specific genes to develop molecular assays to differentiate between *Crocosphaera* types. The second goal was to apply those assays to water column samples and thereby provide an initial assessment of the distribution of the two *C. watsonii* types in natural populations.

Methods

Design and testing of qPCR assays

Previously completed comparisons of six *Crocosphaera watsonii* genomes enabled identification of genes unique to each phenotype (Bench et al., 2011; Bench et al., 2012). Two genes were identified for each phenotype that were found in all of the genomes of that type, but were absent from all genomes of the alternate type. Primer-probe set (pps) was designed for each gene using Primer 3 (Rozen and Skaletsky, 1999) with a goal T_m of 64°C for primers, and 74°C for probes. All four genes and corresponding primer and probe sequences were used in nucleotide BLAST searches against the CAMERA (Sun et al., 2011) and GenBank NT and WGS (Benson et al., 2003) databases to verify that they did not have significant sequence similarity to other known organisms. The genes used as template sequences and the resulting primers and probes for all four loci are listed in Table 1, and the design of the pps for the *nifH* locus was previously described (Moisander et al., 2010). Dual-label probes were synthesized with FAM fluorescent tags and TAMRA quenchers. Reactions were set up in sterile PCR hoods using UV sterilized optical tubes or plates and contained 1.5 - 2 µl of template DNA plus 1 µl of each primer (10 µM), 0.5 µl of probe (10 µM), 12.5 µl TaqMan Gene Expression 2X Master Mix (Applied Biosystems), and water to a final volume of 25 µl. Amplification and detection was carried out on an ABI 7500 instrument using the following 2-step reaction: initial steps of 50°C for 2 minutes, then 95°C for 10 minutes, then 45 cycles of 90°C for 15 seconds, then 60°C for 60 seconds. Each run included 3 or 4 no template controls (NTCs) and a set of standards, in triplicate, with known gene copies from 10^0 to 10^7 per reaction. Following each run, the C_t of each standard was plotted versus the log of its gene copy number to create a standard curve. The equation for that standard curve was used to calculate the gene copies in each of the sample reactions from the same run.

Standards were made from amplified genomic DNA from strains of the appropriate phenotype. To avoid amplifying DNA from contaminants in non-axenic cultures, cells were sorted using a flow cytometer prior to whole genome amplification (WGA) with Repli-g (Qiagen). The sorting and WGA were carried out as described in the methods used for genome sequencing of *C. watsonii* strains (Bench et al., 2011; Bench et al., 2012). Amplified genomic DNA was quantified using Pico Green (Life Technologies), and genome copies/µl were calculated based on the DNA concentration and the draft genome sizes (Bench et al., 2012). Appropriate dilutions were made to generate a set of standards that contained 10^0 to 10^7 genome copies in 2µl (the volume used in each reaction). Multiple sets of the prepared genomic standards were compared in triplicate to *nifH* linearized plasmid standards to verify the DNA quantification, and relative reaction efficiency, and no significant differences were observed between the plasmid and any of the genomic standards (Supp. Fig. 1 and 2).

Tests for cross reactivity and inhibition were carried out for all loci using multiple mixtures of DNA from different *C. watsonii* strains. Names and phenotypes of *C. watsonii* strains used are listed in table 2. Four test mixtures contained genomic

DNA from WH8501 and WH0003 strains in the following ratios: 1:3, 3:1, 1:10, and 10:1. Eight additional test mixtures contained WH8501 or WH0003 DNA mixed with 3-fold more DNA (final ratio of 1:3) from one of four additional strains (WH8502, WH0401, WH0401, and WH0005). The pps for each locus was tested for amplification and inhibition in triplicate qPCR reactions with the 12 different mixtures, which ranged over an order in magnitude in target DNA and non-target DNA concentrations, and included samples that contained only target DNA as well as only non-target DNA. Copy numbers from qPCR reactions were very close (± 30% for nearly all samples) the expected numbers based on DNA concentrations used in each reaction, and there was no amplification in any of the samples that did not contain target DNA. Because there was no apparent cross-reactivity or inhibition from non-target strains, all four pps were determined to be appropriate for use in environmental samples.

Sample collection, DNA extraction and qPCR of cruise samples

South Pacific samples were collected during the R/V Kilo Moana cruise KM0703 in March and April of 2007. Cruise station locations as well as methods for water sample collection and processing and DNA extraction were described previously (Moisander et al., 2010). North Pacific samples were collected during the BioLINCS cruise in September of 2011, which took place north of Station Aloha (Figure 1). At each station, water samples were collected from multiple discrete depths (5 to 175 m) in Niskin bottles mounted on a CTD rosette. Two to three liters of collected water was filtered through two in-line Durapore filters (10 µm pore size, followed by 0.2 µm pore size). Filters were placed in bead beater tubes with 0.1 g sterile glass beads, immediately frozen and subsequently stored at -80°C until DNA extraction.

The DNA extraction protocol used for N. Pacific samples is a further modification of the modified DNeasy Plant MiniKit (Qiagen) protocol used in to extract the S. Pacific samples (Moisander et al., 2008; Moisander et al., 2010). Filters were thawed and 400 µl of AP1 buffer (provided in kit) was added to each tube. Samples then underwent three freeze-thaw cycles of rapid freezing in liquid N_2, followed by rapid thawing in a 65°C heat block. The samples were then bead-beat in Mini-Beadbeater-96 (Biospec Inc.) for 2 minutes. Tubes were centrifuged briefly prior to addition of 45 µl (20 mg/ml) of Proteinase K (Qiagen), and then vortexed briefly and incubated (with rocking) at 55°C for 1 hour. An RNaseA digestion was then carried out by adding 4 µl of RNaseA to each sample, vortexing and incubating at 65°C for 10 minutes. The filters were removed from the tubes, and 130 µl of AP2 buffer (provided in kit) was added to each tube followed by a brief vortex and a 10 minute incubation on ice. Tubes were spun for 5 minutes at 14,000 RPM to pellet beads large precipitates, and the supernatant for each sample was then transferred to sterile 2 ml locking Sample tubes RB (Qiagen). DNA was then extracted from the transferred supernatant using the standard reagents and protocols for “Plant Cell & Tissues” with the “DNeasy Plant Mini” kit in the QIAcube instrument. The final elution volume for each sample was 100 µl.

The *capD* and UDPhydro assays of environmental samples used the same qPCR reaction contents (except template DNA), genomic DNA standards, and cycling conditions as those used in the pps testing described above. DNA extracts were diluted 1:5 (N. Pacific samples) or 1:1 (S. Pacific samples) and 1.5 µl of the dilution was used in triplicate reactions. For the N. Pacific samples, *nifH* reactions contained 2 µl of undiluted DNA extract in duplicate reactions, and for the S. Pacific samples, previously reported *nifH* gene copies (Moisander et al., 2010) were used for comparison to *capD* and UDPhydro gene copies. The diluted template DNA reactions had a higher limit of quantification (LOQ) in the resulting qPCR calculations, and as such a number of samples from 100 m and below were under this LOQ. The good correlation with the undiluted *nifH* copy numbers, which were well above the LOQ, provided enough confidence in those data to include them in this analysis. However, some or all of the samples that were below the LOQ will need to be re-run with undiluted template DNA to verify those numbers.

Results and Discussion

Despite known phenotypic differences in cultivated strains, *Crocosphaera watsonii* has been treated as a single population in environmental samples, because available molecular markers were designed for genes that showed no genetic variation between strains or within natural populations. However, genomic comparisons of multiple cultivated strains revealed strain-specific and phenotype-specific genes that could be used to differentiate between strains (Bench et al., 2011; Bench et al., 2012). Based on that genomic data, four qPCR assays were designed as part of this study; two designed to target large-cell type strains, and two targeted to small-cell strains. Each pps was designed to a separate gene, resulting in assays for 4 different genomic loci (Table 1). Using the established Group B *nifH* pps (Moisander et al., 2010) as a positive quantitative control, each qPCR locus was thoroughly tested using DNA from six *C. watsonii* cultivated strains (see methods). All loci amplified quantitatively as expected for DNA from strains with the targeted phenotype, and did not amplify with DNA from non-target phenotype strains (Table 2 and Supp. Fig. 1 and 2). For example, the two loci designed to target large-cell strains (*capD* and polyExp), amplified as predicted only with DNA from large-cell type strains (WH0003, WH0005, and WH0402). This was true for all loci, even when DNA from non-targeted phenotype strain (i.e. not expected to amplify) was ten-fold greater than DNA from the targeted strain (see methods). This testing demonstrated that, when using DNA from isolates, all four qPCR assays were robust without cross-reactivity or noticeable inhibition from un-targeted *C. watsonii* phenotypes.

Following testing with DNA from isolated strains, the qPCR assays were applied to two sets of research cruise samples to investigate the natural distributions of the two phenotypes. One sample set was collected during September 2011 north of Hawaii in the oligotrophic gyre of the N. Pacific Ocean (Figure 1). Collected water samples were filtered through two in-line filters; first one with a 10µm pore size, followed by one with a 0.2µm pore size. Because all *C. watsonii* isolates have cell diameters that range from 3.5 µm to ~6 µm (Webb et al., 2009; Sohm et al., 2011), it is predicted that the vast majority of naturally occurring *C. watsonii* should pass through the 10 µm filter and be captured on the 0.2 µm filter. However, the large-cell phenotype is also known to produce copious amount of EPS, and to form multi-cell aggregates as a result (Webb et al., 2009; Sohm et al., 2011). Such aggregates could contribute to retention of the large-cell type on the larger pore size filter, as was observed in microscopic examination of the *Crocosphaera* community the western South Pacific (Webb et al., 2009). To test that prediction, phenotype specific assays were run on both filter sizes and the relative contribution of the 0.2 µm filter sample to the total for each sample was calculated. For the vast majority of samples assayed with the UDPhydro, (small-cell specific) locus, nearly 100% of gene copies were found on the 0.2 µm filter, and only four samples out of 58 (plus three that had no amplification) showed less than 80% of gene copies on the 0.2 µm filter (right panel in Figure 2).

The large-cell specific assay (*capD*) showed a very different pattern, with less than 40% of the total copies found on the 0.2 µm filter for many samples collected from 75 m and shallower (left panel in Figure 2). The samples collected from deeper than 75 m had a pattern more similar to the small-cell locus, but with a slightly higher fraction of samples in the 50- 85% range. The distinction in samples from above and below 75 m could be explained by the presence of two subgroups of large-cell type *Crocosphaera*, such as a shallow population that forms large aggregates, as observed in isolates, and is adapted to higher light, and a second deeper population that is adapted to lower light levels and has less aggregation. This would be similar to what is observed in sympatric *Prochlorococcus* strains that have established separate high-light and low-light niches in the water column (West and Scanlan, 1999; Rocap et al., 2003; Scanlan et al., 2009). Alternatively, the pattern could be explained by a process where aggregates break up as they sink and/or are partially grazed. This would release some cells from aggregates, thereby allowing a larger proportion to pass through the 10 µm filter. If that were the explanation for the observed pattern, it would mean that the presence of *C. watsonii* below 75 m is not necessarily indicative of those cells being active and dividing, but is primarily a result of sinking and mixing. Because the two proposed scenarios have vastly different predictions for the metabolic state of the deeper population of cells, it is unclear what contribution that population may be making to the microbial ecosystem at those depths. Future experiments could possibly differentiate between the two scenarios by measuring sinking rates using sediment traps, and assessing the viability of the deeper population through measurements such as relative cellular fluorescence and relative gene expression levels of the two groups.

Comparing abundances of the two *Crocosphaera* types along the cruise track revealed that the small-cell type accounted for the vast majority of total *Crocosphaera* abundance in samples shallower than 75 m, while the large-ell type dominated in samples deeper than 100 m (Figure 3). Total *Crocosphaera*, as measured by *nifH* qPCR with the two filter sizes summed for each sample (see methods), was very high (often >10^6 gene copies per liter) in samples from 50 m and shallower, and rapidly declined by 2 to 3 orders of magnitude between 50 and 100 meters, decreasing to abundances of ~10^3 gene copies per liter at and below 150 m (upper left panel of Figure 3). The small-cell *Crocosphaera* showed a similar pattern, with UDPhydro gene abundances ranging from over 10^6 gene copies per liter at the surface to less than 10^3 gene copies per liter in samples at or below 125 m (Figure 3, upper right panel). By comparison, the large-cell type *Crocosphaera* were much less abundant in the upper water column (<75 m depth) at 10^4 to 10^5 gene copies (*capD* locus) per liter (lower left panel of Figure 3). However, many samples below 75 m had *capD* abundances measured well above 10^3 copies per liter, which was often much higher than the small-cell abundance in the same samples. As such, the ratio of large-cell to small-cell *Crocosphaera* (as calculated by dividing *capD* gene copies by UDPhydro gene copies for each sample) was typically between 0.01 and 0.1 in samples above 50 m, while it was between 4 and 8 in samples below 75 m (lower right panel of Figure 3). This ratio was the highest at station 2 (~20 km distance in figure 3), in the 100 and 125 m samples and second highest at station 7 (~110 km distance) in the 125 and 150m samples. These stations also had some of the highest abundances of large-cell type *Crocosphaera* in the upper water column, which apparently supports the proposed scenario of a sinking process driving abundances of large-cells in deeper samples.

Flow cytometry measurements (FCM) of *Crocosphaera* abundances also showed the highest counts in the shallowest samples (Supp. Table 1). However, while there was some correlation in relative abundances, FCM cell counts were far below qPCR counts (7% to 35% of *nifH* gene copies) and no *Crocosphaera* cells were found by FCM below 75 m. This may be due to the much higher limit of detection and much smaller sample size for FCM, or it could indicate that the cells in deeper water are less metabolically active and have a corresponding low level of fluorescence that prevents proper assessment by FCM. More detailed comparisons of qPCR and FCM data will be needed to clarify the reason for the apparent discrepancy between the two counts.

Abundances of the two *Crocosphaera* phenotypes were also measured in samples taken at three stations during a research cruise in the South Pacific Ocean in the spring of 2007. In these samples, the surface abundances of *Crocosphaera* were nearly as high as those measured in the N. Pacific samples, with *nifH* and UDPhydro assays measuring nearly 10^5 gene copies per liter to well over 10^6 copies per liter (Figure 4). In contrast to the N. Pacific samples, large-cell abundances were much closer to the small-cell abundances in surface samples. In fact, at station 23, the large-cells were more abundant at all depths except for 75 m. The tendency, observed in N. Pacific upper water column samples, to capture a substantial portion of large-cell type *Crocosphaera* on the 10 µm filter was not seen in S. Pacific samples. Based on *capD* gene copies, only one sample showed less than 60% of large-cells on the 0.2 µm filter, two samples were between 60-70% and all remaining samples were over 80%, with an average of 89% for all samples (Supp. Table 2). In the small-cell qPCR assays (UDPhydro locus) over 80% of total abundance was measured on the 0.2 µm filter for all samples, with many close to 100% and an overall average of 92%. The much higher abundance of the large-cells in surface coupled with their lower rate of capture on the 10 µm filter could be explained by the presence of different *Crocosphaera* strains in the S. Pacific samples than those in the N. Pacific samples. If the S. Pacific strains produced less EPS, there would be less aggregation. Alternatively, the physical and/or chemical conditions in the S. Pacific could simply have been more favorable to the large-cell types than the conditions in the N. Pacific, thereby enabling the large-cell types to grow to higher abundances.

While the small-cell type *Crocosphaera* were less dominant in the S. Pacific upper water column samples relative to the N. Pacific samples, at depths below 75 m both sample sets showed a similar dominance of large-cell type *Crocosphaera*, supporting the theory that a faster sinking rate and longer persistence of large-cell type, which are typically protected by EPS, could result in an over-representation of that phenotype in deeper samples. However, with only two sample sets, and only minimal understanding of the conditions that favor the two phenotypes, future experiments will be needed to support or rule out that proposed process.

Conclusions

The qPCR assays developed for this study provide a straightforward and robust method to assess the relative abundances of two sub-types of *Crocosphaera* in natural samples. As this is the first study to examine depth profiles of the two types, and only included samples from two regions and seasons, it is impossible to draw broad conclusions about natural abundances. However, in both of these samples, there was a dramatic decline in overall *Crocosphaera* abundance between 40 m and 100 m. In addition, the small-cell types are usually dominant by 1-2 orders of magnitude in the upper water column (above 75 m), while the large-cell types dominate in samples from below 75 m. The N. Pacific samples also showed more evidence of aggregation in the large-cell types than samples from the S. Pacific. In future work, it will be important to determine if the same patterns are observed in all regions where *Crocosphaera* is abundant. Because both sample sets had very high *Crocosphaera* numbers, the relative abundance patterns of the two types will also need to be assessed in samples with lower population counts. If these patterns are determined to be common, or universal, they will be important to consider when estimating *Crocosphaera* contributions to marine nitrogen and carbon cycling. An important part of that determination will be identifying the processes that drive abundances of the two *Crocosphaera* phenotypes. It is possible that there are two sub-types of large-cell *Crocosphaera* adapted to different niches (one shallow, and one deep), or that there are significant differences in sinking and decay processes for the small- and large-cell types. It may also be that the distributions of the two *Crocosphaera* types results from a more complicated combination of both explanations. Future experiments using additional field samples will be required to better understand *Crocosphaera* ecology, and the assays developed during this study will help researchers carry out those experiments.

References

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 Microbio* **2**.

Bench SR, Heller P, Frank I, Arciniega M, Shilova IN, Zehr JP. (2012). Investigation of *Crocosphaera watsonii* phenotypes through whole genome comparison of six strains. *In Prep*.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. (2003). GenBank. *Nucleic Acids Res* **31**: 23-27.

Church MJ, Jenkins BD, Karl DM, Zehr JP. (2005). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat Microb Ecol* **38**: 3-14.

Church MJ, Bjorkman KM, Karl DM, Saito MA, Zehr JP. (2008). Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol Oceanogr* **53**: 63-77.

Dyhrman ST, Haley ST. (2006). Phosphorus scavenging in the unicellular marine diazotroph *Crocosphaera watsonii*. *Appl Environ Microbiol* **72**: 1452-1458.

Falcon LI, Carpenter EJ, Cipriano F, Bergman B, Capone DG. (2004). N_2 Fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: phylogeny and in situ rates. *Appl Environ Microbiol* **70**: 765-770.

Kitajima S, Furuya K, Hashihama F, Takeda S, Kanda J. (2009). Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol Oceanogr* **54**: 537-547.

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.

Moisander PH, Beinart RA, Voss M, Zehr JP. (2008). Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME J* **2**: 954-967.

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.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. (2004). High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**: 1027-1031.

Passow U, Shipe RF, Murray A, Pak DK, Brzezinski MA, Alldredge AL. (2001). The origin of transparent exopolymer particles (TEP) and their role in the sedimentation of particulate matter. *Cont Shelf Res* **21**: 327-346.

Pereira S, Zille A, Micheletti E, Moradas-Ferreira P, Philippis RD, Tamagnini P. (2009). Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. *FEMS Microbiol Rev* **33**: 917-941.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.

Rozen S, Skaletsky H. (1999). Primer3 on the WWW for general users and for biologist programmers. In Bioinformatics Methods and Protocols. Misener S, and Krawetz SA (eds). Totowa, NJ: Humana Press, pp. 365-386.

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.

Sohm JA, Edwards BR, Wilson BG, Webb EA. (2011). Constitutive extracellular polysaccharide (EPS) production by specific isolates of *Crocosphaera watsonii*. *Front Microbio* **2**.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546-D551.

Webb EA, Ehrenreich IM, Brown SL, Valois FW, Waterbury JB. (2009). Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, *Crocosphaera watsonii*, isolated from the open ocean. *Environ Microbiol* **11**: 338-348.

West NJ, Scanlan DJ. (1999). Niche-partitioning of Prochlorococcus populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65**: 2585-2591.

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 U S A* **104**: 17807-17812.

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.

Table 1. Primer and probe sequences and gene information for each qPCR assay locus.

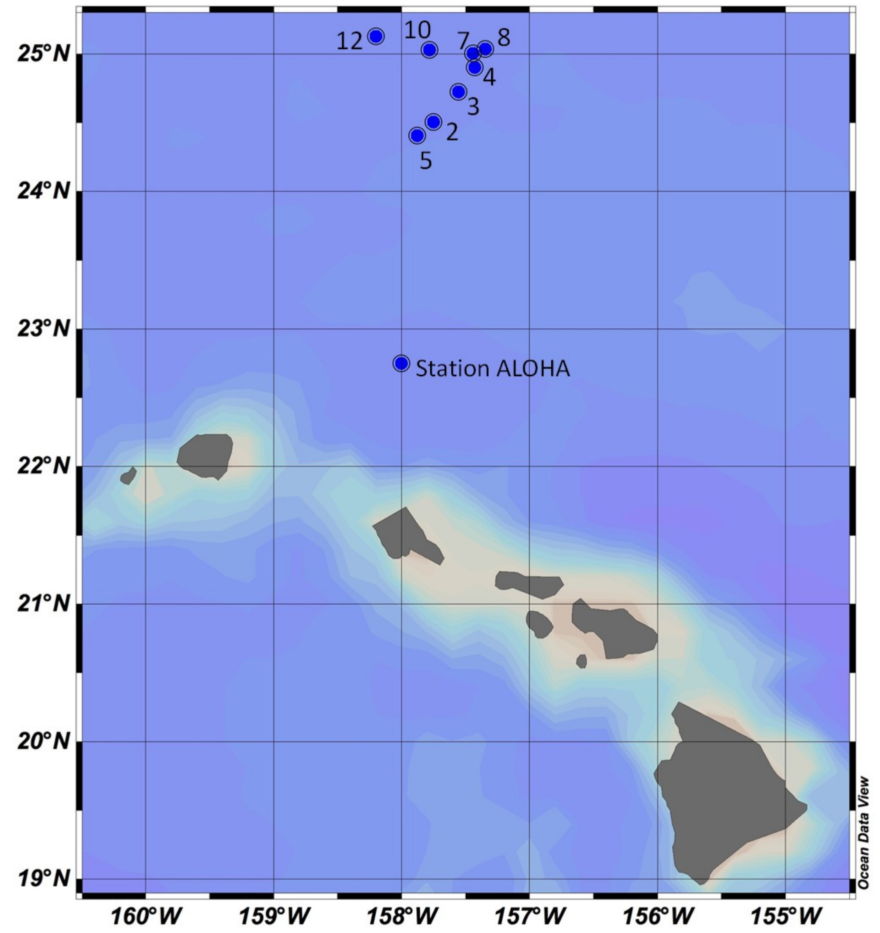
Locus name (target phenotype)	**source NCBI accession # (strain)**	**source ORF function**	**Forward primer (5' to 3')**	**Reverse primer (5' to 3')**	**Probe (5' to 3')**
capD (Large-cell)	EHJ11763 (WH0003)	polysaccharide biosynthesis protein CapD	TGCTGCTCATAAGCACGTTCC	GCCGCTAATTGTGCTAGATTACCC	GCAAGAAAATCCCACGGAATCCCTGG
polyExp (Large-cell)	EHJ11764 (WH0003)	polysaccharide export protein	CCTAGCACGGACGAAATTAGGC	TGTTCCTGGACGACGTACTTGC	TGCGCCAGTTGGCCTCATCCA
UDPhydro (Small-cell)	ZP_00518025 (WH8501)	5'-Nucleotidase / UDP-sugar hydrolase	AAATCCTCGGTGCTGACTTCG	CCACCGGCAACCTCTAAACC	CGACTTCTCCACCGATGAAAACCTGGC
ATPhelica (Small-cell)	ZP_00514180 (WH8501)	ATP-dependent helicase	TTTCGTTTTACCATTCCTTTAGTACGC	TTGAGGGTATCCACAGAAACAATAGC	GGCATTCAGCGTTTGCGTAATCGGA
nifH (both types)	ZP_00516386 (WH8501)	Nitrogenase iron protein	CGTAATGCTCGAAGGGTTTGA	CACGACCAGCACAACCAACT	CAAGTGTGTAGAATCTGGTGGTCCTGAGCC

Table 2. qPCR results for all loci with DNA from cultivated *C. watsonii* strains

		qPCR amplification results for each locus[a]				
Phenotype	**Strain Name**	***capD***	**polyExp**	**UDPhydro**	**ATPhelica**	***nifH***
Large-cell	WH0003	(+)	(+)	no amp.	no amp.	(+)
	WH0005	(+)	(+)	no amp.	no amp.	(+)
	WH0402	(+)	(+)	no amp.	no amp.	(+)
Small-cell	WH8501	no amp.	no amp.	(+)	(+)	(+)
	WH8502	no amp.	no amp.	(+)	(+)	(+)
	WH0401	no amp.	no amp.	(+)	(+)	(+)

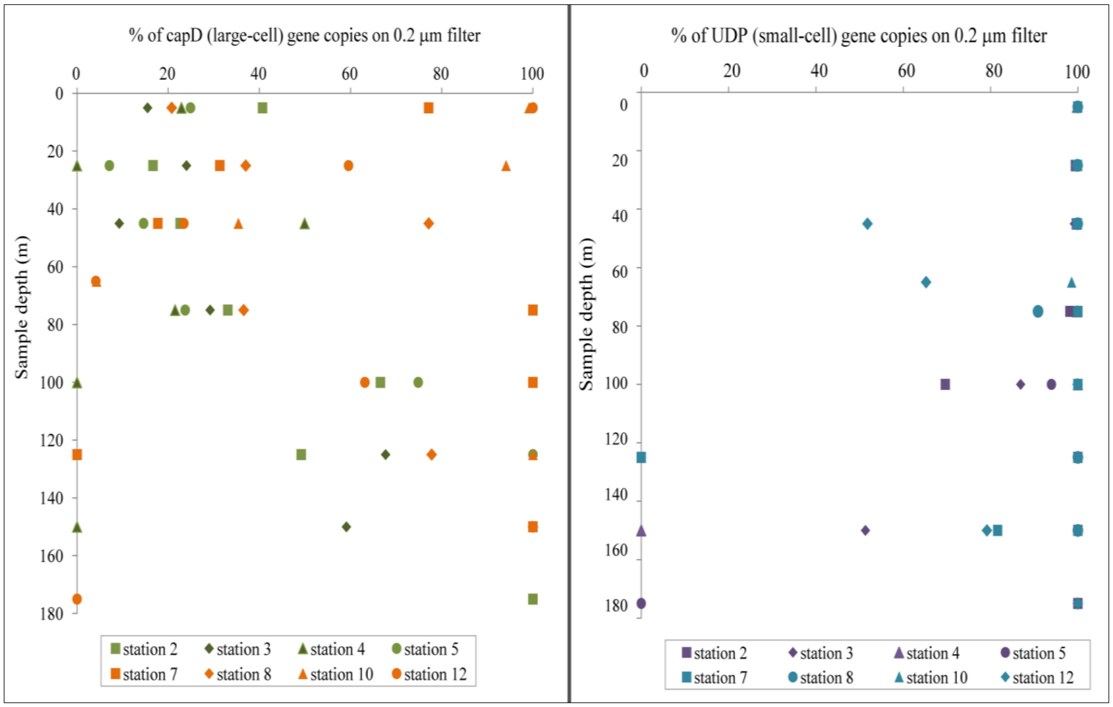
a: (+) indicates that amplification was quantitative between 10^0 and 10^7 gene copies for all loci. See supplemental figures 1 and 2 for C_t vs log gene copy plots.
"no amp." indicates that in all tested reactions, amplification was completely undetected (i.e. no C_t value could be calculated).

Figure 1. North Pacific cruise sample locations



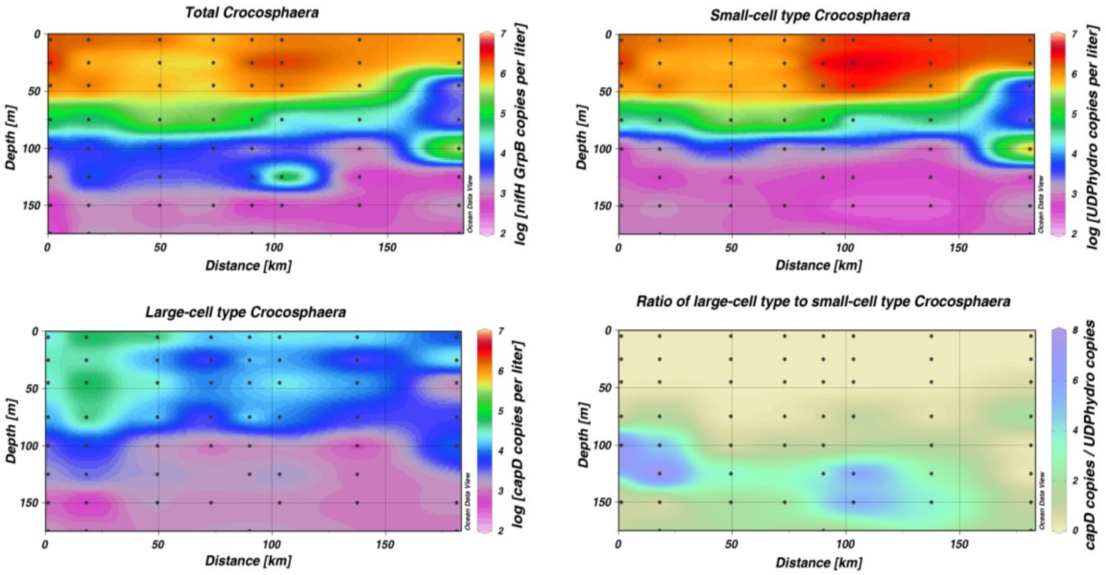
Numbers indicate the eight stations where samples for this study were collected during the September 2011 BioLINCS cruise in the N. Pacific. The Hawaiian Islands and Station ALOHA are shown for reference only.

Figure 2. Percent of each type of *C. watsonii* found on the 0.2 µm filter (remaining fraction was found on the 10 µm filter)



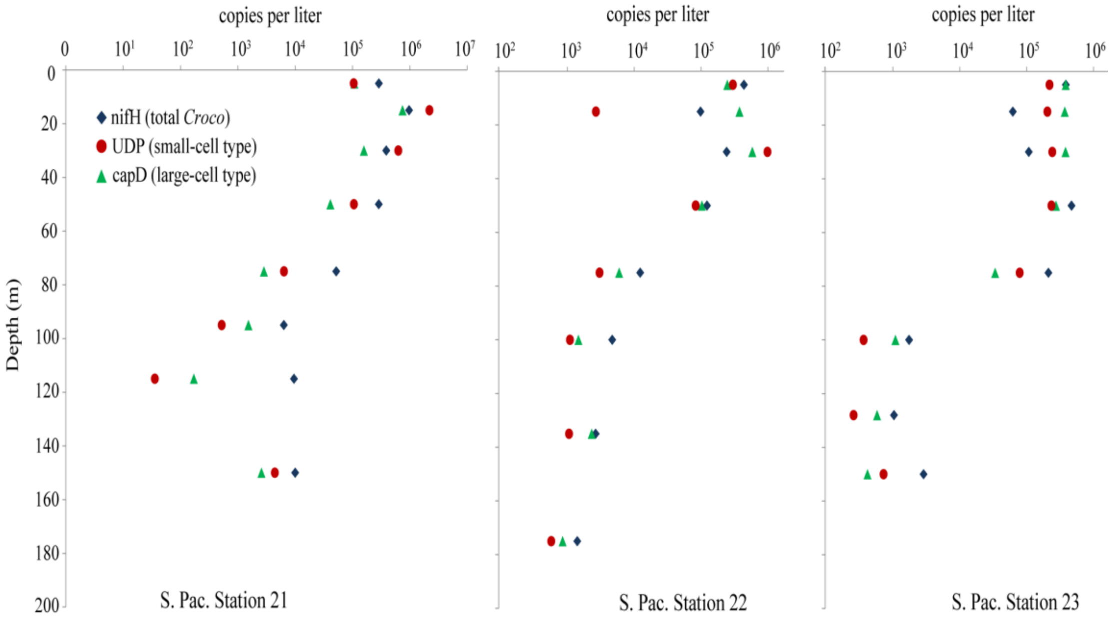
For each locus, the total gene copies were calculated for each sample by summing gene copies on both filters (10 µm and 0.2 µm), and the contribution of gene copies on the 0.2 µm filter was calculated as a percent of that total. The results were plotted as percentages (x-axis) vs depth in the water column (y-axis). The stations where samples were collected are designated by the symbols in the legends at the bottom of the plots. The UDPhydro locus (specific for small-cell strains) is on the right, and the *capD* locus (specific for large-cell strains) is on the left.

Figure 3. Patterns of *Crocosphaera* abundance in the N. Pacific



Contour plots of qPCR gene copies of total *Crocosphaera* (*nifH*, upper left), small-cell type *Crocosphaera* (upper right), large-cell type *Crocosphaera* (lower left) and the ratio of the two subtypes to each other (lower right). The data was plotted according to the distance between samples, with the X-axis as km along the cruise transect starting from the south most station (station 5). The station order from 0 to ~200 km is: 5, 2, 3, 4, 8, 7, 10, 12 (See Figure 1).

Figure 4. Depth profiles of *C. watsonii* and two sub-types at three stations in the S. Pacific



Depth profiles of qPCR gene copies of total *Crocosphaera* (*nifH*, blue diamonds), small-cell type *Crocosphaera* (UDPhydro, red circles), and large-cell type *Crocosphaera* (*capD*, green pyramids), at three different stations in the S. Pacific during the KM0703 cruise in March to April of 2007.

Chapter 5: Conclusions and future directions

It is well established that *Crocosphaera watsonii* are crucial members of the marine phytoplankton community and substantially contribute to marine biogeochemical cycling, principally via N_2-fixation. Through work using cultured isolates, researchers have also begun to understand the physiological and genetic characteristics of *Crocosphaera*, and have determined that *C. watsonii* strains have at least two distinct phenotypes with ecologically important differences. The research described in this dissertation builds upon and significantly expands the current understanding of *C. watsonii* genetics, physiology, and ecology.

Work done using the *Crocosphaera* WH8501 genome and large-insert environmental clones indicated the species may utilize a novel evolutionary paradigm that involves exceptionally high levels of DNA sequence conservation in combination with large-scale genome rearrangements mediated by abundant mobile genetic elements. The genomic comparison of two phenotypically different strains confirmed that this paradigm occurs throughout the genomes, and revealed that genome of one small-cell strain (WH8501) harbored many times the number of genomic transposase genes. That comparison also identified a number of genes unique to each strain, including a genomic fragment that was specific to the WH0003 (large-cell) genome, which encoded genes critical for the biosynthesis of EPS, a characteristic of that phenotype.

While many of the observed differences between the *C. watsonii* WH8501 and WH0003 genomes appeared likely to distinguish each phenotype, genomic sequence from additional strains would be needed to verify those conclusions. As such, the following study which compared six *Crocosphaera* genomes, three of each phenotype, was a logical extension of the first genomic comparison. The larger genomic comparison found an extremely high abundance of transposases in the *C. watsonii* WH8501 genome that was, in fact, unique to that genome, and not indicative of the small-cell phenotype. Other features of the WH8501 genome distinguished it from other small-cell strains, and suggested that WH8501 may not be appropriate for use as the type-strain to represent the species in future physiological studies.

The six-genome comparison also verified a number of findings from the original two genome comparison, such as the genome-wide paradigm of remarkably high DNA sequence conservation amidst abundant large-scale rearrangements. As a result of this pattern, much of all six *C. watsonii* genomes were essentially identical to each other, despite being isolated from across wide temporal and geographic distances. Some sets of genes were also found to be strain- or phenotype-specific, such as a large-cell specific region containing the EPS biosynthesis genes originally identified in the *C. watsonii* WH0003 genome. The small-cell strains have smaller genomes with corresponding loss of genetic capabilities, while the large-cell strains have larger genomes with type-specific functional capabilities that may make them better adapted to nutrient limitations, especially iron and phosphorus. In some respects, the *Crocosphaera* genomes resemble those of non-cyanobacterial facultative pathogens, which often harbor abundant mobile genetic elements and show evidence of genome decay. The accumulation of such deleterious genome features is typically associated with a reduced effective population size and the corresponding increased likelihood of these features becoming fixed within the population. As such, it may be that the small-cell strains have, or recently had a smaller effective population which resulted in the higher levels of genome decay in that type as well as the higher transposase abundance in WH8501. Future studies with strains of both types will be needed to verify both this hypothesis and the proposed adaptive differences between strains. In addition, further sequencing efforts of more *Crocosphaera* strains will be crucial to verify the unique nature of the WH8501 genome, as well as the proposed phenotype-specific features. Finally, as the relatively limited collection of isolated strains are unlikely to fully represent the diversity in natural populations, large-scale sequencing efforts should also be conducted using DNA from natural samples of *Crocosphaera* in order to potentially identify heretofore unknown genotypes and phenotypes.

Because the molecular assays that have been used to investigate natural *Crocosphaera* populations were designed for genes that have no genetic variation among known *C. watsonii* strains, the genome sequences included in this dissertation were instrumental in the development of molecular (i.e. qPCR) assays that could distinguish between the two phenotypes. Applying those newly developed assays to natural samples revealed, for the first time, relative abundance patterns of the two *Crocosphaera* types in the water column. Two sets of samples, from the North and

South Pacific, exhibited some similar patterns such as high abundances above 40 m followed by rapidly declining counts between 40 and 100 m. There was also a dominance of the small-cell types in the shallower (≤ 75 m) samples at both locations, while large-cells dominated the deeper (≥100 m) samples, which also had overall lower abundances. This may be a result of differences between the two types in light-level adaptation, sinking rates and/or cellular resilience. An intriguing difference found between sample sets was that, based on the fraction of gene copies found on the larger pore-size filter, the N. Pacific large-cell types appeared to aggregate to a greater extent than the S. Pacific large-cell population. This hints at the possibility of two different sub-types of large-cell *Crocosphaera*, one in each Ocean Basin. However, the difference could also be a result of changes in environmental conditions. The *Crocosphaera* abundance patterns observed in the initial data described here are only a first step in understanding the distributions of the two types, and the processes that control those distributions. Many additional locations, times, and conditions will need to be sampled and assayed for the *Crocosphaera* sub-types before basin-scale patterns are confidently identified. However, once those patterns are identified, they will be crucial for better estimating the global contribution of *Crocosphaera* to marine biogeochemical cycles.

Appendix 1: Additional analysis of WH0003 and WH8501 transposase ORFs

Methods

IS family distribution was compared between the two *C. watsonii* genomes using two methods: 1) the total abundance of each family in the two genomes, and 2) examining the relative contribution of each IS family to the total in each genome (Figure S1). For those comparisons, the IS families were assigned as described in the methods section of the main text. In addition, transposase genes in the WH8501 genome were grouped according to their IS family and plotted according to their position in the genome to investigate potential genomic patterns within families, or across all families (Figure S2). The genome positions are based on a concatenation of all WH8501 contigs as described in the methods section of the main text.

In order to identify potentially functional transposons in the most highly repeated IS family in the WH8501 genome (IS5 family ISCwa22_aa1, 283 copies), the upstream and downstream regions were examined. A BLASTn search (Altschul et al., 1990) of the 1,527 bp transposase gene plus 250 bp upstream and downstream of the ORF against the nucleotide sequence of the WH8501 genome revealed that there were 135 copies of the full ORF with additional conserved sequences flanking both ends in the genome. These 135 sequences were aligned using ClustalX Version 2.0 (Higgins and Sharp, 1988; Thompson et al., 1994; Thompson et al., 1997; Larkin et al., 2007) and showed high sequence conservation in the ORF as well as 144 bp upstream and 36 bp downstream (≥ 99% nucleotide ID, see Table S6, and Figures S3 and S4). The 19 bp inverted repeat with a single mismatch between the two sequences was identified at both ends of the alignment using "einverted", a program that is part of the EMBOSS package (Rice et al., 2000). A search using "etandem" (also part of the EMBOSS package (Rice et al., 2000)) with a minimum repeat length of 5bp found no direct repeats in the ORF or in the conserved flanking regions.

Results and Discussion

Because IS elements are typically over 700 bp (Mahillon and Chandler, 1998; Mahillon et al., 1999), and both strains had a similar percentage of transposase ORFs that were over 700 bp, IS sequences have undergone a similar amount of sequence degradation in both strains, despite the large difference in abundance of these genes. In order for transposases to be widely found in all domains of life (Aziz et al., 2010), there must be some selective pressure that enhances their integration into genomes. Indeed, a study showed evidence of positive selection in one IS family in the WH8501 genome (Mes and Doeleman, 2006). However, other research has shown that IS elements typically have negative impacts on prokaryotic cellular fitness due to the metabolic cost of expressing the genes and the high probability of gene inactivation upon IS insertion in gene-dense genomes (Touchon and Rocha, 2007). So, while the selective pressures on mobile elements in prokaryotes are complicated and not fully understood, it is possible that the much higher number of transposases in the WH8501 genome have resulted in a lower level of fitness when compared to the WH0003 strain.

Large differences were found between the two genomes in transposase and IS family abundances (as discussed in the main text, and seen in Table 3 and Figure S1). For example while WH0003 had a number of transposases that were similar to those which were highly replicated in WH8501, the family with the fewest replicated sequences in WH8501 (IS200/IS605) made up the majority of transposase genes in WH0003, and actually had fewer total representatives in the WH8501 genome (Figure S1). In addition to differences in the numbers of each IS family in the two genomes, the genomic distribution (i.e. positions in the genomes) of transposase genes, and each IS family was also quite different between the two *Crocosphaera* strains. Transposase genes were grouped by IS family and mapped onto the WH8501 proxy genome (concatenated contigs as described in the methods of the main text), the genes in all of the IS families were generally evenly distributed, but in some regions there were multiple transposases that were very close or adjacent to each other (Figure S2). This may partly be explained by the fact that many transposons insert at specific recognition sequences, so it is possible that a single recognition sequence could facilitate multiple insertion events resulting in a string of adjacent transposons. Also, because mobile elements are more likely to be maintained in regions of low selective pressure, such as duplicated genes, or redundant newly acquired laterally transferred sequences (Touchon and Rocha, 2007), those regions would also be more likely to accumulate multiple IS elements. Finally, some IS elements require two or more adjacent ORFs for activity (Mahillon and Chandler, 1998), so those ORFs would be expected to be found in pairs or groups. Despite such mechanisms that could result in uneven transposase distribution, most transposase ORFs were well distributed throughout the WH8501 genome (Figure S2). In contrast, the smaller number of transposase genes in the WH0003 genome were more unevenly distributed (data not shown) indicating that the transposases have either had a shorter residence time in that genome, or been less actively replicated. Other studies have observed that IS element abundance diverges much faster than the rest of the genome, such that no correlation can be found between IS abundance and evolutionary distance, even among closely related strains of prokaryotes (Touchon and Rocha, 2007). The very closely related *Crocosphaera* strains in this study with vastly different abundances of transposase genes seem to be an example of this phenomenon.

The dissimilar distributions and abundances of transposases and their IS families in the two *C. watsonii* genomes indicate that they are under different regulatory controls and/or evolutionary pressures. There is evidence that the some of the WH8501 transposase genes are regulated on a diel cycle (Figure 5), but additional experiments will be required to determine if the WH0003 transposase genes are similarly regulated. Coordination of transposase expression has been observed in other cyanobacteria (Labiosa et al., 2006), and is consistent with the fact that transposition is often modulated by or requires host factors such as the DnaA protein, DNA gyrase, chaperones, and the Dam protein (Mahillon and Chandler, 1998). The pattern of transposase expression is also consistent with a number of studies that have shown diel regulation of many metabolic and cellular processes in the *Crocosphaera* genome (Church et al., 2005; Mohr et al., 2010; Pennebaker et al., 2010; Shi et al., 2010). While more work is required to investigate the full range and activity of transposase genes in these strains, the available data suggest that transposases have been integrated with other cellular processes and are regulated on a diel cycle.

The most highly repeated IS family in the WH8501 genome, IS5 family ISCwa22_aa1, was examined to determine how many of its 283 copies appeared to be complete transposons, and therefore likely to be currently or recently involved in genome movement. Based on nucleotide sequence alignments, 135 copies of the transposase also shared >99% sequence identity upstream (144 bp) and downstream (36 bp) of the ORF (Figure S3 and S4 and Table S6). A 19bp inverted repeat with a single nucleotide mismatch was identified at the beginning of the upstream, and the end of the downstream conserved sequence, but no direct repeats were found in the gene or the flanking conserved regions (see methods above). The lack of direct repeats was not surprising because these sequences are most similar to the ISCwa2 family which is described in IS finder (Siguier et al., 2006) as having members without direct repeats. The presence of a conserved sequence containing inverted repeat flanking both ends of the ORF suggests that nearly 50% (135 of 283) of the genes coding for the most repeated transposase family are part of functional transposons. This indicates that the replication of this sequence is recent and ongoing, which is also indicated by the high genome copy number of the IS family.

Similar analysis of the other 49 IS families would be required to determine which of those families are also currently active, and to establish a correlation between genome copy number and proportion of functional transposons, but such extensive analyses are beyond the scope of this genome comparison.

References

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215**,** 403-410.

Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38**,** 4207-4217.

Church, M.J., Short, C.M., Jenkins, B.D., Karl, D.M., and Zehr, J.P. (2005). Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. *Appl. Environ. Microbiol.* 71**,** 5362-5370.

Higgins, D.G., and Sharp, P.M. (1988). CLUSTAL - a package for performing multiple sequence alignment on a microcomputer. *Gene* 73**,** 237-244.

Labiosa, R.G., Arrigo, K.R., Tu, C.J., Bhaya, D., Bay, S., Grossman, A.R., and Shrager, J. (2006). Examination of diel changes in global transcript accumulation in *Synechocystis* (cyanobacteria). *J. Phycol.* 42**,** 622-636.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., Mcwilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23**,** 2947-2948.

Mahillon, J., and Chandler, M. (1998). Insertion sequences. *Microbiol. Mol. Biol. Rev.* 62**,** 725-774.

Mahillon, J., Léonard, C., and Chandler, M. (1999). IS elements as constituents of bacterial genomes. *Res. Microbiol.* 150**,** 675-687.

Mes, T.H.M., and Doeleman, M. (2006). Positive selection on transposase genes of insertion sequences in the *Crocosphaera watsonii* genome. *J. Bacteriol.* 188**,** 7176-7185.

Mohr, W., Intermaggio, M.P., and Laroche, J. (2010). Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. *Environ. Microbiol.* 12**,** 412-421.

Pennebaker, K., Mackey, K.R.M., Smith, R.M., Williams, S.B., and Zehr, J.P. (2010). Diel cycling of DNA staining and *nifH* gene regulation in the unicellular cyanobacterium *Crocosphaera watsonii* strain WH 8501 (Cyanophyta). *Environ. Microbiol.* 12**,** 1001-1010.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends Genet.* 16**,** 276-277.

Shi, T., Ilikchyan, I., Rabouille, S., and Zehr, J.P. (2010). Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* 4**,** 621-632.

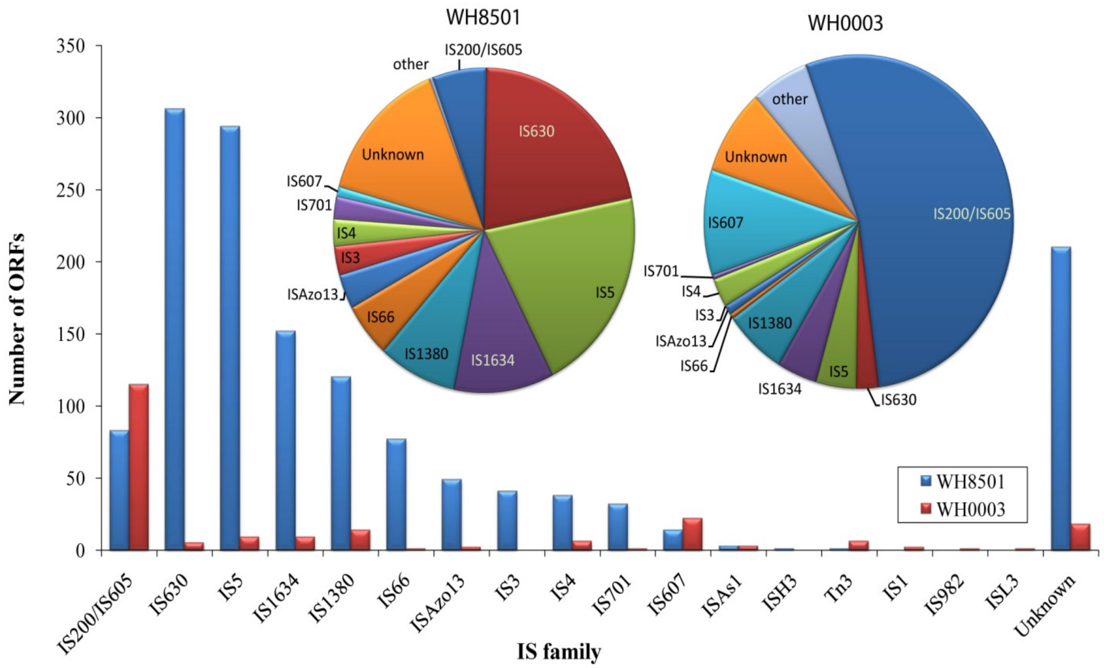
Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34**,** D32-36.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25**,** 4876-4882.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22**,** 4673-4680.

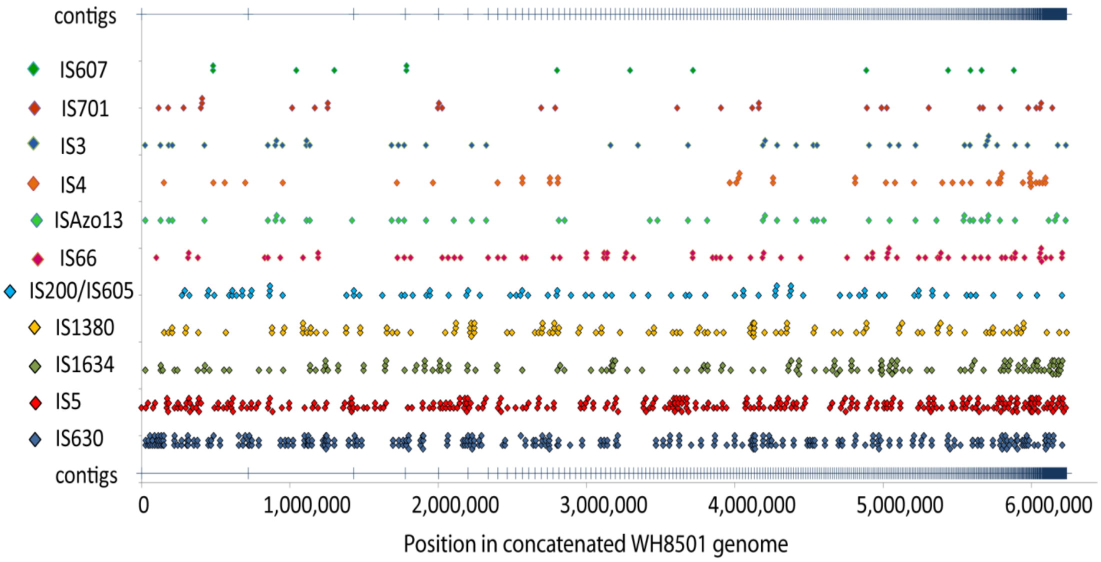
Touchon, M., and Rocha, E.P.C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24**,** 969-981.

Figure S1. Distribution of transposase IS families in both genomes.



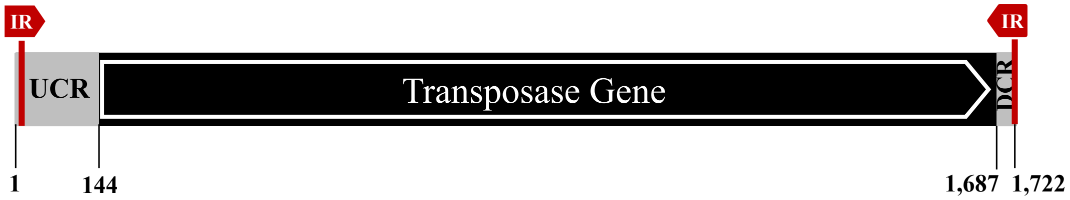
Transposase genes were categorized according to IS families using the IS finder tool. Bars represent the total number of genes in each IS family, and pie charts show the proportion of the total that each family represents within each genome.

Figure S2. Distribution of IS families in the *Crocosphaera* WH8501 genome.



Rows correspond to IS families, with some ORFs spread vertically to show symbols that would overlap due to dense spacing. Horizontal axis is the nucleotide position of the concatenated WH8501genome. The break points between contigs are indicated at the top and bottom of the figure.

Figure S3. Schematic representation of transposase gene and conserved flanking regions.



Proposed transposon region of 1,722 bp, including a transposase gene as indicated by a black background with a white arrow showing ORF direction. The 144 bp upstream conserved region (UCR) and 36 bp downstream conserved region (DCR) are shown in gray, and the two corresponding 19 bp inverted repeats (IR) are shown as red lines with red arrows above indicating the direction of each repeat.

Figure S4. Fasta formatted ORF and flanking conserved regions for the IS5 family ISCwa22_aa1 transposase.

```
> IS5 family ISCwa22_aa1 transposase alignment consensus sequence
cataattagggtttgctgaataaaagcaaaaccctattctttaaaaggttacacccatat
tcgcgatcgcaaaaaagatcagctttgtcggctacaaaccgctataattggtagaaacac
ctcccagttgttggtcaagaacaaATGTATCGAAAAGAGGAGCAACCTTTACCGCCCCCA
GAAAAATTTGAATTACCTTTCGAGGGCAAATTGTC-CCCCAATAATCGTTGGGTAATCAT
GGCAGAGTTGATACCTTGGGATGATTTTGAGGAAGAATATGCTAAACTTTTTT-CAGCAG
AAAAAGGTGCGCCAGCCAAACTATTTAGAATGGCATTAGGGACATTAATTATTAAGGAAA
AATTAGGAACAAGTGATAGAGAAACTATAGAACAGATTAGAGAAAATCCCTATCTACAAT
ACTTTATAGGTTTAAATTGTTATCAACAAGAGCAACCACTGGAATCTTCAATGCTAGTTC
ATTTTAGGAAAAGAATTGATAGAGAATTGATAAACAAAA-TCAATAAAAAAATAGTAAAA
AGAGAAATAGACAAGAGTGACAAAGAAGTAAAAAAAAA-GGATTGTCTCCAAGAAAAAGG
AGAAAAAATAAAAAATAAAGGTAAACTAATTTTAGATGCGACTTGTG-CGCT-AGCAGAC
ATCAAATACCCAACGGATTTAGGCATATTAAATCAAGCCAGAATTGAGACAGAAAGAATA
ATAGATAATC-TTTATAAACCATTAAGAGTAAAATTGAGAAAAAAGCCGAGAA-CTTCTA
GAAAAATTGCTAGAAAGGAATACTTAAAAGTAGCTAAAAAACGAAAATTATCTTATCAAG
AA-AGAAGAAAAGCTATCGGGAAACAGTTAAAATACCTTAAGAAAAATTTAGGAAATATA
GAAGACTTAATTCAAGCTGGGGCTTCC-CTGGAAAATCTGAGTAAAAGACAG-AAAAATT
GTCTAGAGACTA-TCAAAAAAGTTTATGAACAACAACAGTCAATGT-GGGAGAATAAGAC
CCAGAGTGTTCCTCAAAGAATTGTAAGTTTAACTCAACCGCATATTCGTCCAATAGTGAG
GGGAAAAGCAGGAAAACCAATAGAGTTTGGAGCTAAACTCTCAGTAAGCTGTGTAGATAA
CTATATGTTTCTAGACAAAATAAGTTGGGAAAACTTCAATGAATCTTATCATTTAAAAGA
ACAAGTAGAAAAGTATAAAGAAACGTTTGGCTATTATCCCGAATCCGTCCATGTAGATAA
AATTTATCGAACTAGGGAAAACAGAAATTGGTGTAAGGAAAGAGGGATAAGAATCAGTGG
TCCGAAGTTAGGAAGACCTCCGAAAAATGTCAGTGAAGAAGAAAAGAAAGAAGCCCACTC
CGATGAATGCTTCCGTAATGCTATTGAGGGAAAGTTTGGACAAGCTAAACGAAGATTTAG
CCTAAATCTCGTGATGACAAAACTCCCTGAAACCTCCATTACATCTATTGCTATTACATT
TTTAGTTGTCAATCTTTCTAAACTTCTGAGGCAGTTTTTGTCGCTTTTTTTGTCCTTATT
TACTAATAATAGAACA-GGGGAACTCATCAAACCGCCTTTCATTAATTTTGATTATACTT
TAAACAA-TTTTTGTGTACTAAAACTTATTGATTTGTCAGAAAATTCTTGGTCAAAAGTG
GCATAAattttggagaaactttttcagcaaaccctaatatag
```

Text colors are coordinated with figure S3 above; black upper-case text is the transposase coding sequence, and grey lower-case text is the conserved flanking sequence. Start and stop codons for the transposase gene are highlighted in green and red respectively. The 15 positions where a small number of sequences (usually one or two) had single base insertions are indicated with a dash. 19bp inverted repeat (IR) sequences are underlined and highlighted in maroon, with the single bp mismatch between the two IR indicated by yellow text.

Appendix 2: Chapter 2 supplemental tables

Table S2. WH8501 Transposase Isoform Counts

Transposase Isoform	**Number of ORFs**
Transposase IS5 family ISCwa22_aa1	283
Transposase IS5 family ISCwa21_aa1	8
Similar to Transposase IS5 family	3
IS5 total	*294*
Transposase IS630 family ISCwa33_aa1	142
Transposase IS630 family ISCwa35_aa1	127
Transposase IS630 family ISCwa29_aa1	15
Transposase IS630 family ISCwa30_aa2	5
Transposase IS630 family ISCwa31_aa1	3
Transposase IS630 family ISCwa28_aa1	2
Transposase IS630 family ISCwa32_aa2	2
Transposase IS630 family ISCwa34_aa1	2
Transposase IS630 family ISCwa27_aa1	1
Similar to Transposase IS630 family	7
IS630 total	*306*
Transposase IS1380 family ISCwa5_aa1	119
Similar to Transposase IS1380 family	1
IS1380 total	*120*
Transposase IS1634 family ISCwa1_aa1	142
Transposase IS1634 family ISCwa6_aa1	9
Similar to Transposase IS1634 family	1
IS1634 total	*152*
Transposase IS66 family ISCwa4_aa1	75
Similar to Transposase IS66 family	2
IS66 total	*77*
Transposase IS200/IS605 family ISCwa8_aa2	22
Transposase IS200/IS605 family ISCwa15_aa2	17
Transposase IS200/IS605 family ISCwa13_aa2	14
Transposase IS200/IS605 family ISCwa11_aa2	5
Transposase IS200/IS605 family ISCwa13_aa1	4
Transposase IS200/IS605 family ISCwa7_aa2	4
Transposase IS200/IS605 family ISCwa12_aa2	3
Transposase IS200/IS605 family ISCwa14_aa2	3
Transposase IS200/IS605 family ISCwa10_aa2	2
Transposase IS200/IS605 family ISCwa16_aa2	2
Transposase IS200/IS605 family ISCwa9_aa2	2

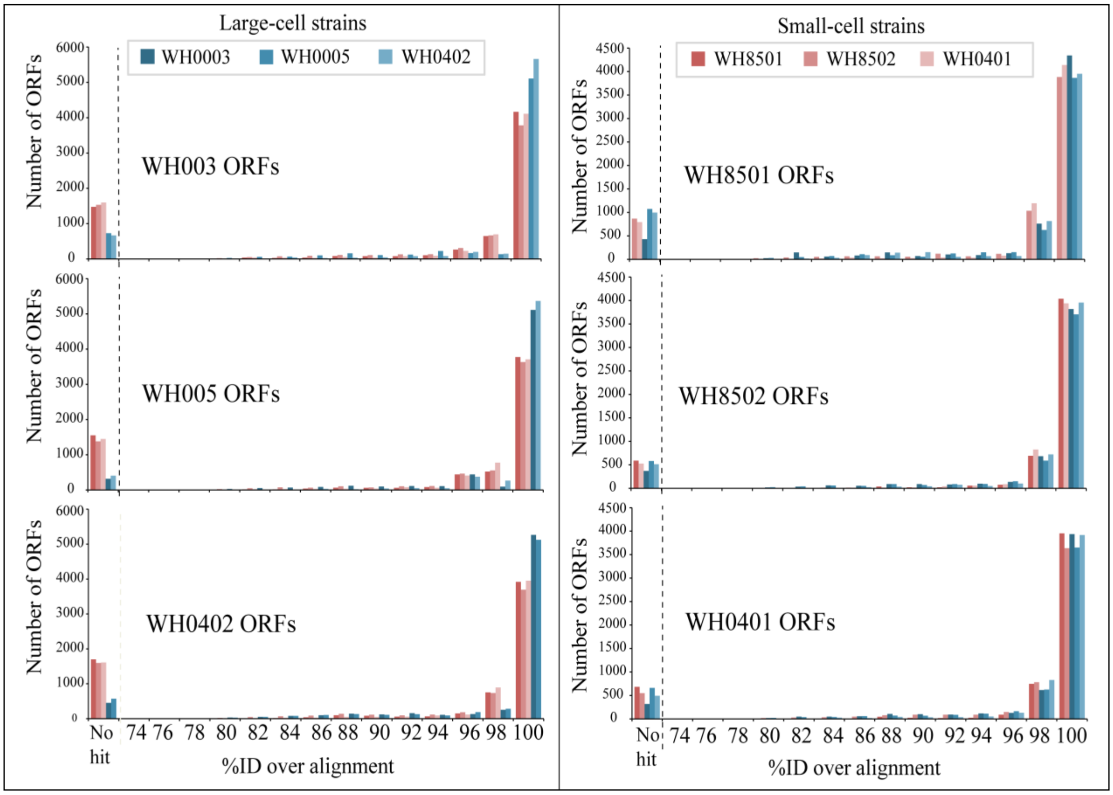
Similar to Transposase IS200/IS605 family	5
IS200/IS605 total	*83*
Transposase ISAzo13 family ISCwa38_aa1	49
Similar to Transposase IS3 family	41
Transposase IS4 family ISCwa19_aa1	24
Transposase IS4 family ISCwa18_aa1	10
Transposase IS4 family ISCwa17_aa1	1
Transposase IS4 family ISCwa20_aa1	1
Similar to Transposase IS4 family	2
IS4 total	*38*
Transposase IS607 family ISCwa26_aa2	7
Transposase IS607 family ISCwa24_aa2	4
Transposase IS607 family ISCwa23_aa2	1
Transposase IS607 family ISCwa25_aa2	1
Similar to Transposase IS607 family	1
IS607 total	*14*
Transposase IS701 family ISCwa37_aa1	29
Transposase IS701 family ISCwa36_aa1	3
IS701 total	*32*
Transposase ISAs1 family ISCwa3	1
Similar to Transposase ISAs1 family	2
ISAs1 total	*3*
Transposase ISH3 family ISCwa39_aa1	1
Transposase Tn3 family ISCwa40_aa1	1

Table S5. WH0003 ORFs in longest strain-specific region

WH0003 Locus Tag (genome accession #AESD01000000)	**ORF Length (bp)**	**Annotated function**
CWATWH0003_4887	314	DNA polymerase beta domain protein region
CWATWH0003_4888	356	hypothetical protein
CWATWH0003_4889	341	protein of unknown function DUF86
CWATWH0003_4890	2108	hypothetical protein
CWATWH0003_ 4891a4	1184	Primase 2
CWATWH0003_ 4891b4	348	hypothetical protein
CWATWH0003_4892	269	hypothetical protein
CWATWH0003_4893	1802	hypothetical protein
CWATWH0003_4894	1064	hypothetical protein
CWATWH0003_4895	707	hypothetical protein
CWATWH0003_4896	1583	hypothetical protein
CWATWH0003_4897	794	hypothetical protein
CWATWH0003_4898	1139	NAD/NADP transhydrogenase beta subunit
CWATWH0003_4899	737	hypothetical protein
CWATWH0003_4900	2714	hypothetical protein
CWATWH0003_4901	308	hypothetical protein
CWATWH0003_4902	626	hypothetical protein
CWATWH0003_4903	515	hypothetical protein
CWATWH0003_4904	878	mobilization protein BmgA
CWATWH0003_4905	329	hypothetical protein
CWATWH0003_4906	125	hypothetical protein
CWATWH0003_4907	1004	hypothetical protein
CWATWH0003_4908	446	hypothetical protein
CWATWH0003_4909	788	hypothetical protein
CWATWH0003_4910	743	hypothetical protein
CWATWH0003_4911	758	single-strand DNA-binding protein
CWATWH0003_4912	197	hypothetical protein
CWATWH0003_4913	350	hypothetical protein
CWATWH0003_4914	413	hypothetical protein
CWATWH0003_4915	407	Single-stranded DNA-binding protein
CWATWH0003_4916	347	hypothetical transcriptional regulator

<u>Appendix 3: Chapter 3 Supplemental tables and figures</u>

Figure S1. Percent identity bins of all ORFs in each genome vs other 5 genomes



Nucleotide BLAST similarity of ORFs in each of the six *C. watsonii* genomes as compared to the other five genomes. All ORFs in each genome were used as query sequnecs, and top BLAST hits were binned by % identity. The number of ORFs in each bin are represented by the height of the bars. The color of each bar, as listed in the legends, corresponds to the genome that was used as the BLAST reference sequence.

Table S2. LysM gene PCR primer sequences and product sizes

Primer Name	Primer Sequence	Product size with reverse primer (bp)
Reverse_all	5' CCT GTR YTR CCA TTT CDG C	n/a
LysM_Fwd_3	5' CCC TTA TAT CCC ACA CCA TAG	397
LysM_Fwd_4	5' CGA AGA AGC CAA CAG TAT TGT G	1401
LysM_Fwd_6	5' CTA AAC AGT GTC AGA AAT GAA AAA GCT G	518
LysM_Fwd_7	5' CTA GAG ACT GTC AAC GAA CAG	638

Table S3. ORF IDs and lengths for the 4 LysM gene forms in 6 *C. watsonii* genomes and PCR results [(+) or (–) amplification] for 2 additional strains.

***C. watsonii* strain**	**Pheno-type**	**LysM_3 ORFs** (600 bp)	**LysM_4 ORFs** (2061 bp)	**LysM_6 ORFs** (1185 bp)	**LysM_7 ORFs** (1185 bp)
WH0402	*large cell*	*WH0402_3003* *WH0402_3004*	*WH0402_1788*	*WH0402_0610*	*WH0402_1003*
WH0003	*large cell*	*WH0003 _0900*	*WH0003_3214*	*WH0003_5014*	*WH0003_5018*
WH0005	*large cell*	*WH0005_2591*	*WH0005_4551*	*WH0005_5028 (truncated at end of contig)*	*767 bp on contig1683 (not annotated as ORF)*
WH0004	*large cell*	***(+)** PCR* *no genome*	***(+)** PCR* *no genome*	***(+)** PCR* *no genome*	***(+)** PCR* *no genome*
WH0401	**small cell**	WH0401_0089	WH0401_4464	WH0401_1496	**none**
WH8502	**small cell**	WH8502_1480	WH8502_4505	WH8502_1556	**none**
WH8501	**small cell**	CwatDRAFT _5403	CwatDRAFT _2860	CwatDRAFT _0871	**none**
WH0002	**small cell**	**(+)** PCR no genome	**(+)** PCR no genome	**(+)** PCR no genome	**(-) PCR** **no genome**

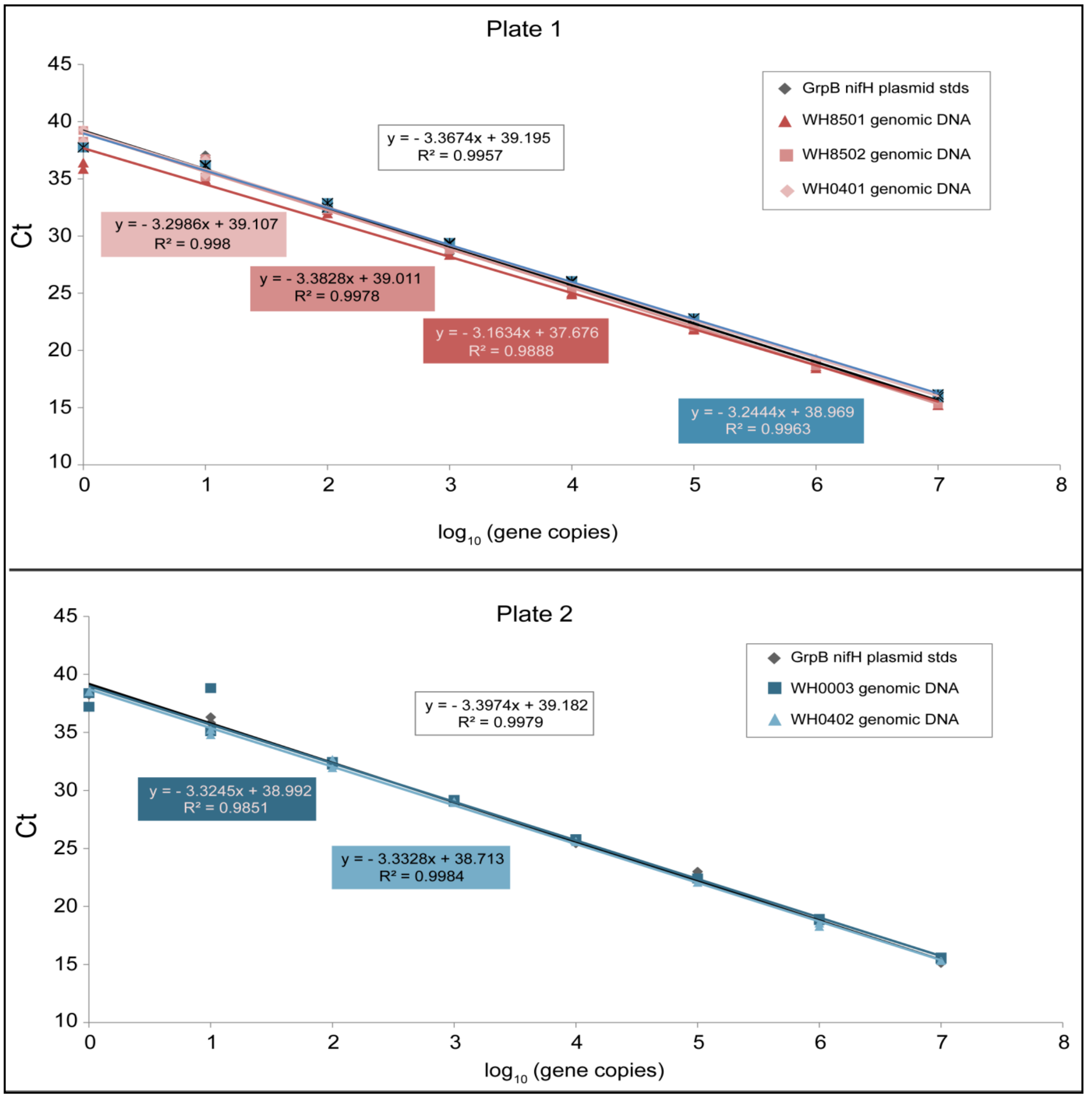
Table S5. Counts of sequences in each category, based on genomes in which the sequences was found.

total count	large-cell count	small-cell count	WH0402	WH0005	WH0003	WH0401	WH8502	WH8501	count of category	expected count	difference from expected	% difference from expected
2	2	0	0	1	1	0	0	0	175	115.1	59.9	52.0%
2	2	0	1	0	1	0	0	0	268	115.1	152.9	132.8%
2	2	0	1	1	0	0	0	0	308	115.1	192.9	167.5%
2	1	1	0	0	1	1	0	0	30	115.1	-85.1	-73.9%
2	1	1	0	0	1	0	0	1	144	115.1	28.9	25.1%
2	1	1	0	0	1	0	1	0	15	115.1	-100.1	-87.0%
2	1	1	0	1	0	1	0	0	18	115.1	-97.1	-84.4%
2	1	1	0	1	0	0	0	1	101	115.1	-14.1	-12.3%
2	1	1	0	1	0	0	1	0	5	115.1	-110.1	-95.7%
2	1	1	1	0	0	1	0	0	29	115.1	-86.1	-74.8%
2	1	1	1	0	0	0	0	1	95	115.1	-20.1	-17.5%
2	1	1	1	0	0	0	1	0	37	115.1	-78.1	-67.9%
2	0	2	0	0	0	1	1	0	69	115.1	-46.1	-40.1%
2	0	2	0	0	0	0	1	1	166	115.1	50.9	44.2%
2	0	2	0	0	0	1	0	1	267	115.1	151.9	131.9%
3	3	0	1	1	1	0	0	0	781	92.2	688.9	747.5%
3	2	1	0	1	1	1	0	0	10	92.2	-82.2	-89.1%
3	2	1	0	1	1	0	0	1	35	92.2	-57.2	-62.0%
3	2	1	0	1	1	0	1	0	13	92.2	-79.2	-85.9%
3	2	1	1	0	1	1	0	0	90	92.2	-2.2	-2.3%
3	2	1	1	0	1	0	0	1	78	92.2	-14.2	-15.4%
3	2	1	1	0	1	0	1	0	13	92.2	-79.2	-85.9%
3	2	1	1	1	0	1	0	0	85	92.2	-7.2	-7.8%
3	2	1	1	1	0	0	0	1	64	92.2	-28.2	-30.5%
3	2	1	1	1	0	0	1	0	23	92.2	-69.2	-75.0%
3	1	2	0	0	1	1	0	1	67	92.2	-25.2	-27.3%
3	1	2	0	1	0	1	0	1	91	92.2	-1.2	-1.2%
3	1	2	1	0	0	1	0	1	93	92.2	0.8	0.9%
3	1	2	0	0	1	1	1	0	20	92.2	-72.2	-78.3%
3	1	2	0	1	0	1	1	0	16	92.2	-76.2	-82.6%
3	1	2	1	0	0	1	1	0	76	92.2	-16.2	-17.5%

3	1	2	0	0	1	0	1	1	35	92.2	-57.2	-62.0%
3	1	2	0	1	0	0	1	1	34	92.2	-58.2	-63.1%
3	1	2	1	0	0	0	1	1	66	92.2	-26.2	-28.4%
3	0	3	0	0	0	1	1	1	153	92.2	60.9	66.0%
4	3	1	1	1	1	1	0	0	103	107.1	-4.1	-3.8%
4	3	1	1	1	1	0	0	1	142	107.1	34.9	32.6%
4	3	1	1	1	1	0	1	0	78	107.1	-29.1	-27.1%
4	2	2	0	1	1	1	0	1	29	107.1	-78.1	-72.9%
4	2	2	0	1	1	1	1	0	31	107.1	-76.1	-71.0%
4	2	2	0	1	1	0	1	1	60	107.1	-47.1	-44.0%
4	2	2	1	0	1	1	0	1	367	107.1	259.9	242.8%
4	2	2	1	0	1	1	1	0	39	107.1	-68.1	-63.6%
4	2	2	1	0	1	0	1	1	41	107.1	-66.1	-61.7%
4	2	2	1	1	0	1	0	1	201	107.1	93.9	87.7%
4	2	2	1	1	0	1	1	0	39	107.1	-68.1	-63.6%
4	2	2	1	1	0	0	1	1	25	107.1	-82.1	-76.7%
4	1	3	0	0	1	1	1	1	226	107.1	118.9	111.1%
4	1	3	0	1	0	1	1	1	22	107.1	-85.1	-79.5%
4	1	3	1	0	0	1	1	1	203	107.1	95.9	89.6%
5	2	3	1	1	0	1	1	1	98	244.0	-146.0	-59.8%
5	2	3	1	0	1	1	1	1	272	244.0	28.0	11.5%
5	2	3	0	1	1	1	1	1	223	244.0	-21.0	-8.6%
5	3	2	1	1	1	0	1	1	212	244.0	-32.0	-13.1%
5	3	2	1	1	1	1	1	0	331	244.0	87.0	35.7%
5	3	2	1	1	1	1	0	1	328	244.0	84.0	34.4%
1	1	0	0	0	1	0	0	0	223	195.0	28.0	14.4%
1	1	0	0	1	0	0	0	0	167	195.0	-28.0	-14.4%
1	1	0	1	0	0	0	0	0	315	195.0	120.0	61.5%
1	0	1	0	0	0	1	0	0	132	195.0	-63.0	-32.3%
1	0	1	0	0	0	0	0	1	229	195.0	34.0	17.4%
1	0	1	0	0	0	0	1	0	104	195.0	-91.0	-46.7%
6	3	3	1	1	1	1	1	1	3825	n/a	n/a	n/a

Appendix 4: Chapter 4 supplemental tables and figures

Figure S1. qPCR Standard curves of *nifH* linearized plasmids and genomic DNA from *C. watsonii* strains using *nifH* primer-probe set.



Genomic DNA from *C. watsonii* strains was amplified (see methods) and copies calculated according to DNA concentration and genome size. Sets of standards were prepared in triplicate for each strain from 10^0 to 10^7 genome copies, and qPCR amplification was carried out. C_t's were compared to triplicate reactions of the well-established *nifH* linearized plasmids with the same number of copies per reaction that were run on the same plate.

Table S1. Flow cytometry (FCM) counts of *Crocosphaera* cells compared to *nifH* gene copies (qPCR) for samples in the North Pacific.

station	**Depth (m)**	***nifH* gene copies/ liter**	**FCM Croco cells/liter**	**ratio of FCM count to *nifH* copies**
2	5	1.83E+06	1.45E+05	0.08
2	25	7.31E+05	2.51E+05	0.34
2	45	1.64E+06	2.43E+05	0.15
3	5	1.85E+06	1.50E+05	0.08
3	25	5.52E+05	1.97E+05	0.36
3	45	7.70E+05	1.70E+05	0.22
3	75	4.09E+05	5.76E+04	0.14
4	5	5.06E+05	6.01E+04	0.12
4	25	2.77E+05	2.95E+04	0.11
4	45	3.92E+05	5.86E+04	0.15
4	75	4.62E+05	8.86E+04	0.19
5	5	1.96E+06	2.01E+05	0.10
5	25	3.00E+06	2.25E+05	0.08
5	45	1.41E+06	9.78E+04	0.07
7	5	1.26E+06	2.09E+05	0.17
7	25	3.72E+06	5.00E+05	0.13
7	45	1.55E+06	3.16E+05	0.20
7	75	2.31E+05	2.27E+04	0.10
8	5	1.71E+06	3.52E+05	0.21
8	25	2.84E+06	4.66E+05	0.16
8	45	2.10E+06	4.73E+05	0.23
12	5	9.79E+05	2.18E+05	0.22
12	25	1.07E+06	1.32E+05	0.12
12	45	1.27E+03	1.09E+05	86.12

Table S2. Flow cytometry (FCM) counts of *Crocosphaera* cells compared to *nifH* gene copies (qPCR) for samples in the South Pacific.

station	**depth (m)**	**capD (large-cell) % on 0.2um filter**	**UDP (small-cell) % on 0.2um filter**
6	15	91.5%	98.1%
6	30	91.7%	98.1%
21	5	35.8%	79.5%
21	15	90.1%	99.1%
21	30	98.4%	98.2%
21	50	91.5%	91.5%
21	75	80.3%	99.0%
21	95	88.3%	98.4%
21	150	91.4%	99.1%
22	5	95.4%	98.8%
22	15	96.0%	0.0%
22	30	95.8%	99.7%
22	50	95.9%	98.7%
22	75	94.5%	97.9%
22	100	59.3%	85.1%
22	135	98.5%	97.3%
22	175	85.8%	87.6%
23	5	99.2%	99.9%
23	15	99.0%	99.7%
23	30	98.4%	99.9%
23	50	97.7%	99.3%
23	75	97.2%	99.7%
23	100	100.0%	100.0%
23	128	68.7%	96.9%
23	150	91.7%	94.9%

List of additional supplementary files

Chapter 2 Table S1.	WH8501 Re-annotated Transposase ORFs
Chapter 2 Table S3.	WH8501 strain-specific ORFs
Chapter 2 Table S4.	WH0003 strain-specific ORFs
Chapter 2 Table S6.	WH8501 sequences included in transposase alignment and % ID of each to alignment consensus sequence shown in Fig. S4
Chapter 3 Table S1.	ORF IDs for sequences included in alignment of 25 genes for phylogenetic analysis
Chapter 3 Table S4.	Strain-specific ORFs from six *C. watsonii* genomes
Chapter 3 Table S6.	Phenotype-specific ORF IDs and functions from six *C. watsonii* genomes