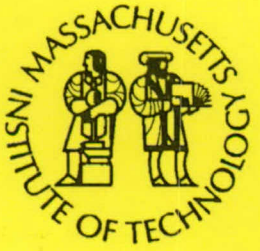
MIT/WHOI 2004-09

Massachusetts Institute of Technology
Woods Hole Oceanographic Institution



Joint Program in Oceanography/ Applied Ocean Science and Engineering

DOCTORAL DISSERTATION

Ecology, Diversity and Comparative Genomics of Oceanic Cyanobacterial Viruses

by

Matthew B. Sullivan

June 2004

MIT/WHOI
2004-09

Ecology, Diversity and Comparative Genomics of Oceanic Cyanobacterial Viruses

by

Matthew B. Sullivan

Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

and

Woods Hole Oceanographic Institution
Woods Hole, Massachusetts 02543

June 2004

DOCTORAL DISSERTATION

Funding was provided by National Science Foundation grant OCE-9820035 and DOE GTL grant DE-FG02-02ER63445 to S.W. Chisholm. Support was also available by an Education Fellowship from the Woods Hole Oceanographic Institution, NASA Grant NAG5-7538 to Heidi Sosik, The Seaver Foundation, NIH Genome training grant 7-T32-HG0039-05 and Anonymous YS funds from the Massachusetts Institute of Technology.

Approved for publication; distribution unlimited.

Approved for Distribution:

John J. Stegeman, Chair
Department of Biology

Paola Malanotte-Rizzoli
MIT Director of Joint Program

John W. Farrington
WHOI Dean of Graduate Studies

ECOLOGY, DIVERSITY AND COMPARATIVE GENOMICS OF OCEANIC CYANOBACTERIAL VIRUSES

by

Matthew Brian Sullivan

B.S. Marine Biology, LIU/Southampton College 1997
M.Phil. Biology, Queens University of Belfast, Northern Ireland 1998
M.S. Biology MIT/WHOI 2001

Submitted in partial fulfillment of the requirements for the degree of

PhD. Biology

at the

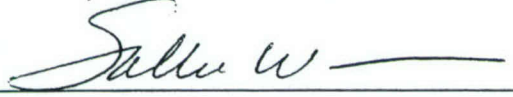
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

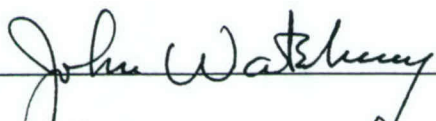
and the

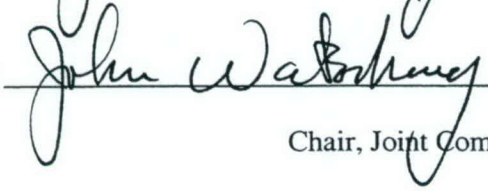
WOODS HOLE OCEANOGRAPHIC INSTITUTION

June 2004

Signature of Author 
Joint Program in Biological Oceanography
MIT/WHOI

Certified by 
Sallie W. Chisholm
Co-Thesis Supervisor, MIT

Certified by 
John Waterbury
Co-Thesis Supervisor, WHOI

Accepted by 
John Waterbury
Chair, Joint Committee for Biological Oceanography, WHOI

ECOLOGY, DIVERSITY AND COMPARATIVE GENOMICS OF OCEANIC CYANOBACTERIAL VIRUSES

By

Matthew Brian Sullivan

Submitted to the Department of Biology in April, 2004 in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

The marine cyanobacteria *Prochlorococcus* and *Synechococcus* are numerically dominant primary producers in the oceans. Each genera consists of multiple physiologically and genetically distinct groups (termed "ecotypes" in *Prochlorococcus*). Cyanobacterial viruses (cyanophages) that infect *Synechococcus* are abundant (to 10^4-10^6 phage ml^{-1}) in the oceans and calculations suggest that they play a small but significant role in host mortality. Cyanophages are also thought to shape their host populations through regulation of sub-populations and through transfer of genes.

Here we describe the isolation of *Prochlorococcus* cyanophages and the assembly of a culture collection established using a broadly diverse suite of *Prochlorococcus* and *Synechococcus* hosts. The collection contains three morphological families, *Myoviridae, Podoviridae* and *Siphoviridae,* known to infect marine bacteria and cyanobacteria. Host strains of similar ecotypes often yielded cyanophages of the same family. Host-range analyses of these isolates demonstrated varying levels of specificity among the different morphological types, ranging from infection of a single strain to infection across ecotypes and even across both cyanobacterial genera. Strain-specific cyanophage titers were low in open ocean waters where total cyanobacterial abundances were high, suggesting low phage titers might be a feature of open oceans. Investigations of the underlying cause(s) of this trend require culture-independent assays for quantifying phage that infect particular hosts. We used the phage g20 gene, which encodes the portal protein, to examine the diversity of *Myoviridae* isolates and found that g20 sequences from our isolates had high similarity to those from other cultured isolates, but not to six phylogenetic clusters of environmental g20 sequences that lacked cultured representatives.

Three *Prochlorococcus* cyanophage genomes were sequenced and analysis of these genomes show striking similarity to the well-studied T7- and T4-like phages, but additionally suggest that these *Prochlorococcus* cyanophages are modified for infection of photosynthetic hosts, that live in nutrient-limited environments. All three cyanophage genomes contain, among other novel genes of interest, photosynthetic genes that are full-length, conserved, and clustered in the genome suggesting they are functional during infection. Phylogenetic inference suggests that some of these genes were horizontally transferred between host and phage influencing the evolution and ecology of both host and phage.

This thesis was co-supervised by: Sallie W. Chisholm (MIT Professor of Civil and Environmental Engineering and Biology) and John B. Waterbury (WHOI Professor of Biology)

ACKNOWLEDGEMENTS

As the stories in this thesis grew, the efforts for uncovering and understanding the meaning of these stories became more and more collaborative. Like a good developing graduate student, I began learning just what I knew and what I did not know. Such realization of my intellectual (and time) limits led to the involvement of many friends and colleagues in the chapters that follow.

First, I thank my thesis co-advisers, John Waterbury and Penny Chisholm, committee members Jonathan King and Martin Polz, and my unofficial adviser Forest Rohwer. John took me in at a very early and ignorant stage to take advantage of the finer points of a summer in Woods Hole. John taught me to begin to think and to question. And most importantly he taught me to grow and love the cyanobacterial hosts and their phages that would become the basis of my thesis. Penny had the much un-enviable task of shaping my sporadically focused excitement and youthful enthusiasm towards a finished product. Think, read, think. What is the question? I'm still learning, still practicing, but I am making progress towards understanding how to understand. Martin has challenged me to get more out of my phylogenetic analyses and to think deeper about the details. Jonathan has offered countless insights and suggestions from his myriad experiences from many years of thinking about phages. Forest took me in as an 'afraid of anything molecular' student and molded my mind and my perspectives to realize just how valuable molecular tools could be for answering questions I had about my phages in the oceans. Forest has also set me on the right path to getting clone libraries made by David Mead at Lucigen that has enabled the sequencing of my cyanophage genomes as seemless as possible.

Second, a graduate thesis is a trying and testing time full of challenges at a professional and personal level. I've been fortunate to be surrounded by some fellow graduate students, post-docs, and administrators that made life easier. Vanja Klepac-Ceraj, Frederic Chagnon, Zack Johnson, Erik Zinser, Gabrielle Rocap, Mak Saito, Pat Friel, John Doll, Ken Olive and Arvind Govindarajan, have put the most time in mentoring me as both a colleague and a friend. Their comments and criticisms over the years have been invaluable. Two superb undergraduate students, Grace Tan and Jessica Rosenkrantz, have contributed significantly to the development of the g20 manuscript. Sheila Frankel and Vicki Murphy have made various administrative and other things significantly easier within the confines of Parsons Lab. Finally, Virginia Rich has single-handedly made it her quest to help me graduate faster. During this last, recklessly fast-paced year, she has provided a very critical and objective mind for polishing everything from ideas to manuscripts.

Third, I would like to thank the colleagues who have contributed invaluably to manuscripts and story development. PhIG (Phage Interest Group) and committee member Jonathan King were a valuable sounding board for realistic interpretations in the context of phage biology. Debbie Lindell shaped the way our photosynthetic genes in cyanophage manuscript developed and offered one-on-one criticisms, brainstorming and open-mindedness to some relatively crazy ideas as well as reinforced that so often the devil really is in the details. David Mead (Lucigen), Jane Grimwood (Stanford finishing group) and Jan-Fang Cheng (DOE JGI) among others are responsible for administering, sequencing and finishing the three cyanophage genomes. Peter Weigele and Maureen Coleman have been a breath of fresh air in the writing of the cyanophage genomics manuscript. Peter leapt at the opportunity to dive into a deeper understanding of phage tail fiber genes in the cyanophage genomics work and his efforts inspire

future development of a more meaningful index of phage diversity. Maureen responded equally excitedly about building a bioinformatics pipeline for going from phage sequence to viewable annotation and her scripts will prove useful not only to my work but to much of the work coming out of this lab for a while.

Finally, my family has offered support like no other. Mary and Terry Sullivan are perhaps the finest parental units on the planet, while my brother Jeff pushes me up the next mountain in our hiking so that I can think of obstacles in my thesis simply as the next mountain.

I shudder to think what this thesis would be without the help of so many people. Thanks.

This research was supported by an NSF grant OCE-9820035 and DOE GTL grant DE-FG02-02ER63445 to S.W. Chisholm. Matthew Sullivan was supported by an Education Fellowship from the Woods Hole Oceanographic Institution, NASA Grant NAG5-7538 to Heidi Sosik, Anonymous YS funds from the Massachusetts Institute of Technology, The Seaver Foundation and an NIH Genome training grant 7-T32-HG0039-05 and the NSF grant cited above.

TABLE OF CONTENTS

Chapter I

Introduction

CHAPTER I. INTRODUCTION

The bacterial viruses (phages) contained in the oceans are the most abundant "biological entities" on the planet (Hendrix et al., 1999). We are interested in understanding the interplay between marine phages and their hosts and how this interplay affects the physiological differentiation and evolution of each. In this thesis, we isolated and examined cyanophage whose hosts are the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. These cyanobacteria are the numerically dominant primary producers (Partensky, Hess, and Vaulot, 1999; Waterbury et al., 1986) responsible for 32-89% of primary production in the vast surface oligotrophic oceans thus contributing to carbon fixation on a global scale (Goericke, 1993; Li, 1994; Liu, Campbell, and Landry, 1995; Liu et al., 1998; Liu and Landry, 1999; Liu, Nolla, and Campbell, 1997; Veldhuis et al., 1997).

Early culture-based studies using terrestrial microbes as hosts suggested that phages were rare in the oceans (ZoBell, 1946). It was not until appropriate marine host strains were used that indigenous marine phages were clearly demonstrated (Spencer, 1955; Spencer, 1960; Spencer, 1963). Later, electron microscopy (EM) of concentrated seawater samples revealed virus-like particles (VLPs) occurring at concentrations up to 10^8 VLPs ml^{-1} in aquatic environments (Bergh, 1989; Borsheim, Bratbak, and Heldal, 1990; Bratbak, 1990; Bratbak et al., 1992; Heldal and Bratbak, 1991; Proctor and Fuhrman, 1990; Sieburth, Johnson, and Hargraves, 1988; Torella and Morita, 1979). Subsequent field studies using EM and epifluorescence microscopy showed that VLP abundances were approximately an order of magnitude higher than total prokaryote counts in seawater – leading to the suggestion that phage may be a significant source of host mortality in marine systems (Bratbak, 1990; Bratbak, 1993; Cottrell and Suttle, 1991; Moebus, 1987; Nagasaki, Tarutani, and Yamaguchi, 1999; Proctor and Fuhrman, 1990; Suttle, 1999; Wommack et al., 1992).

To begin to understand phage-host dynamics and to extrapolate ecological significance, model marine phage-host systems are critical. Among the few marine microbes that have been successfully cultured (Fuhrman and Campbell, 1998), *Prochlorococcus* and *Synechococcus* are not only ecologically important genera (Partensky, Hess, and Vaulot, 1999; Waterbury et al., 1986), but are also unique in having many (80+) strains now in culture. The genetic diversity of *Prochlorococcus* and *Synechococcus* cultured isolates has been well characterized using multiple molecular markers and all suggest similar phylogenetic topologies (Ferris and Palenik, 1998; Fuller et al., 2003; Moore, Rocap, and Chisholm, 1998; Rocap et al., 2002; Zeidner et al., 2003).

Among *Prochlorococcus,* laboratory work with cultured isolates shows that they form at least two physiologically (Moore and Chisholm, 1999; Moore, Goericke, and Chisholm, 1995; Moore, Rocap, and Chisholm, 1998) and genetically (Moore, Rocap, and Chisholm, 1998; Rocap et al., 2002) distinct groups which are termed ecotypes. At this broad ecotype level, ribosomal DNA genetic differences are reflected in copper sensitivity (Mann et al., 2002), nutrient utilization (Moore et al., 2002) and genetic composition of the genome (Dufresne et al., 2003; Rocap et al., 2003). It has been suggested that cyanophage may be partly responsible for this diversity through the regulation of subpopulations of these cyanobacteria as well as through horizontal gene transfer (Marston and Sallee, 2003; Suttle and Chan, 1994; Waterbury and Valois, 1993).

Building a diverse cyanophage collection

Because of the breadth of known genetic diversity among these marine cyanobacteria (Fig. 1), we were interested in developing a cyanophage collection using hosts that represented the rDNA-based genetic diversity known at the time (Rocap et al., 2002). We note that even in these well-studied marine cyanobacteria, culture collections may not yet represent their naturally occurring genetic and physiological diversity (Ahlgren, Hook, and Rocap, 2004; Fuller et al., 2003; Rocap, McKay, and Ahlgren, 2004), but data from Q-PCR analyses from seasonal sampling at the Bermuda Atlantic Time Series and Hawaii Ocean Time series suggests we often come close (E. Zinser, Unpubl. Results).

Synechococcus cyanophage have been isolated using a number of host strains, but primarily focusing on strain WH7803 (Table 1) (Fuller et al., 1998; Lu, Chen, and Hodson, 2001; Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993). *Synechococcus* cyanophage isolates belong to the three phage morphological familes (*Myoviridae, Podoviridae* and *Siphoviridae*) that have been identified for freshwater cyanobacteria (Safferman et al., 1983), marine bacteria (Moebus, 1987; Wichels et al., 1998) and terrestrial bacteria (Bradley, 1967; Murphy et al., 1995). In this thesis, we created a cyanophage collection by isolating cyanophages using a diverse group of *Prochlorococcus* and *Synechococcus* strains as hosts. We characterized these isolates using host range, morphology, a family specific gene marker, and genomes.

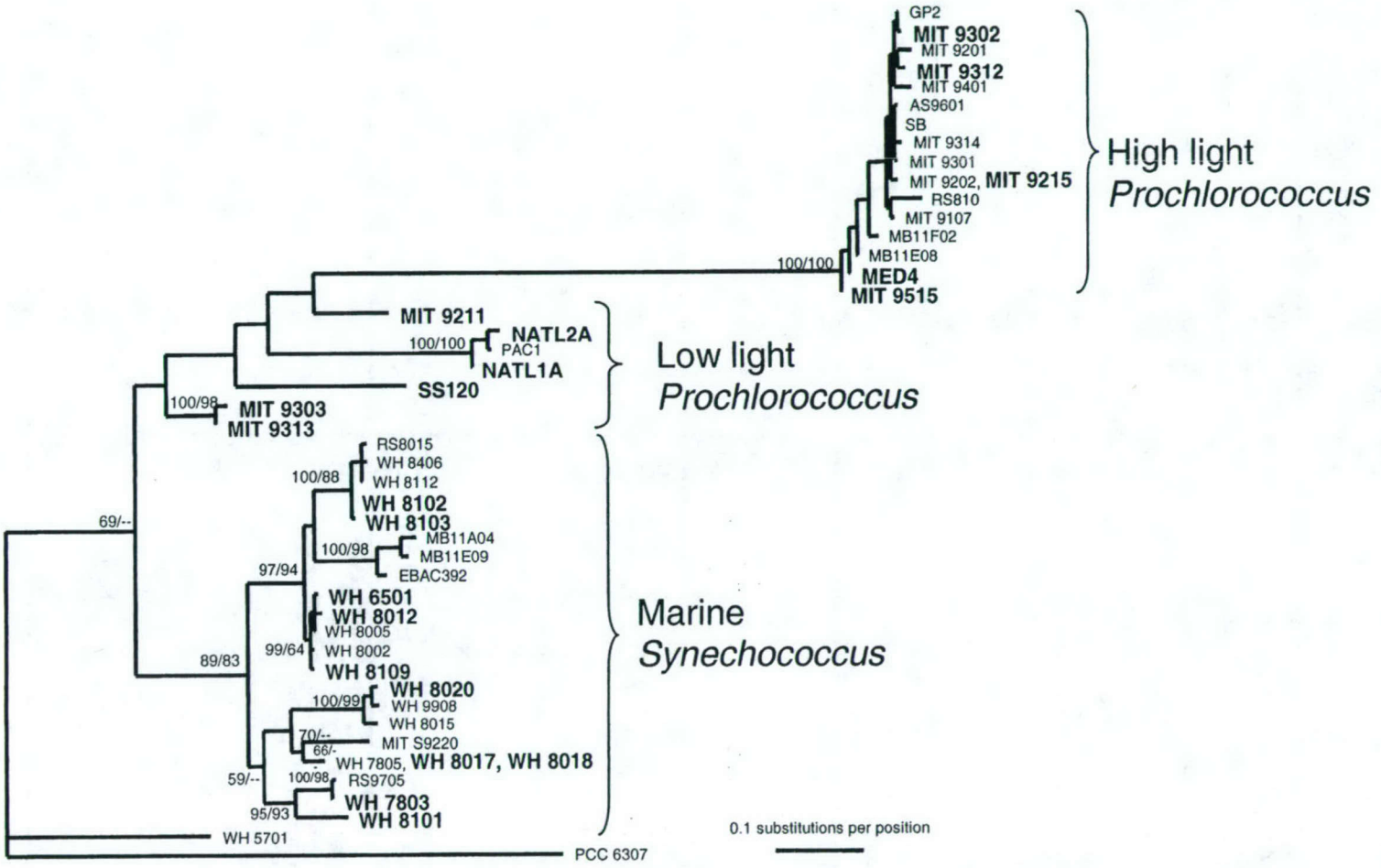


Figure 1: Evolutionary relationships of cyanobacterial isolates and environmental sequences from Monterey Bay (designated by MB and a number or EBAC) inferred by maximum likelihood analysis using 233 positions of the 16S-23S rDNA spacer. Strains used during this thesis for host range analyses and cyanophage isolations are in designated in bold and larger print. The high light and low light ecotype designations for *Prochlorococcus* are based upon physiological work done on select laboratory isolates examining growth rate and chlorophyll b/a ratios for cells grown over various light intensities. Reprinted with permission from (Moore et al., 2002).

Table 1: Details of published cyanophage isolates isolated using strains of marine *Synechococcus* spp. M/P/S represent this phage isolate belongs to the family *Myoviridae, Podoviridae* or *Siphoviridae,* respectively. Morphological measurements and location information are included where available.

Cyanophage	Host Strain	Location	M/P/S?	Head (nm)	Tail (nm)	Reference
S-BM1	WH7803	Bermuda, BBSR	M	88	57 x 33	Wilson et al. 1993
S-PM1	WH7803	Plymouth Sound (50°18'N, 4°12'W)	M	88	57 x 33	Wilson et al. 1993
S-PM2 (formerly S-PS1)	WH7803	Plymouth Sound (50°18'N, 4°12'W)	M	90	165 x 20	Wilson et al. 1993
S-WHM1	WH7803	WHOI	M	88	108 x 23	Wilson et al. 1993
S-BM2 (formerly S-BS1)	WH7803	Bermuda, BBSR	M	90	165 x 20	Wilson et al. 1993
S-BM3	WH7803	Hydrostation S (30°10'N, 64°30'W)	M	110	230 x 25	Fuller et al. 1998
S-BM4	WH7803	West coast of Bermuda	M	125	190 x 35	Fuller et al. 1998
S-BM5	WH7803	West coast of Bermuda	M	125	190 x 35	Fuller et al. 1998
S-BM6	WH7803	Hydrostation S (30°10'N, 64°30'W)	M	110	200 x 25	Fuller et al. 1998
S-BP1	WH8018	West coast of Bermuda	P	125		Fuller et al. 1998
S-BP2	WH8018	West coast of Bermuda	P	75		Fuller et al. 1998
S-BP3	WH7803	West coast of Bermuda	P	70		Fuller et al. 1998
S-BnM1	WH7803	Bergen, Norway	M			Wilson PhD thesis 1994
S-RSM1	WH7803	Red Sea, Eilat, Israel	M			Wilson PhD thesis 1994
S-RSM2	WH7803	Red Sea, Eilat, Israel	M			Wilson PhD thesis 1994
S-MM1	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-MM2	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-MM3	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-MM4	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-MM5	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-MM7	WH7803	Miami, FL	M			Wilson PhD thesis 1994
S-PWM1	WH7803		M			Suttle & Chan 1993
S-PWM2	WH7803		M			Suttle & Chan 1993
S-PWM3			M			Suttle & Chan 1993
S-PWM4			M			Suttle & Chan 1993
S-BBS1	BBC1		S	50	230 x ?	Suttle & Chan 1993
S-BBP1	BBC2		P	50		Suttle & Chan 1993
S-PWP1	SNC1		P	65		Suttle & Chan 1993
Syn1	WH8101	WHOI	M			Waterbury & Valois 1993
Syn2	WH8012	Sargasso Sea (34°06'N, 61°01'W)	M	66	149x19	Waterbury & Valois 1993
Syn5	WH8109	Sargasso Sea (34°06'N, 61°01'W)	P			Waterbury & Valois 1993
Syn7	WH8012	WHOI	M			Waterbury & Valois 1993
Syn9	WH8012	WHOI	M	87	153x19	Waterbury & Valois 1993
Syn10	WH8017	Gulf Stream (36°58'N, 73°42'W)	M	100	145x19	Waterbury & Valois 1993
Syn12	WH8017	Gulf Stream (36°58'N, 73°42'W)	P	45	8x10	Waterbury & Valois 1993
Syn14	WH8103	Gulf Stream (36°58'N, 73°42'W)	M	93	136x21	Waterbury & Valois 1993
Syn16	WH8018	Sargasso Sea (34°06'N, 61°01'W)	M			Waterbury & Valois 1993
Syn17	WH8018	Gulf Stream (36°58'N, 73°42'W)	M			Waterbury & Valois 1993
Syn18	WH8108	Sargasso Sea (34°06'N, 61°01'W)	M			Waterbury & Valois 1993
Syn19	WH8109	Sargasso Sea (34°06'N, 61°01'W)	M			Waterbury & Valois 1993
Syn20	WH8109	Gulf Stream (36°58'N, 73°42'W)	M			Waterbury & Valois 1993
Syn22	WH8109	Gulf Stream (36°58'N, 73°42'W)	M			Waterbury & Valois 1993
P3	WH7805	Sapelo Island, GA	M			Lu, Chen & Hodson 2001
P5	WH7803	Sapelo Island, GA	M			Lu, Chen & Hodson 2001
P6	WH7805	Dauphin Island, GA	M			Lu, Chen & Hodson 2001
P8	WH8101	Sapelo Island, GA	M			Lu, Chen & Hodson 2001
P12	WH8101	Sayll Estuary, Ala	M			Lu, Chen & Hodson 2001
P16	WH7803	Savannah River Estuary, GA	M			Lu, Chen & Hodson 2001
P17	WH7803	Qingdao Coast, China	M			Lu, Chen & Hodson 2001
P39	WH7805	Savannah River Estuary, GA	M			Lu, Chen & Hodson 2001
P61	WH7803	Satilla River Estuary, GA	M			Lu, Chen & Hodson 2001
P66	WH7803	Satilla River Estuary, GA	M			Lu, Chen & Hodson 2001
P73	WH7805	Satilla River Estuary, GA	M			Lu, Chen & Hodson 2001
P76	WH8007	Altahama River Estuary, GA	M			Lu, Chen & Hodson 2001
P77	WH8007	Altahama River Estuary, GA	M			Lu, Chen & Hodson 2001
P79	WH7805	Satilla River Estuary, GA	M			Lu, Chen & Hodson 2001
P81	WH7805	Altahama River Estuary, GA	M			Lu, Chen & Hodson 2001
P1	WH7803	Sayll Estuary, Ala.	S			Lu, Chen & Hodson 2001

Cyanophage host range and field abundance

As the number of fully-sequenced phage and microbial genomes increases, their analyses are revealing the importance of horizontal gene transfer (HGT) as an evolutionary force among prokaryotes and their phages (Hendrix et al., 1999; Ochman, Lawrence, and Groisman, 2000). Although we know that some phages can move genes between hosts (Jiang and Paul, 1998; Miller, 2001; Miller et al., 1992; Paul, 1999), and that phages can help shape microbial population structure (Fuhrman, 1999; Mann, 2003; Suttle and Chan, 1994; Waterbury and Valois, 1993), we are still working to understand the mechanisms and extent to which phages influence the evolution of their hosts. To begin to understand the basic natural history of the cyanophages in our collection, we examined the host range of 45 cyanophage isolates using 21 *Prochlorococcus* and *Synechococcus* cultured strains (Chapter 2). **Do phages isolated using *Prochlorococcus* cross-infect other strains of *Prochlorococcus*? *Synechococcus*?**

Among the broad diversity of phages, there are two well-described classes of phage infective lifestyles: lytic (discussed first) and temperate (discussed in the following section). Lytic phages infect their host, use the host cellular processes to build new phage particles, and then burst the host cell releasing progeny phage. Strain-specific lytic *Synechococcus* cyanophage titers measured in mostly coastal marine systems showed these cyanophage are abundant (up to 10^4-10^6 ml^{-1}) with the highest *Synechococcus* cyanophage titers remaining within an order of magnitude of the total *Synechococcus* concentrations (Lu, Chen, and Hodson, 2001; Suttle and Chan, 1994; Waterbury and Valois, 1993). Calculations from such field abundances and other observations (reviewed in (Mann, 2003)) suggest that, in marine systems, lytic phages can be responsible for up to 50% of bacterial mortality per day (Fuhrman, 1999) and a small (<3%) portion of cyanobacterial mortality per day (Mann, 2003; Suttle, 2000; Suttle and Chan, 1994; Waterbury and Valois, 1993). However, the open ocean systems dominated by *Prochlorococcus* (DuRand, Olson, and Chisholm, 2001; Partensky, Hess, and Vaulot, 1999) offer a contrasting environment to the previous studies of *Synechococcus* cyanophage.

These oligotrophic regions likely differ from coastal environments in three ways that could affect assays measuring the phage abundances: the types of dominant cyanobacterial cells (Ferris and Palenik, 1998), the total diversity of cell populations (Fuhrman, 2000), and nutrient limitations (Cavender-Bares, Karl, and Chisholm, 2001; Wu et al., 2000). First, culture-based assays might better represent naturally occurring cells in one environment relative to another, creating the appearance of cyanophage titers that change along such a transect. Second, if host cell diversity changes along such a gradient, then phage titers may change accordingly as they are suggested to decrease where host cell diversity increases (Thingstad, 2000). Finally, decreased

nutrient availability along the transect (Cavender-Bares, Karl, and Chisholm, 2001) might result in sub-optimal growth of host cells in the Sargasso Sea (Mann et al., 2002) relative to *Synechococcus* at the coastal site along this transect (Waterbury et al., 1986). Viral production is correlated with host growth rates in chemostats (Bohannan and Lenski, 2000) and in the field (Steward, Smith, and Azam, 1996), which could result from nutrient limitation causing physiological changes in the host that stall the lytic process of obligately lytic phage (Stent, 1963), or favour lysogeny in temperate phage (Rohwer et al., 2000; Suttle, 2000; Wilson, Carr, and Mann, 1996).

Previous work suggests that phosphate, in particular, is the most likely nutrient to influence phage abundances along such a transect. Mesocosm experiments with *Emiliani huxleyi* viruses showed that viral production did not occur under phosphate-deplete conditions (Bratbak, 1993). *Synechococcus* WH7803 infection by cyanophage S-PM2 under various nutrient stress conditions showed a significant (80%) reduction in burst size (i.e., phage produced per cell) occurs in phosphate-deplete (but not nitrate-deplete) conditions relatve to phosphate-replete conditions (Wilson, Carr, and Mann, 1996). Subsequent work in seawater mesocosms with natural communities of *Synechococcus* showed that phosphate additions to phosphate-deplete enclosures terminated a *Synechococcus* bloom and was interpreted as a phosphate stimulation of viral production through the induction of prophage within the cyanobacterial genomes (Wilson and Mann, 1997). Finally, the *phoH* gene, which in *E. coli* is a phosphate-induced gene that encodes a putative ATPase, is found in the genomes of two phages that infect marine bacterial hosts (Miller et al., 2003; Rohwer et al., 2000). Together, these observations suggest that phosphate is important to production of phages in marine systems. We examined strain-specific cyanophage titers along a coastal-to-open Atlantic Ocean transect (Chapter 2). **How do such *Prochlorococcus* and *Synechococcus* phage titers vary along such a transect that includes gradients in the types of dominant cells, total cell diversity and bioavailable nutrients?**

Phage portal protein gene diversity in myoviruses

An understanding of the roles of marine phage (indeed all phage) in their communities has been hampered by the lack of a universal gene (analogous to the 16S rRNA gene that is used as the taxonomic marker for all microbes) that could be used to investigate phage genetic diversity (Paul et al., 2002). For this reason, studying the genetic diversity of both cultured and wild phages has proven difficult. Recently, the use of family-specific genes has been proposed for use as taxonomic tools (Rohwer and Edwards, 2002). For the *Myoviridae,* which are highly represented among *Synechococcus* cyanophage isolates (Fuller et al., 1998; Lu, Chen, and

Hodson, 2001; Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993), the family-specific gene commonly used is a gene homologous to the coliphage T4 portal protein gene, g20 (Fuller et al., 1998; Zhong et al., 2002). Many field studies have now examined myophage (phage of the morphological family *Myoviridae*) g20 sequence diversity in a variety of aquatic environments using DGGE banding patterns (Dorigo, Jacquet, and Humbert, 2004; Frederickson, Short, and Suttle, 2003; Wilson et al., 1999; Wilson et al., 2000), cloning and sequencing (Dorigo, Jacquet, and Humbert, 2004; Zhong et al., 2002) and/or culturing and sequencing (Marston and Sallee, 2003; Zhong et al., 2002). We examined the diversity of g20 sequences from the *Myoviridae* in this cyanophage collection (Chapter 3). **Do the g20 sequences of new cyanophage isolates represent novel g20 sequence types? Are they genetically related to g20 sequences from known cyanophage isolates or from unknown environmental g20 sequences? Are existing PCR primer sets adequately capturing the full g20 sequence diversity? Does g20 diversity correspond to ecologically relevant phage traits?**

Searching for prophage in host cyanobacterial genomes

In contrast to lytic phages, temperate phages do not always immediately lyse their hosts, but instead can temporarily insert their DNA into their host genome as a prophage. The existence of inducible prophage (prophage that can be 'induced' out of the host genome to enter the lytic phase of the temperate phage life cycle) was first elegantly demonstrated at the single cell level over half a century ago (Lwoff, 1953). Prophage induction has since been repeatedly demonstrated in natural microbial isolates and recently identified in most microbial genomes (reviewed in (Casjens, 2003)). Prophage are so common that they often account for a significant fraction of the "strain-specific" DNA between closely related microbial strains (Baba et al., 2002; Simpson et al., 2000; Smoot et al., 2002) and, furthermore, their genes are highly expressed as seen using genome-wide expression arrays under varying conditions (Smoot et al., 2001; Whiteley et al., 2001).

Together, these findings emphasize that prophage are not only widespread in prokaryotes, but also frequently account for strain diversification at both the genome and transcriptome (expression) levels often altering the host cell's physiology. However, the genomes of currently available freshwater cyanobacterial genomes lack intact prophage (Canchaya et al., 2003; Casjens, 2003) and no direct observation has confirmed the existence of prophage in marine cyanobacteria. We examined the genome sequences of 3 marine cyanobacteria (*Prochlorococcus* strains MED4 and MIT9313 and *Synechococcus* WH8102) for the presence of intact prophage (Appendix B). **Are there intact prophage genomes within marine cyanobacterial genomes?**

If yes, what type of phage do their genomes most resemble? If not, could potential prophage have been induced during cultivation? Is there any evidence suggesting prophage ever integrated into these genomes?

Cyanophage genomics

Phage genomes represent the largest unexplored reservoir of sequence information in the biosphere (Pedulla et al., 2003). Calculations using two independent methods suggest that less than 0.0002% of this reservoir has been sampled (Rohwer, 2003). First, using the estimated number of phage types (100 million), an average genome size of 50 ORFs and extrapolation from the number of phage-encoded ORFs from uncultured genomic DNA sequencing that lack known function suggests that 2.5 billion phage-encoded ORFs remain to be discovered. Second, using the non-parametric estimator Chao1 to evaluate every phage-encoded ORF compared against every other, one predicts that 2 billion phage-encoded ORFs remain to be discovered. In fact, nearly every new phage genome sequenced leads to novel insight into the role phages play in their host physiology and niche differentiation (Brussow and Hendrix, 2002).

Since the first phage was sequenced over a quarter of a century ago (Sanger et al., 1977), observations from over 150 phage genomes (primarily pathogen related) have taught us a great deal. First, our models of phage evolution have changed to reflect the fact that phages are modular and evolve through the homologous and non-homologous exchange of genes from a common gene pool (Hendrix, 2002; Hendrix et al., 1999). Second, the genetic sloppiness of phages, which is compensated by their high numbers of progeny, allows the acquisition of "non-phage" genes that most often will decrease phage fitness and be lost from the phage population, but occasionally provides advantage to the phage and/or host and thus will become fixed in the phage population (Hendrix et al., 2000). These events, though rare, can create entirely new phage-host evolutionary dynamics even defining new host capabilities, and thus are vital to the physiology, ecology and evolution of both the phage and its host.

Marine phage genomics is in its fledgling stage, with, at the time of this writing, only a handful of published genomes (Paul et al., 2002). Currently, there are over 170 phage genomes in GenBank, but only seven infect marine hosts (cyanophage P60; vibriophages VpV262, KVP40, VP16T, VP16C; roseophage SIO1; *Pseudoalteromonas* phage PM2), and only one is a cyanophage (P60). However, genomic analyses have already provided tantalizing hints that, just as phages are connected to their host's ecology in phagen phage-host systems, so too marine phages appear intimately tied to the ecological setting and physiological state of their host. For example, phosphate is a commonly limiting nutrient in marine systems (Cavender-Bares, Karl,

and Chisholm, 2001; Wu et al., 2000) and the discovery of phosphate-inducible genes in *Roseophage* SIO1 (Rohwer et al., 2000) and *Vibriophage* KVP40 (Miller et al., 2003) suggests that these phages respond to the phosphate status of their hosts and/or environment. Data in this thesis from analyses of three cyanophage genomes also supports a strong connection between the phage and their host's ecology (Chapter 4, 5). Through collaboration with Forest Rohwer (San Diego State University), David Mead (Lucigen) and the Department of Energy Joint Genome Institute we sequenced the genomes of three *Prochlorococcus* cyanophage isolates representing two viral families (*Podoviridae* and *Myoviridae*). **Will the gene complements of these cyanophages isolated from nutrient-limited waters provide insight into properties of these phages that are unique to the oligotrophic ocean environment of their hosts?** Recently, core photosynthetic genes were discovered in a *Synechococcus* cyanophage and it was hypothesized that these genes were expressed enabling photosynthesis to continue during infection (Mann et al., 2003). **Do the gene complements of these *Prochlorococcus* cyanophage also photosynthetic genes? Can we use the genetic information in these genomes to infer whether phage influence the genomes of their hosts? Vice versa?**

Comparative phage genomics suggests that dsDNA phages evolve through the exchange of genetic material, in the form of modular functional cassettes, through a global phage genome pool (Hendrix et al., 1999). To explain the observation that phage genomes appear to be mosaics, containing a large number of fixed, essential genes interspersed with highly variable, non-essential genes (Desplats and Krisch, 2003; Hendrix et al., 1999; Molineux, in press), it has been suggested that access to this global phage genome pool must be limited in some manner (Hendrix et al., 1999). This theory that phage evolve primarily through the horizontal exchange of genes appears to be applicable to the few phage types (which are predominantly temperate phages, i.e., capable of integrating into their host genomes) whose genomes are well represented among the phage genome databases. These include such phages as the lambdoid phages (capable of genetically recombining with coliphage lambda) (Hendrix et al., 1999), dairy phages (phages that infect lactic acid bacteria) (Brussow, 2001) and the mycobacteriophages (phages that infect mycobacteria) (Pedulla et al., 2003). However, fundamental differences in phage infection strategies (i.e., temperate phages integrate into host genomes wheras lytic phages do not) may lead to quantitatively different opportunities for horizontal gene exchange and it remains an open question whether lytic phage genomes might be less prone to mosaicism (Kovalyova and Kropinski, 2003). Further, if all phages are extensively horizontally transferring genes, this leads one to wonder whether significantly new phage types will be discovered as more phage genomes are sequenced or whether there might be a limited number of phage types are possible. Because

cyanobacterial phages are underrepresented in the phage genome database (only 1 cyanophage out of over 170 phage genomes), the *Prochlorococcus* cyanophage genomes sequenced during this thesis offer comparison of a phage that infects phylogenetically disparate hosts from those contained in the database. **How do these new *Prochlorococcus* phage genomes compare to those already sequenced? Are their genomes organized similarly to well-characterized phages? What kinds of inferences can we draw about phage-host interactions from the presence or absence of 'non-phage-like' genes in these phage genomes?**

References

Ahlgren, N., Hook, H. E., and Rocap, G. (2004). *Ocean Research Conference, Honolulu, Hawaii.*

Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., Nagai, Y., Iwama, N., Asano, K., Naimi, T., Kuroda, H., Cui, L., Yamamoto, K., and Hiramatsu, K. (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359**(9320), 1819-27.

Bergh, O. (1989). High abundance of viruses found in aquatic environments. *Nature* **340,** 467-468.

Bohannan, B. J. M., and Lenski, R. E. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecol. Letters* **3,** 362-377.

Borsheim, K. Y., Bratbak, G., and Heldal, M. (1990). Enumeration and biomass estimation of planktonic bacteria and viruses by transmission electron microscopy. *Appl Environ Microbiol* **56**(2), 352-6.

Bradley, D. E. (1967). Ultrastructure of bacteriophages and bacteriocins. *Bacteriol Rev* **31,** 230-314.

Bratbak, G. (1990). Viruses as partners in spring bloom microbial trophodynamics. *Applied and Environmental Microbiology* **56,** 1400-1405.

Bratbak, G. (1993). Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. *Marine Ecological Progress Series* **93,** 39-48.

Bratbak, G., Heldal, M., Thingstad, T. F., Riemann, B., and Haslund, O. H. (1992). Incorporation of viruses into the budget of microbial C-transfer: A first approach. *Mar. Ecol. Prog. Ser.* **83,** 273-280.

Brussow, H. (2001). Phages of Dairy Bacteria. *Annu Rev Microbiol* **55,** 283-303.

Brussow, H., and Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell* **108**(1), 13-6.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev* **67**(2), 238-76.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* **49**(2), 277-300.

Cavender-Bares, K. K., Karl, D. M., and Chisholm, S. W. (2001). Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure, and comparison to patterns in the Pacific Ocean. *Deep-Sea Research, Part I* **48,** 2373-2395.

Cottrell, M. T., and Suttle, C. A. (1991). Widespread occurrence and clonal variation in viruses which caused lysis of a cosmopolitan, eukaryotic marine phytoplankter, *Micromonas pusilla. Mar. Ecol. Prog. Ser.* **78,** 1-9.

Desplats, C., and Krisch, H. M. (2003). The diversity and evolution of the T4-type bacteriophages. *Res Microbiol* **154,** 259-267.

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004). Cyanophage diversity, inferred from g20 gene analyses, in the Largest Natural Lake in France, Lake Bourget. *Appl Environ Microbiol* **70,** 1017-22.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I., and Hess, W. R. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**(17), 10020-5.

DuRand, M. D., Olson, R. J., and Chisholm, S. W. (2001). Phytoplankton population dynamics at the Bermuda Atlantic Time-series Station in the Sargasso Sea. *Deep Sea Research II* **48,** 1983-2003.

Ferris, M. J., and Palenik, B. (1998). Niche adaptation in ocean cyanobacteria. *Nature* **396,** 226-228.

Frederickson, C. M., Short, S. M., and Suttle, C. A. (2003). The physical environment affects cyanophage communities in British Columbia Inlets. *Microb Ecol* **46,** 348-357.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399,** 541-548.

Fuhrman, J. A. (2000). Impact of viruses on bacterial processes. *In* "Microbial Ecology of the Oceans" (D. L. Kirchman, Ed.), pp. 327-350. Wiley-Liss, Inc.

Fuhrman, J. A., and Campbell, L. (1998). Microbial microdiversity. *Nature* **393,** 410-411.

Fuller, N. J., Marie, D., Partensky, F., Vaulot, D., Post, A. F., and Scanlan, D. J. (2003). Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine Synechococcus clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69**(5), 2430-43.

Fuller, N. J., Wilson, W. H., Joint, I. R., and Mann, N. H. (1998). Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl Environ Microbiol* **64**(6), 2051-60.

Goericke, R. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Research* **40,** 2283-2294.

Heldal, M., and Bratbak, G. (1991). Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser.* **72,** 205-212.

Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor Popul Biol* **61**(4), 471-80.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends in Microbiology* **8,** 504-508.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* **96**(5), 2192-7.

Jiang, S. C., and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol* **64**(8), 2780-7.

Kovalyova, I. V., and Kropinski, A. M. (2003). The complete genomic sequence of lytic bacteriophage gh-1 infecting Pseudomonas putida--evidence for close relationship to the T7 group. *Virology* **311**(2), 305-15.

Li, W. K. W. (1994). Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography* **39**(1), 169-175.

Liu, H., Campbell, L., and Landry, M. R. (1995). Growth and mortality rates of *Prochlorococcus* and *Synechococcus* measured with a selective inhibitor technique. *Marine Ecology Progress Series* **116,** 277-287.

Liu, H., Campbell, L., Landry, M. R., Nolla, H. A., Brown, S. L., and Constantinou, J. (1998). *Prochlorococcus* and *Synechococcus* growth rates and contributions to production in the Arabian Sea during the 1995 Southwest and Northeast Monsoons. *Deep-Sea Research II* **45,** 2327-2352.

Liu, H., and Landry, M. R. (1999). *Prochlorococcus* growth rates in the central equatorial Pacific: An application of the *fmax* approach. *J. Geophys. Res.* **104,** 3391-3399.

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. *Aquatic Microbial Ecology* **12,** 39-47.

Lu, J., Chen, F., and Hodson, R. E. (2001). Distribution, isolation, host specificity, and diversity of cyanophages infecting marine Synechococcus spp. in river estuaries. *Appl Environ Microbiol* **67**(7), 3285-90.

Lwoff, A. (1953). Lysogeny. *Bacteriological Review* **17,** 269-337.

Mann, E. L., Ahlgren, N., Moffett, J. W., and Chisholm, S. W. (2002). Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47,** 976-988.

Mann, N. H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev* **27**(1), 17-34.

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature* **424,** 741.

Marston, M. F., and Sallee, J. L. (2003). Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters. *Appl Environ Microbiol* **69**(8), 4639-47.

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., Lee, J., Szczypinski, B., and Fraser, C. M. (2003). Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185**(17), 5220-33.

Miller, R. V. (2001). Environmental bacteriophage-host interactions: Factors contribution to natural transduction. *Antonie Van Leeuwenhoek* **79,** 141-147.

Miller, R. V., Ripp, S., Replicon, J., Ogunseitan, O. A., and Kokjohn, T. A. (1992). Virus-mediated gene transfer in freshwater environments. *In* "Gene transfers and environment" (M. J. Gauthier, Ed.). Springer-Verlag, Berlin, Germany.

Moebus, K. (1987). Ecology of marine bacteriophages. *In* "Phage Ecology" (S. M. Goyal, C. P. Gerba, and G. Bitton, Eds.), pp. 136-156. john Wiley & Sons, Inc., New York.

Molineux, I. (in press). The T7 group. *In* "The Bacteriophages" (R. Calendar, Ed.).

Moore, L. R., and Chisholm, S. W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnology and Oceanography* **44**(3), 628-638.

Moore, L. R., Goericke, R., and Chisholm, S. W. (1995). Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine Ecology Progress Series* **116,** 259-275.

Moore, L. R., Post, A. F., Rocap, G., and Chisholm, S. W. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* **47,** 989-996.

Moore, L. R., Rocap, G., and Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**(6684), 464-467.

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A., and Summers, M. D. (1995). "Virus Taxonomy: Classification and Nomenclature of Viruses." Springer, Vienna, Austria.

Nagasaki, K., Tarutani, K., and Yamaguchi, M. (1999). Growth characteristics of Heterosigma akashiwo virus and its possible use as a microbiological agent for red tide control. *Appl Environ Microbiol* **65**(3), 898-902.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**(6784), 299-304.

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

Paul, J. H. (1999). Microbial gene transfer: An ecological perspective. *J. Mol. Microbiol. Biotechnol.* **1,** 45-50.

Paul, J. H., Sullivan, M. B., Segall, A. M., and Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**(4), 463-76.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Hendrix, R. W., and Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**(2), 171-82.

Proctor, L. M., and Fuhrman, J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature* **343,** 60-62.

Rocap, G., Distel, D., Waterbury, J. B., and Chisholm, S. W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Applied and Environmental Microbiology* **68**(3), 1180-1191.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., 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., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042-1047.

Rocap, G., McKay, J. F., and Ahlgren, N. (2004). *Ocean Research Conference, Honolulu, Hawaii.*

Rohwer, F. (2003). Global phage diversity. *Cell* **113**(2), 141.

Rohwer, F., and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**(16), 4529-35.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45,** 408-418.

Safferman, R. S., Cannon, R. E., Desjardins, P. R., Gromov, B. V., Haselkorn, R., Sherman, L. A., and Shilo, M. (1983). Classification and nomenclature of viruses of cyanobacteria. *Intervirology* **19,** 61-66.

Sanger, F., Air, G. M., Barrell, B. G., and al., e. (1977). Nucleotide sequence of bacteriophage phiX174 DNA. *Nature* **265,** 687-695.

Sieburth, J. M., Johnson, P. W., and Hargraves, P. E. (1988). Ultrastructure and ecology of *Aureococcus anophagefferens* gen. et sp. nov. (Chrysophycaea): the dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, summer 1985. *J. Phycol.* **24,** 416-425.

Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., Alves, L. M., Araya, J. E., Baia, G. S., Baptista, C. S., Barros, M. H., Bonaccorsi, E. D., Bordin, S., Bove, J. M.,

Briones, M. R., Bueno, M. R., Camargo, A. A., Camargo, L. E., Carraro, D. M., Carrer, H., Colauto, N. B., Colombo, C., Costa, F. F., Costa, M. C., Costa-Neto, C. M., Coutinho, L. L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A. P., Ferreira, A. J., Ferreira, V. C., Ferro, J. A., Fraga, J. S., Franca, S. C., Franco, M. C., Frohme, M., Furlan, L. R., Garnier, M., Goldman, G. H., Goldman, M. H., Gomes, S. L., Gruber, A., Ho, P. L., Hoheisel, J. D., Junqueira, M. L., Kemper, E. L., Kitajima, J. P., Krieger, J. E., Kuramae, E. E., Laigret, F., Lambais, M. R., Leite, L. C., Lemos, E. G., Lemos, M. V., Lopes, S. A., Lopes, C. R., Machado, J. A., Machado, M. A., Madeira, A. M., Madeira, H. M., Marino, C. L., Marques, M. V., Martins, E. A., Martins, E. M., Matsukuma, A. Y., Menck, C. F., Miracca, E. C., Miyaki, C. Y., Monteriro-Vitorello, C. B., Moon, D. H., Nagai, M. A., Nascimento, A. L., Netto, L. E., Nhani, A., Jr., Nobrega, F. G., Nunes, L. R., Oliveira, M. A., de Oliveira, M. C., de Oliveira, R. C., Palmieri, D. A., Paris, A., Peixoto, B. R., Pereira, G. A., Pereira, H. A., Jr., Pesquero, J. B., Quaggio, R. B., Roberto, P. G., Rodrigues, V., de, M. R. A. J., de Rosa, V. E., Jr., de Sa, R. G., Santelli, R. V., Sawasaki, H. E., da Silva, A. C., da Silva, A. M., da Silva, F. R., da Silva, W. A., Jr., da Silveira, J. F., Silvestri, M. L., Siqueira, W. J., de Souza, A. A., de Souza, A. P., Terenzi, M. F., Truffi, D., Tsai, S. M., Tsuhako, M. H., Vallada, H., Van Sluys, M. A., Verjovski-Almeida, S., Vettore, A. L., Zago, M. A., Zatz, M., Meidanis, J., and Setubal, J. C. (2000). The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* **406**(6792), 151-7.

Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F., Parkins, L. D., Beres, S. B., Campbell, D. S., Smith, T. M., Zhang, Q., Kapur, V., Daly, J. A., Veasy, L. G., and Musser, J. M. (2002). Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99**(7), 4668-73.

Smoot, L. M., Smoot, J. C., Graham, M. R., Somerville, G. A., Sturdevant, D. E., Migliaccio, C. A., Sylva, G. L., and Musser, J. M. (2001). Global differential gene expression in response to growth temperature alteration in group A Streptococcus. *Proc Natl Acad Sci U S A* **98**(18), 10416-21.

Spencer, R. (1955). A marine bacteriophage. *Nature* **175,** 690.

Spencer, R. (1960). Indigenous marine bacteriophages. *J. Bacteriol.* **79,** 614.

Spencer, R. (1963). Bacterial viruses in the sea. *In* "Symposium on Marine Microbiology" (C. H. Oppenheimer, Ed.), pp. 350-365. Charles C. Thomas, Springfield.

Stent, G. S. (1963). "Molecular biology of bacterial viruses." W.H. Freeman and Company, San Francisco.

Steward, G., Smith, D. C., and Azam, F. (1996). Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Mar. Ecol. Prog. Ser.* **131,** 287-300.

Suttle, C. A. (1999). Do viruses control the oceans? *Natural History* **Feb,** ???

Suttle, C. A. (2000). Cyanophages and their role in the ecology of cyanobacteria. *In* "The Ecology of Cyanobacteria" (B. A. Whitton, and M. Potts, Eds.), pp. 563-589. Kluwer Academic Publishers, Netherlands.

Suttle, C. A., and Chan, A. M. (1993). Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92,** 99-109.

Suttle, C. A., and Chan, A. M. (1994). Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Applied. Environ. Microbiol.* **60**(9), 3167-3174.

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnol. Oceanogr.* **45**(6), 1320-1328.

Torella, F., and Morita, R. Y. (1979). Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon: ecological and taxonomical implications. *Appl. Env. Microbiol.* **37,** 774-778.

Veldhuis, M. J. W., Kraay, G. W., van Bleijswijk, J. D. L., and Baars, M. A. (1997). Seasonal and spatial variability in phytoplankton biomass, productivity and growth in the northwestern Indian Ocean: the southwest and northeast monsoon, 1992-1993. *Deep Sea Research* **44,** 425-449.

Waterbury, J. B., and Valois, F. W. (1993). Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**(10), 3393-3399.

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.* **214,** 71-120.

Whiteley, M., Bangera, M. G., Bumgarner, R. E., Parsek, M. R., Teitzel, G. M., Lory, S., and Greenberg, E. P. (2001). Gene expression in Pseudomonas aeruginosa biofilms. *Nature* **413**(6858), 860-4.

Wichels, A., Biel, S. S., Gelderblom, H. R., Brinkhoff, T., Muyzer, G., and Schutt, C. (1998). Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* **64**(11), 4128-33.

Wilson, W. H., Carr, N. G., and Mann, N. H. (1996). The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp. WH7803. *J. Phycol.* **32,** 506-516.

Wilson, W. H., Fuller, N. J., Joint, I. R., and Mann, N. H. (1999). Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. *In* "Marine Cyanobacteria" (L. Sharpy, and A. W. D. Larkum, Eds.), Vol. 19, pp. 209-216. Bulletin de l'Institut oceanographique, Monaco.

Wilson, W. H., Fuller, N. J., Joint, I. R., and Mann, N. H. (2000). Analysis of cyanophage diversity in the marine environment using denaturing gradient gel electrophoresis. *In* "Microbial Biosystems: New Fronteirs. Proceedings of the 8th International Symposium on Microbial Ecology" (C. R. Bell, M. Brylinksky, and P. Johnson-Green, Eds.). Atlantic Canada Society for Microbial Ecology, Halifax, Nova Scotia, Canada.

Wilson, W. H., Joint, I. R., Carr, N. G., and Mann, N. H. (1993). Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl Environ Microbiol* **59,** 3736-3743.

Wilson, W. H., and Mann, N. H. (1997). Lysogenic and lytic viral production in marine microbial communities. *Aq. Microb. Ecol.* **13,** 95-100.

Wommack, K. E., Hill, R. T., Kessel, M., Russek-Cohen, E., and Colwell, R. R. (1992). Distribution of viruses in the Chesapeake Bay. *Appl Environ Microbiol* **58**(9), 2965-70.

Wu, J., Sunda, W., Boyle, E. A., and Karl, D. M. (2000). Phosphate depletion in the Western North Atlantic Ocean. *Science* **289,** 759-762.

Zeidner, G., Preston, C. M., Delong, E. F., Massana, R., Post, A. F., Scanlan, D. J., and Beja, O. (2003). Molecular diversity among marine picophytoplankton as revealed by psbA analyses. *Environ Microbiol* **5**(3), 212-6.

Zhong, Y., Chen, F., Wilhelm, S. W., Poorvin, L., and Hodson, R. E. (2002). Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl Environ Microbiol* **68**(4), 1576-84.

ZoBell, C. (1946). "Marine Microbiology." Chronica Botanica, Waltham, MA.

Chapter II

Phages infecting the oceanic cyanobacterium *Prochlorococcus*

Reprinted with permission from *Nature*

Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W. (2003). Cyanophage infecting the oceanic cyanobacterium, *Prochlorococcus*. *Nature* **424,** 1047-1051.

Acknowledgements This research was funded by the Biological and Environmental Research Program of the US Department of Energy's Office of Science. The Joint Genome Institute managed the overall sequencing effort. Genome finishing was carried out under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory. Computational annotation was carried out at the Oak Ridge National Laboratory, managed by UT-BATTELLE for the US Department of Energy. Additional support was provided by the DOE, NSF and the Seaver Foundation to S.W.C., the Israel–US Binational Science Foundation to A.F.P. and S.W.C., and FP5-Margenes to W.R.H. and A.F.P. We thank the *Synechococcus* WH8102 annotators (B. Palenik, B. Brahamsha, J. McCarren, E. Allen, F. Partensky, A. Dufresne and I. Paulsen) for their help with curating the *Prochlorococcus* genomes and E. V. Armbrust and L. Moore for critical reading of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu). The complete nucleotide sequences and sequences of predicted open reading frames have been deposited in the EMBL/GenBank/DDBJ databases under accession numbers BX548174 (MED4) and BX548175 (MIT9313).

Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*

Matthew B. Sullivan[1], John B. Waterbury[3] & Sallie W. Chisholm[2]

[1]*MIT/Woods Hole Oceanographic Institution Joint Program in Biological Oceanography, and* [2]*Department of Civil and Environmental Engineering and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Woods Hole Oceanographic Institution, Department of Biology, Woods Hole, Massachusetts 02543, USA*

***Prochlorococcus* is the numerically dominant phototroph in the tropical and subtropical oceans, accounting for half of the photosynthetic biomass in some areas[1,2]. Here we report the isolation of cyanophages that infect *Prochlorococcus*, and show that although some are host-strain-specific, others cross-infect with closely related marine *Synechococcus* as well as between high-light- and low-light-adapted *Prochlorococcus* isolates, suggesting a mechanism for horizontal gene transfer. High-light-adapted *Prochlorococcus* hosts yielded *Podoviridae* exclusively, which were extremely host-specific, whereas low-light-adapted *Prochlorococcus* and all strains of *Synechococcus* yielded primarily *Myoviridae*, which has a broad host range. Finally, both *Prochlorococcus* and *Synechococcus* strain-specific cyanophage titres were low ($<10^3$ ml^{-1}) in stratified oligotrophic waters even where total cyanobacterial abundances were high ($>10^5$ cells ml^{-1}). These low titres in areas of high total host cell abundance seem to be a feature of open ocean ecosystems. We hypothesize that gradients in cyanobacterial population diversity, growth rates, and/or the incidence of lysogeny underlie these trends.**

Phages are thought to evolve by the exchange of genes drawn from a common gene pool through differential access imposed by host range limitations[3]. Similarly, horizontal gene transfer, important in microbial evolution[4,5], can be mediated by phages[6] and is probably responsible for many of the differences in the genomes of closely related microbes[5]. Recent detailed analyses of molecular phylogenies constructed for marine *Prochlorococcus* and *Synechococcus*[7,8] (Fig. 1) show that these genera form a single group within the marine picophytoplankton clade[9] (>96% identity in 16S ribosomal DNA sequences), yet display microdiversity in the form of ten well-defined subgroups[8]. We have used members of these two groups to study whether phage isolated on a particular host strain cross-infect other hosts, and if so, whether the probability of cross-infection is related to rDNA-based evolutionary distance between the hosts.

Analyses of host range were conducted (Fig. 1) with 44 cyanophages, isolated as previously described[10] from a variety of water depths and locations (see Supplementary Information) using 20 different host strains chosen to represent the genetic diversity of *Prochlorococcus* and *Synechococcus*[8]. Although we did not examine how these patterns would change if phage were propagated on different hosts, this would undoubtedly add another layer of complexity due to host range modifications as a result of methylation of phage DNA[6]. Similar to those that infect other marine bacteria[11] and *Synechococcus*[10–14], our *Prochlorococcus* cyanophage isolates fell into three morphological families: *Myoviridae*, *Siphoviridae* and *Podoviridae*[15].

As would be predicted[10–14], *Podoviridae* were extremely host specific with only two cross-infections out of a possible 300 (Fig. 1). Similarly, the two *Siphoviridae* isolated were specific to their hosts. In instances of extreme host specificity, *in situ* host abundance would need to be high enough to facilitate phage–host contact. It is noteworthy in this regard that members of the high-light-adapted *Prochlorococcus* cluster, which yielded the most host-specific cyanophage, have high relative abundances *in situ*[16]. The *Myoviridae* exhibited much broader host ranges, with 102 cross-infections out of a possible 539. They not only cross-infected among and between *Prochlorococcus* ecotypes but also between *Prochlorococcus* and *Synechococcus*. Those isolated with *Synechococcus* host strains have broader host ranges and are more likely to cross-infect low-light-adapted than high-light-adapted *Prochlorococcus* strains. The low-light-adapted *Prochlorococcus* are less diverged from *Synechococcus* than high-light-adapted *Prochlorococcus*[7,8], suggesting a relationship, in this instance, between the probability of cross-infection and rDNA relatedness of hosts. Finally, we tested the *Myoviridae* for cross-infection against marine bacterial isolates closely related to *Pseudoalteromonas*, which are known to be broadly susceptible to diverse bacteriophages (bacterial strains HER1320, HER1321, HER1327, HER1328)[11]. None of the *Myoviridae* cyanophages infected these bacteria.

Phage morphotypes isolated were determined, to some degree, by the host used for isolation (Fig. 1). For example, ten of ten cyanophages isolated using high-light-adapted *Prochlorococcus* strains were *Podoviridae*. In contrast, all but two cyanophages isolated on *Synechococcus* were *Myoviridae*, a bias that has been reported by others[14], and over half of those isolated on low-light-adapted *Prochlorococcus* belonged to this morphotype. We further substantiated these trends by examining lysates (as opposed to plaque-purified isolates) from a range of host strains, geographic locations and depths—of 58 *Synechococcus* lysates 93% contained *Myoviridae*, of 43 low-light-adapted *Prochlorococcus* lysates 65% contained *Myoviridae*, and of 107 high-light-adapted *Prochlorococcus* lysates 98% contained *Podoviridae* (see Supplementary Information).

Maximum cyanophage titres, using a variety of *Synechococcus* hosts, are usually found to be within an order of magnitude of the total *Synechococcus* abundance[10,14,17,18], and can be as high as 10^6 phage ml^{-1}. One study[17] has shown, for example, that along a transect in which total *Synechococcus* abundance decreased from 10^5 cells ml^{-1} to 250 cells ml^{-1}, maximum cyanophage titres remained at least as high as the total number of *Synechococcus*. We wondered whether titres of *Prochlorococcus* cyanophage in the Sargasso Sea, where *Prochlorococcus* cells are abundant (10^5 cells ml^{-1}), would be comparable to those measured in coastal oceans for *Synechococcus* where total *Synechococcus* host abundances are of similar magnitude. We assayed cyanophage titres in a depth profile in the Sargasso Sea at the end of seasonal stratification using 11 strains of *Prochlorococcus* (Fig. 2), choosing at least one host strain from each of the six phylogenetic clusters that span the rDNA-based genetic diversity of our culture collection[8].

Three *Prochlorococcus* host strains (MIT 9303, MIT 9313 and SS120) yielded low or no cyanophage. Other hosts yielded titres

that reached a maximum at 70 m (NATL2A-phage) or 100 m (MIT 9302-, MIT 9515-, MED4-, MIT 9211-, NATL1A-phage) near the depth of maximum *Prochlorococcus* abundance (Fig. 2). All *Prochlorococcus* cyanophage titres were low (<350 cyanophage ml^{-1}) compared with those reported for *Synechococcus* in coastal regions (approximately 10^4–10^6 cyanophage ml^{-1}) even though total host abundances were similar between these regions (approximately 10^5 cells ml^{-1})[10,14,17,18]. *Prochlorococcus* cyanophage titres are comparable to those of *Synechococcus* from oligotrophic waters in the Gulf of Mexico—but in that instance the total *Synechococcus* abundance was also low (<250 cells ml^{-1})[17].

Cyanophage titres were also examined along a surface water transect from coastal (mesotrophic) to open ocean (oligotrophic) in the Atlantic Ocean to better understand the relationship between maximum phage titre and total host abundance along a trophic gradient. Titres were assayed with 12 strains of *Synechococcus* and *Prochlorococcus* that represented the known rDNA-based genetic diversity at the time that we began the study[8] (but see also ref. 19). We found that *Synechococcus* cyanophage titres decreased by an order of magnitude or greater in surface waters between the coastal and open ocean (Sargasso) sites, whereas total *Synechococcus* abundance decreased from 3×10^4 to 7×10^3 cells ml^{-1} (Fig. 3). *Prochlorococcus* hosts did not yield cyanophage in coastal samples where there are no *Prochlorococcus* cells, and yielded relatively low titres (0 to 1.5×10^3 phage ml^{-1}) at the shelf, slope and Sargasso stations where total *Prochlorococcus* abundance was between 4.5×10^4 and 1.4×10^5 cells ml^{-1}. Even though total *Prochlorococcus* abundance at the Sargasso site was similar to that of *Synechococcus* at the coastal site (Fig. 3i, j), *Prochlorococcus* and *Synechococcus* cyanophage titres were significantly lower at the open ocean site (Fig. 3a–h). Moreover, regardless of the host used, titres never exceeded 3×10^3 cyanophage ml^{-1} at any depth throughout the photic zone even though total *Prochlorococcus* abundances exceeded 10^5 cells ml^{-1} (see Supplementary Information). Thus it seems that cyanophage titres at the end of summer stratification are relatively low in open ocean ecosystems, where the total possible host cell abundances are relatively high. Low titres lead to reduced contact rates and lowered mortality rates[6].

Although it is difficult to draw definitive conclusions about causality from such trends because of the complexity of the phage–host interaction, there are some factors that might be implicated. If, for example, host strain microdiversity increased along the transect and cross-infection ability did not increase concurrently, this would lead to lower phage titres yielded by a suite of host strains[20]. Indeed, we know that the relative abundance of *Synechococcus* ecotypes changes from coastal to oligotrophic waters[16]. However, we observed a systematic decrease in cyanophage titres for all five *Synechococcus* hosts (note WH 8020 yielded no

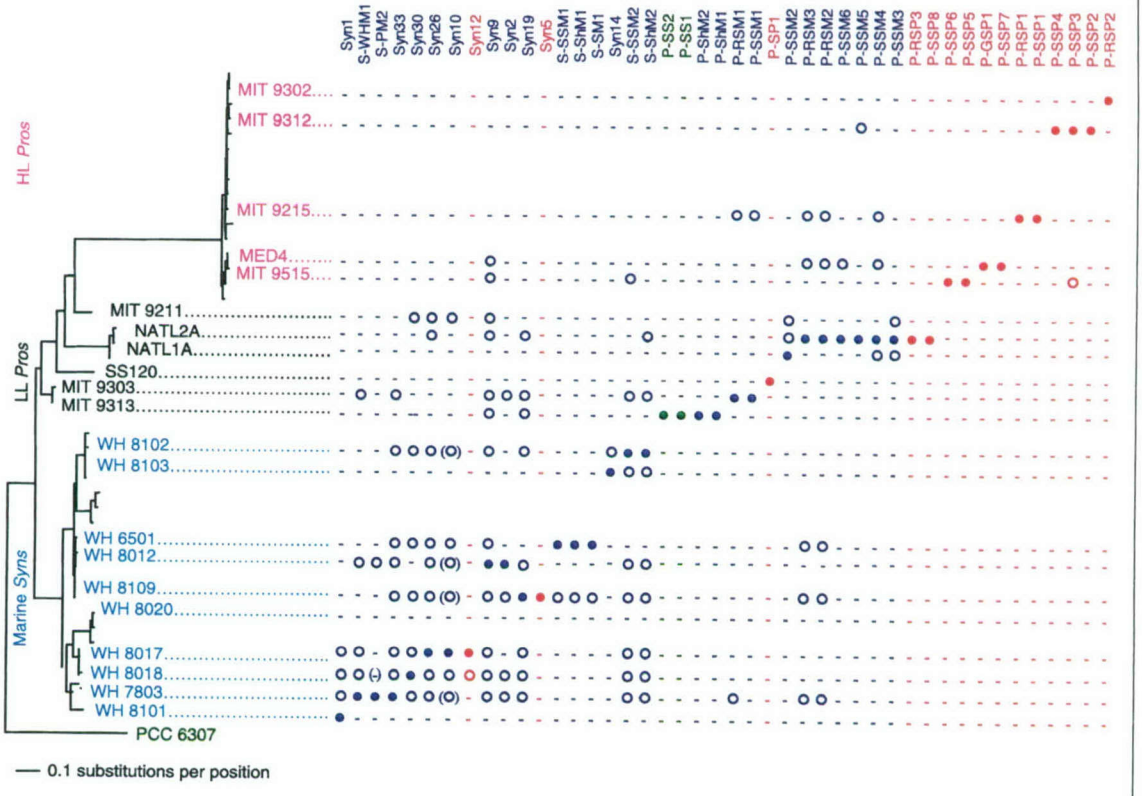


Figure 1 Host ranges of 44 clonal cyanophages exposed to marine *Prochlorococcus* and *Synechococcus* cultured isolates. The evolutionary relationships between the 21 host strains are shown in the phylogenetic tree inferred using 16S–23S rDNA spacer regions[8]. For the cyanophages, red indicates *Podoviridae*, blue indicates *Myoviridae* and green indicates *Siphoviridae*. Filled circles, host strain used to isolate a particular cyanophage; open circles, cross-infection of cyanophage with another host; dash, no infection (that is, lysis). Symbols in parentheses indicate that the results do not match earlier studies[10,13] with these phage and hosts (see Methods for details). HL *Pros*, high-light-adapted *Prochlorococcus*; LL *Pros*, low-light-adapted *Prochlorococcus*; Marine *Syns*, marine *Synechococcus*.

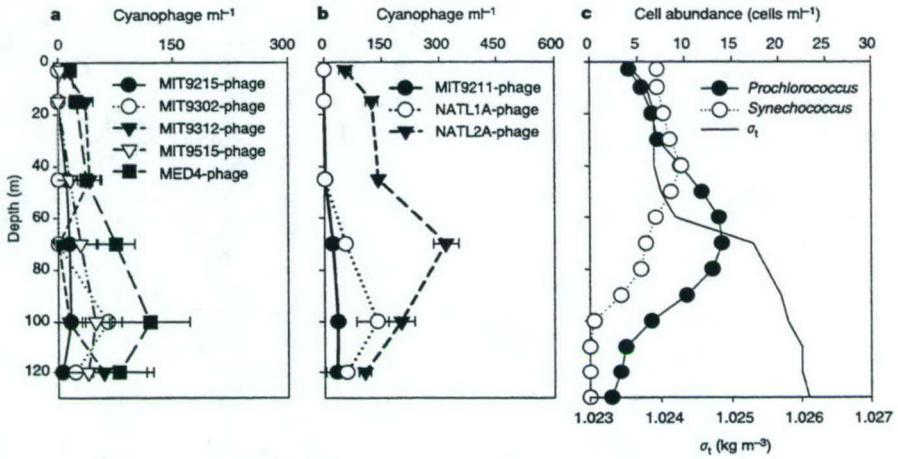


Figure 2 Cyanophage titres, measured using *Prochlorococcus* host strains, as a function of depth at the Bermuda Atlantic Time Series Station in the Sargasso Sea on 26 September 1999. Nine of the cyanophages used in host range analyses were isolated from this depth profile (see Supplementary Information). **a–c**, Titres measured using high-light-adapted *Prochlorococcus* hosts (**a**) and low-light-adapted hosts (**b**), and total *Prochlorococcus* and *Synechococcus* cell abundances (*Prochlorococcus* cells $\times 10^4$ ml^{-1}; *Synechococcus* cells $\times 10^3$ ml^{-1}) and σ_t (a proxy for water density used to measure the depth of the mixed layer) (**c**). Titres were undetectable for low-light-adapted *Prochlorococcus* strains SS120, MIT9303 and MIT9313. Error bars represent the s.d. of assays.

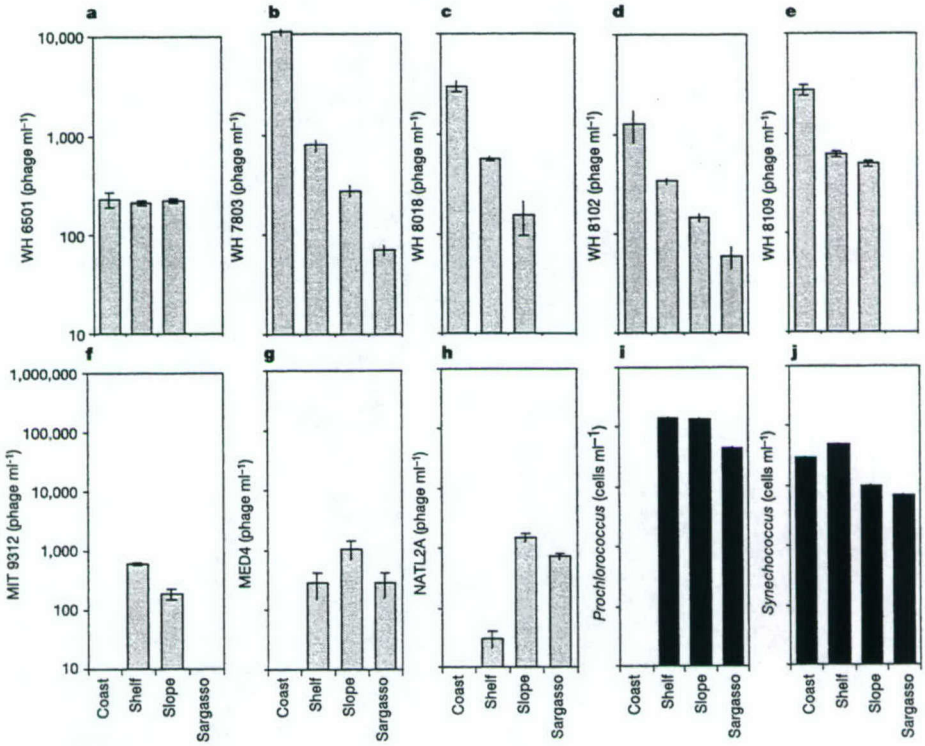


Figure 3 Cyanophage titres measured in *Synechococcus* and *Prochlorococcus* host cells along a surface water transect from coastal (coast, Woods Hole, Massachusetts) to open ocean (Sargasso) conducted in September 2001. Note that the magnitudes of the *y* axes are different for **a–e** and **f–j**. Ten of the cyanophages used in host range analyses were isolated from along this transect (see Supplementary Information). **a–h**, Cyanophage titres represent the averages and s.d. of triplicate plaque assays. **i**, **j**, Cell concentrations represent averages and s.d. of duplicate flow cytometry assays. Where no bar is shown, there were no plaques (**a**, **c**, **e–h**) or no cells (**i**). No plaques were observed at any of the surface samples along the transect for *Synechococcus* strain WH 8020 and *Prochlorococcus* strains MIT 9313, SS120 and MIT 9211 (data not shown).

plaques) at the extremes of this transect (the coastal and Sargasso sites; Fig. 3a–e). If changing host abundance alone explained the change in titres, and if our suite of host strains is representative of natural diversity, then one might expect at least one host strain to yield increasing titres along the gradient (for example, for the 'open ocean strain' WH 8102); however, this was not observed.

Another possible explanation for decreasing phage titres as one goes from coastal to open ocean ecosystems is decreased nutrient availability along the transect[21] resulting in suboptimal growth of host cells in the Sargasso Sea[22] relative to *Synechococcus* at the coastal site[23]. Viral production is correlated with host growth rates in chemostats[24] and in the field[25], which could result from nutrient limitation causing physiological changes in the host that stall the lytic process of obligately lytic phage[6], or favour lysogeny in temperate phage[18,26,27]. Although temperate phage have not been identified for marine *Prochlorococcus* or *Synechococcus*, INT family site-specific recombinases exist in the genomes of *Prochlorococcus* MED4 and MIT 9313, and *Synechococcus* WH 8102 (http://www.jgi.doe.gov/JGI_microbial/html/index.html), suggesting that prophages were once integrated into these host genomes[28,29].

The phage–host system described here should continue to be a useful framework for advancing our understanding of the ecology and evolution of phage–host interactions in marine ecosystems. We have known for some time that cyanophages must have a role in maintaining genetic diversity among hosts[10,17]. The broad host ranges reported here indicate further their potential for mediating horizontal gene transfer, which may help explain the extensive microdiversity[8,19,28,29] seen in these two groups of marine cyanobacteria. The extent to which this potential is realized should become clear as more and more host and phage genomes are sequenced. Of significance also is the coupling between phage morphology and host type. Experiments designed to characterize phage resistance across variable hosts and phages (for example, identification of receptors and restriction and modification systems) should elucidate the underlying mechanisms responsible for these patterns. Finally, our analyses of cyanophage titres along a coastal–open ocean transect suggest that the underlying processes responsible for the production of free cyanophage differ along trophic gradients in the oceans. To fully explain these observations will require the development of approaches that allow one to determine which phage can infect which host(s) in a given community, and an understanding of the relative roles of lytic and lysogenic phases of the viral life cycle in aquatic systems. □

Methods

Sample collection

Water samples for cyanophage titres, cyanophage isolations and cyanobacterial abundances were collected at the Bermuda Atlantic Time Series Station on 26 September 1999 and at four sites along a transect from Woods Hole to the Western Sargasso Sea on 5, 16, 17 and 22 September 2001 (see Supplementary Information). Water for cyanophage isolations was filtered (0.4 μm, Poretics number 13028 in 1999; 0.2 μm, Osmonics number K02CP04700 in 2001) and stored at 4 °C in the dark in acid-washed polycarbonate (1999) or glass (2001) bottles until analysis (up to 15 months later). Cyanophage titres remain stable for at least one year[27]. Control experiments showed that titres were stable over a 15-month period (see Supplementary Information).

Culturing conditions

Prochlorococcus and *Synechococcus* strains were maintained in '75% Pro99' medium, a modification of the 'Pro2' medium[30] with a 75% seawater base and the following final concentrations of N and P: 800 μM NH_4Cl, 50 μM NaH_2PO_4. Cultures were grown at 19–21 °C under constant light 8–12 $\mu E\,m^{-2}\,s^{-1}$ for low-light-adapted *Prochlorococcus* and *Synechococcus*; 35–45 $\mu E\,m^{-2}\,s^{-1}$ for high-light-adapted *Prochlorococcus*.

Cyanophage isolations

Prochlorococcus cyanophage isolations were done initially using an axenic strain of *Prochlorococcus* (MED4ax; M. Saito and J.B.W., unpublished observations). Exponentially growing cells were transferred to fresh medium (1 ml:20 ml) and inoculated with 1 ml of 0.4-μm-filtered sea water. The time course of auto-fluorescence (chlorophyll biomass) of these cultures was then followed with a Turner Designs 10-AU fluorometer. Cultures showing reduced fluorescence relative to controls were filtered and examined for phage particles as previously described[10,17]. Lysates were stored at 4 °C in the dark. Subsequent isolations using 19 additional host strains were done using the same procedures scaled down to small volumes. Cyanophage isolates used in this study were plaque-purified twice before use, classified using morphology described by the ICTV[15], and named according to suggestions made for cyanophage[18].

Cyanophage host range

Host range analyses were conducted over a period of about 2 yr. Each interaction between a cyanophage and its potential host cell was performed with exponentially growing cells in triplicate on at least two different occasions. Marine bacterial strains were purchased from the Felix d'Herelle Reference Center for Bacterial Viruses (contact H. Ackermann). Several of the *Synechococcus* cyanophage used in this study ('Syn' phages; S-PM2 and S-WHM1) had been previously examined for host range cross-infectivity[10,13] and were maintained as lysates at 4 °C in the dark while host cyanobacterial cultures were serially transferred in late exponential and early stationary phase. A total of 103 of 108 cross-infections using these stored cyanophages yielded similar host range results in this study (Fig. 1). Of the differences observed, four of five were for one cyanophage isolate (Syn10), suggesting that it might have evolved an extended host range mutation. Host range can be altered through DNA modifications that can occur during propagation of a phage on an alternative host. Overall, these results suggest that cyanophage susceptibility of these host strains and the cross-infectivity of the cyanophage remained relatively stable throughout the 10 or more years of storage and culture maintenance.

Host cell and cyanophage quantification

Host cell abundance was measured using a modified Becton-Dickinson FACScan flow cytometer[7]. Cyanophage titres were quantified using most probable number (MPN) assays (1999) or plaque assay (2001). MPN assays were monitored for lysis relative to controls for 2–3 weeks depending on the host strain used. For the plaque assays, we plated the host strain in soft agarose (0.4% final concentration; GIBCO BRL, Life Technologies number 5517-014) along with the phage being titred; lawns of cells appeared 8–28 days after inoculation, depending on the strain (M. Saito, M.B.S. and J.B.W., unpublished data). Plaques were counted daily until they no longer appeared (3–14 days after the first plaques). Titres measured using both assays were not significantly different (t-test assuming equal variances, $\alpha = 0.10$).

Host dependency of measured titres

To see whether our standard suite of host cells used in our assays was giving us a representative picture of the maximum phage titre obtainable in a given sample, we used every cultured isolate of *Prochlorococcus* in our collection with a unique ITS rDNA sequence (23 isolates) to assay the cyanophage titre of a 50-m water sample from the Red Sea. The range of titres yielded was representative of the range we measured using our subset of *Prochlorococcus* isolates (see Supplementary Information).

Received 21 April; accepted 15 July 2003; doi:10.1038/nature01929.

1. Liu, H., Nolla, H. A. & Campbell, L. *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. *Aquat. Microb. Ecol.* **12**, 39–47 (1997).
2. Partensky, F., Hess, W. R. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
3. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl Acad. Sci. USA* **96**, 2192–2197 (1999).
4. Woese, C. R. Interpreting the universal phylogenetic tree. *Proc. Natl Acad. Sci. USA* **97**, 8392–8396 (2000).
5. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
6. Stent, G. S. *Molecular Biology of Bacterial Viruses* (W.H. Freeman, San Francisco, 1963).
7. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464–467 (1998).
8. Rocap, G., Distel, D., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* **68**, 1180–1191 (2002).
9. Urbach, E. & Chisholm, S. W. Genetic diversity in *Prochlorococcus* populations flow cytometrically sorted from the Sargasso Sea and Gulf Stream. *Limnol. Oceanogr.* **43**, 1615–1630 (1998).
10. Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Appl. Environ. Microbiol.* **59**, 3393–3399 (1993).
11. Wichels, A. *et al.* Bacteriophage diversity in the North Sea. *Appl. Environ. Microbiol.* **64**, 4128–4133 (1998).
12. Suttle, C. A. & Chan, A. M. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92**, 99–109 (1993).
13. Wilson, W. H., Joint, I. R., Carr, N. G. & Mann, N. H. Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl. Environ. Microbiol.* **59**, 3736–3743 (1993).
14. Lu, J., Chen, F. & Hodson, R. E. Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl. Environ. Microbiol.* **67**, 3285–3290 (2001).
15. Murphy, F. A. *et al. Virus Taxonomy: Classification and Nomenclature of Viruses* (Springer, Vienna, Austria, 1995).
16. Scanlan, D. J. & West, N. J. Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microb. Ecol.* **40**, 1–12 (2002).
17. Suttle, C. A. & Chan, A. M. Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **60**, 3167–3174 (1994).

18. Suttle, C. A. in *The Ecology of Cyanobacteria* (eds Whitton, B. A. & Potts, M.) 563–589 (Kluwer, Netherlands, 2000).
19. Fuller, N. J. *et al.* Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl. Environ. Microbiol.* **69**, 2430–2443 (2003).
20. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
21. Cavender-Bares, K. K., Karl, D. M. & Chisholm, S. W. Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure, and comparison to patterns in the Pacific Ocean. *Deep-Sea Res. I* **48**, 2373–2395 (2001).
22. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47**, 976–988 (2002).
23. Waterbury, J. B., Watson, S. W., Valois, F. W. & Franks, D. G. Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.* **214**, 71–120 (1986).
24. Bohannan, B. J. M. & Lenski, R. E. Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecol. Lett.* **3**, 362–377 (2000).
25. Steward, G., Smith, D. C. & Azam, F. Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Mar. Ecol. Prog. Ser.* **131**, 287–300 (1996).
26. Rohwer, F. *et al.* The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45**, 408–418 (2000).
27. Wilson, W. H., Carr, N. G. & Mann, N. H. The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp. WH7803. *J. Phycol.* **32**, 506–516 (1996).
28. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).
29. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424**, 1037–1042 (2003).
30. Moore, L. R. & Chisholm, S. W. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* **44**, 628–638 (1999).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank the respective Captain and crew of the R/V *Oceanus* and the R/V *Endeavor* for their help on research cruises, which were made possible by an NSF grant to chief scientists J. Moffett and B. Binder, respectively. This work was funded in part by Grants from the Seaver Foundation, NSF and DOE to S.W.C., and an NIH Genome Training Grant, an MIT Research Fellowship and a WHOI Research Fellowship to M.B.S. We thank A. F. Post for Red Sea samples, G. Tan for technical assistance, J. A. King and P. Weigele for TEM suggestions, S. Casjens and F. Rohwer for help with prophage analysis, and W. H. Wilson for providing cyanophage isolates S-PM2 and S-WHM1.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu).

Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem

T. S. Bibby[1], I. Mary[2], J. Nield[1], F. Partensky[2] & J. Barber[1]

[1]*Wolfson Laboratories, Biochemistry Building, Department of Biological Sciences, South Kensington Campus, Imperial College London, London SW7 2AZ, UK*
[2]*Station Biologique, UMR 7127, Centre National de la Recherché Scientifique et Université Paris 6, BP74, F-29682 Roscoff Cedex, France*

***Prochlorococcus*, the most abundant genus of photosynthetic organisms[1], owes its remarkably large depth distribution in the oceans to the occurrence of distinct genotypes adapted to either low- or high-light niches[2,3]. The *pcb* genes, encoding the major chlorophyll-binding, light-harvesting antenna proteins in this genus[4], are present in multiple copies in low-light strains but as a single copy in high-light strains[5]. The basis of this differentiation, however, has remained obscure. Here we show that the moderate low-light-adapted strain *Prochlorococcus* sp. MIT 9313 has one iron-stress-induced *pcb* gene encoding an antenna protein serving photosystem I (PSI)—comparable to *isiA* genes from cyanobacteria[6,7]—and a constitutively expressed *pcb* gene encoding a photosystem II (PSII) antenna protein. By comparison, the very low-light-adapted strain SS120 has seven *pcb* genes encoding constitutive PSI and PSII antennae, plus one PSI iron-regulated *pcb* gene, whereas the high-light-adapted strain MED4 has only a constitutive PSII antenna. Thus, it seems that the adaptation of *Prochlorococcus* to low light environments has triggered a multiplication and specialization of Pcb proteins comparable to that found for Cab proteins in plants and green algae[8].**

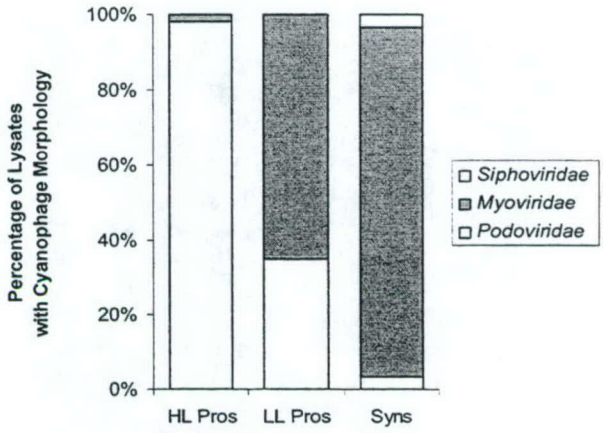
In order to gain a better understanding of the origin, function and localization of the divinyl-chlorophyll *a*/*b*-binding antenna complexes of *Prochlorococcus* species and how these properties relate to the light niche to which different strains are adapted, we have undertaken gene expression and structural studies on the moderate low-light-adapted strain MIT 9313 for which the full genome sequence is available (http://www.jgi.doe.gov/JGI_microbial/html/). This strain contains two *pcb* genes: *pcbA* (PMT 1046) and *pcbB* (PMT 0496). This contrasts with the very low-light-adapted strain SS120 that contains eight *pcb* genes (*pcbA* to *pcbH*)—analysis of its genome sequence (http://www.sb-roscoff.fr/Phyto/ProSS120/) revealed the presence of one more *pcb* gene (*pcbH*) than previously thought[5]—and the high-light-adapted strain MED4, which has only a single *pcb* gene (*pcbA*)[4,5].

Recently it was shown by electron microscopy and single-particle analyses that SS120 contains a giant supercomplex consisting of the PSI reaction centre trimer surrounded by a light-harvesting antenna ring composed of 18 Pcb subunits[9], similar to the 18-mer IsiA–PSI supercomplex induced in cyanobacteria when deprived of iron[6,7]. However studies on MIT 9313 grown under similar conditions to SS120 did not reveal the presence of an 18-mer Pcb–PSI supercomplex but only 'naked' trimeric PSI complexes, which matched with the cyanobacterial X-ray structure[10] (Fig. 1a, b). Instead, electron microscopy (Figs 1c, d and 2a) indicated that Pcb proteins associate with the dimeric reaction centre complex of PSII to form a Pcb–PSII supercomplex having dimensions of approximately 210 × 290 Å. Our interpretation is that this Pcb–PSII supercomplex consists of eight Pcb subunits with four distributed on each side of the PSII dimer as shown in Fig. 1c. This is emphasized by overlaying onto the projection map the published X-ray-derived models of the PSII reaction centre dimer and CP43, a PSII antenna protein structurally similar to Pcb[9,11] (Fig. 1d). In some cases, the four Pcb subunits on one side of the dimer were missing (Fig. 1e, f). The 'naked' PSI trimers and Pcb–PSII complexes shown in Fig. 1 were located in a chlorophyll-containing band (band 2 in Fig. 2b, insert +Fe) obtained by sucrose density centrifugation after solubilizing isolated thylakoid membranes with the detergent β-D-dodecyl maltoside. Also contained in this band were some PSII reaction centre dimers free of Pcb proteins (Fig. 1g, h). Analysis of all discernible particles, taken from band 2 sample micrographs, resulted in 1,192 particles assigned to PSI and PSII, and gave a PSI:PSII ratio of about 2. We assume this to be indicative of the ratio in the intact thylakoid membrane given that most PSI and PSII particles were in band 2. Amino-terminal sequencing of the Pcb protein in band 2 from +Fe conditions showed it to be the product of the *pcbA* gene only, a result which was also found for the free-Pcb proteins in band 1 and for Pcb protein in thylakoid membranes (Fig. 3).

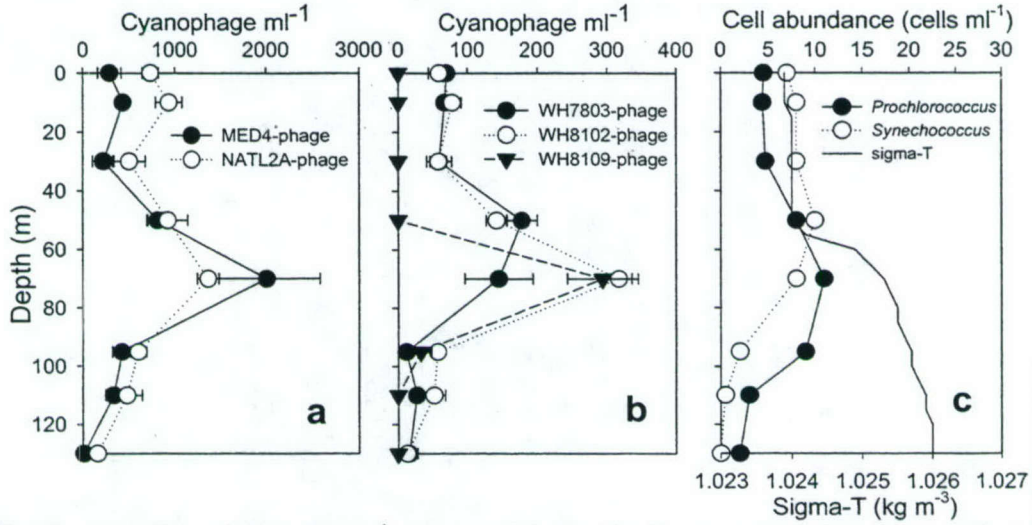
The absence of the PcbB protein and the 18-mer Pcb–PSI supercomplex in iron-replete MIT 9313 cells spurred us to investigate the expression of *pcbA* and *pcbB* genes in this strain. When the cells were grown in medium supplemented with iron (+Fe), we found that only the *pcbA* gene was expressed (Table 1). However, when cells were transferred to culture medium without added iron (−Fe), expression of the *pcbB* gene was activated, a surprising result as *pcb* genes of *Prochlorococcus* were not known to be regulated by iron as is the iron-stress-induced *isiB* gene (Table 1). The latter gene encodes flavodoxin[12], which substitutes for ferredoxin as an electron acceptor to PSI, and its expression indicates that the cells had acclimatized to conditions of iron depletion. On the other hand, the expression of

Supplementary Table 1: Detailed information about the *Prochlorococcus* and *Synechococcus* cyanophage isolates used in the host-range analyses in this study. TEM Morphology designations as follows: "P" = *Podoviridae*, "M" = *Myoviridae*, "S" = *Siphoviridae*, "+" indicates positive stained particles. Morphometric data for cyanophage are approximated where possible from negatively stained (except where indicated by '+') images without internal standards, "n.a." suggests no tail was observed.

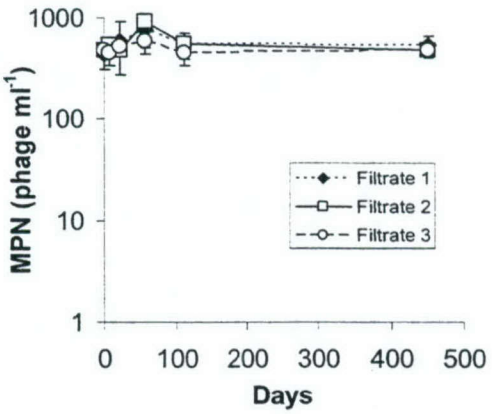
Phage Name	Locale	Latitude and Longitude	Depth (m)	Date Collected	TEM Morphology	Head Diameter (nm)	Tail Length x Width (nm)	Reference
-SSP1	BATS	31°48'N, 64°16'W	100	6 June 2000	P+	45+	n.a.	This study
-RSP1	Red Sea	29°28'N, 34°53'E	0	15 July 2000	P	55	6 x 10	This study
-RSP2	Red Sea	29°28'N, 34°53'E	0	15 July 2000	P+	40+	n.a.	This study
-SSP2	BATS	31°48'N, 64°16'W	120	29 Sep 1999	P+	40+	n.a.	This study
-SSP3	BATS	31°48'N, 64°16'W	100	29 Sep 1999	P	50	n.a.	This study
-SSP4	BATS	31°48'N, 64°16'W	70	26 Sep 1999	P	50	n.a.	This study
-SSP5	BATS	31°48'N, 64°16'W	120	29 Sep 1999	P	55	6 x 10	This study
-SSP6	BATS	31°48'N, 64°16'W	100	26 Sep 1999	P	55	n.a.	This study
-SSP7	BATS	31°48'N, 64°16'W	100	26 Sep 1999	P	50	n.a.	This study
-GSP1	Gulf Stream	38°21'N, 66°49'W	40	6 Oct 1999	P	45	n.a.	This study
-SSP8	BATS	31°48'N, 64°16'W	100	26 Sep 1999	P+	50	n.a.	This study
-RSP3	Red Sea	29°28'N, 34°55'E	50	13 Sep 2000	P+	50	n.a.	This study
-SP1	Slope	37°40'N, 73°30'W	83	17 Sep 2001	P	60	8 x 12	This study
yn-12	Gulf Stream	36°58'N, 73°42'W	0	Dec 1990	P	45	8 x 10	Waterbury & Valois 19
yn-5	Sargasso Sea	34°06'N, 61°01'W	0	July 1990	P			Waterbury & Valois 19
-SSM1	BATS	31°48'N, 64°16'W	100	6 June 2000	M+	60+	160 x 20	This study
-RSM1	Red Sea	29°28'N, 34°53'E	0	15 July 2000	M	80	110 x 27	This study
-ShM1	Shelf	39°60'N, 71°48'W	40	16 Sep 2001	M	75	120 x 18	This study
-ShM2	Shelf	39°60'N, 71°48'W	0	16 Sep 2001	M+			This study
-SSM2	BATS	31°48'N, 64°16'W	100	6 June 2000	M	85	110 x 25	This study
-SSM3	BATS	31°48'N, 64°16'W	100	6 June 2000	M	95	105 x 25	This study
-SSM4	BATS	31°48'N, 64°16'W	10	6 June 2000	M	80	160 x 20	This study
-SSM5	BATS	31°48'N, 64°16'W	15	26 Sep 1999	M+	60+	20 x 80+	This study
-SSM6	BATS	31°48'N, 64°16'W	40	29 Sep 1999	M	90	150 x 28	This study
-RSM2	Red Sea	29°28'N, 34°55'E	50	13 Sep 2000	M	75	170 x 23	This study
-RSM3	Red Sea	29°28'N, 34°55'E	50	13 Sep 2000	M	80	175 x 23	This study
-SM1	Slope	37°40'N, 73°30'W	0	17 Sep 2001	M	90	80+ x 25	This study
-ShM1	Shelf	39°60'N, 71°48'W	0	16 Sep 2001	M	80	120 x 25	This study
-SSM1	Sargasso Sea	34°24'N, 72°03'W	70	22 Sep 2001	M	95	150 x 25	This study
yn-2	Sargasso Sea	34°06'N, 61°01'W	0	July 1990	M	66	149 x 17	Waterbury & Valois 19
yn-9	Woods Hole	41°31'N, 71°40'W	0	Oct 1990	M	87	153 x 19	Waterbury & Valois 19
yn-19	Sargasso Sea	34°06'N, 61°01'W	0	July 1990	M			Waterbury & Valois 19
yn-10	Gulf Stream	36°58'N, 73°42'W	0	Dec 1990	M	100	145 x 19	Waterbury & Valois 19
yn-26	NE Providence Channel	25°53'N, 77°34'W	0	Jan 1992	M			Waterbury & Valois 19
yn-30	NE Providence Channel	25°53'N, 77°34'W	0	Jan 1992	M			Waterbury & Valois 19
yn-33	Gulf Stream	25°51'N, 79°26'W	0	Jan 1995	M			Waterbury & Valois 19
-PM2	English Channel	50°18'N, 4°12'W	0	23 Sep 1992	M	90	165 x 20	Wilson et al. 1993
-WHM1	Woods Hole	41°31'N, 71°40'W	0	11 Aug 1992	M	88	108 x 23	Wilson et al. 1993
yn-1	Woods Hole	41°31'N, 71°40'W	0	August 1990	M			Waterbury & Valois 19
-ShM2	Shelf	39°60'N, 71°48'W	0	16 Sep 2001	M+	70+	100 x 30	This study
-SSM2	Sargasso	34°24'N, 72°03'W	0	22 Sep 2001	M	80	225 x 22	This study
yn-14	Gulf Stream	36°58'N, 73°42'W	0	Dec 1990	M	93	136 x 21	Waterbury & Valois 19
-SS1	Slope	37°40'N, 73°30'W	60	17 Sep 2001	S+	45+ x 90	280 x 15	This study
-SS2	Slope	37°40'N, 73°30'W	83	17 Sep 2001	S	50 x 100	260 x 12	This study



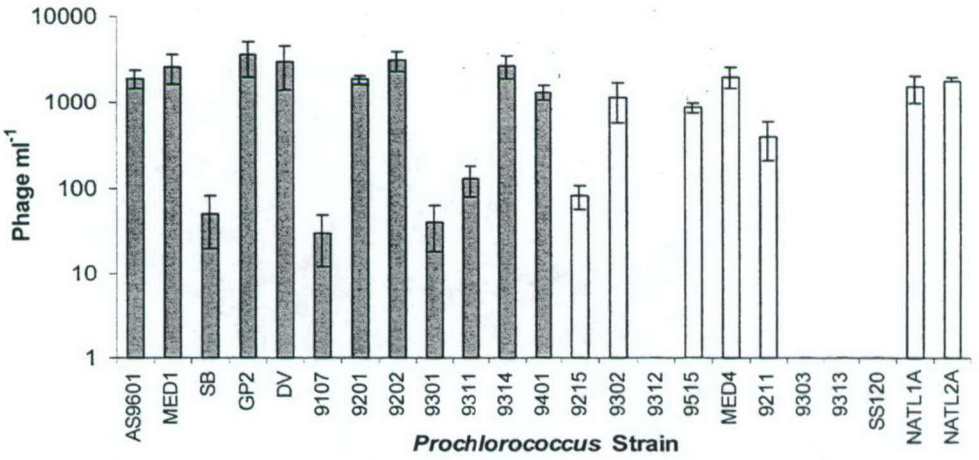
Supplementary Figure 1: Proportion of *Prochlorococcus* and *Synechococcus* cyanobacterial lysates with a given cyanophage morphology. Data represent presence or absence of a given morphology in each lysate and do not include observations of doubly plaque purified cyanophage isolates used in the host range analyses. The following number of observations were made for each category from the lysates: HL *Prochlorococcus* = 107, LL *Prochlorococcus* = 43, *Synechococcus* = 58.



Supplementary Figure 2: Cyanophage titers, measured using *Prochlorococcus* and *Synechococcus* hosts strains, as a function of depth in the oligotrophic western Sargasso Sea (34°24'N, 72°03'W) on 22 Sept. 2001. (a) titers measured using *Prochlorococcus* hosts, (b) titers measured using *Synechococcus* hosts, (c) total *Prochlorococcus* and *Synechococcus* cell abundances (*Prochlorococcus* cells x 10^4 ml^{-1} and *Synechococcus* cells x 10^3 ml^{-1}) and Sigma-T (σ_t — a proxy for water density used to measure the depth of the mixed layer). Titers were insignificant for host strains *Prochlorococcus* MIT9312, MIT9211, MIT9313, SS120 and *Synechococcus* WH6501, WH8018, WH8101. Error bars represent the standard deviation of assays.



Supplementary Figure 3: Water collected from Dyer's Dock, Woods Hole, MA on 3 Oct 2000 was immediately sub-sampled to prepare three separate filtrates to test the effects of long term storage on cyanophage titers using our storage methods. Cyanophage titers were measured using *Synechococcus* WH8012 in a MPN assay periodically (3 Oct 2000, 10 Oct 2000, 24 Oct 2000, 28 Nov 2000, 22 Jan 2001, 28 Dec 2001) over the course of 15 months. Data presented are the average and standard deviation of triplicate MPN assays from each sub-sample.



Supplementary Figure 4: Every isolate in the MIT *Prochlorococcus* culture collection with a unique ITS rDNA sequence was used in plaque assays to determine the strain-specific cyanophage titer from a water sample taken from 50 m depth in the Red Sea on 13 Sep 2000. These data were used to explore, for this one sample only, whether or not the strains we were using to assay phage titers throughout our studies (white bars) were yielding results that might be deemed "typical" with respect to the rest of the host cells in our collection (dark bars). The results show that none of the other host strains yield significantly higher titers than those used in this study. Data shown are average and standard deviation of triplicate plaque assays.

Chapter III

Diversity of portal protein genes from *Prochlorococcus* and *Synechococcus* myophage isolates

Sullivan, M. B., Rosenkrantz, J. R., Tan, G. P., Waterbury, J. B., and Chisholm, S. W. Diversity of portal genes from *Prochlorococcus* and *Synechococcus* myophage isolates. *Applied Environmental Microbiology* **in prep.**

Diversity of portal protein genes from *Prochlorococcus* and *Synechococcus* myophage isolates

Matthew B. Sullivan[1], Jessica E. Rosenkrantz[2], G. Tan[2], J.B. Waterbury[3], S.W. Chisholm[4*]

MIT / Woods Hole Oceanographic Institution Joint Program in Biological Oceanography, MIT, 48-424, Cambridge, Massachusetts 02139[1]; Massachusetts Institute of Technology, Department of Biology, Cambridge, Massachusetts 02139[2]; Woods Hole Oceanographic Institution, Department of Biology, Woods Hole, Massachusetts 02543[3]; Department of Civil and Environmental Engineering and Department of Biology, MIT, 48-425, Cambridge, Massachusetts 02139[4*]

The marine cyanobacteria *Prochlorococcus* and *Synechococcus* are the numerically dominant primary producers in the oceans, are globally distributed and are known to be infected by three morphologies of viruses (phage). The diversity of one of these families of phage, the myophage (phage of the *Myoviridae* morphology), has been assessed in both natural phage communities and cultured *Synechococcus* myophage isolates, using the g20 gene, a homolog to the gene encoding the T4 portal protein. These studies showed that environmental g20 sequences formed 9 distinct phylogenetic clusters, only 3 of which are represented in cultured myophage isolates. The recently assembled MIT phage collection includes a diverse selection of myophage, isolated using a broad range of cyanobacterial host strains (5 *Prochlorococcus* and 8 *Synechococcus* strains) from numerous sites in the Atlantic Ocean and the Gulf of Aqaba. We amplified and sequenced g20 fragments from this collection to see whether phage isolated on *Prochlorococcus* hosts differed from these published g20 sequences. We found that the g20 sequences from our collection clustered with the same three g20 clades represented by other myophage isolates. We suggest the six environmental g20 clades with no cultured representatives are likely to be from myophages that infect as yet uncultured hosts. Finally, the lack of obvious relationships between g20 diversity and a variety of factors associated with the phage isolates suggests the need for new taxonomic markers that allow the tracking of ecologically discrete groups of phage, such as phage that infect particular hosts.

The discovery that virus-like particles occur at high abundances (to 10^8 ml^{-1}) in the oceans (Bergh, 1989; Bratbak et al., 1990; Proctor and Fuhrman, 1990), has prompted efforts to elucidate the roles of viruses in these systems. To this end, there has been extensive work on the

phage-host system of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*, which are globally important primary producers in the oceans (Waterbury et al., 1986; Partensky et al., 1999). Their phages are abundant (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000; Lu et al., 2001; Frederickson et al., 2003; Marston and Sallee, 2003; Sullivan et al., 2003), are small but significant contributors to host mortality (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000), and are thought to play a role in maintaining the extensive microdiversity of the marine cyanobacteria (Waterbury and Valois, 1993; Suttle and Chan, 1994; Marston and Sallee, 2003; Sullivan et al., 2003).

Studying the diversity of phages has proven difficult because no universal gene, analogous to the 16S rRNA gene that is used as the taxonomic marker for all microbes, exists throughout all phage families (Paul et al., 2002). Thus, family-specific genes have recently been proposed for use as taxonomic tools (Rohwer and Edwards, 2002). For the *Myoviridae,* which are highly represented among *Synechococcus* cyanophage isolates (Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993; Sullivan et al., 2003), a fragment of the gene homologous to the coliphage T4 portal protein gene, g20, has been used for this purpose (Fuller et al., 1998; Zhong et al., 2002). The g20 homologue was initially chosen as a diversity indicator because hybridization studies of *Synechococcus* cyanomyophage DNA suggested this region was conserved across 8 phage isolates (Fuller et al., 1998). Subsequent study showed that the g20 homologue is part of a structural cassette (g18, g19, g20, g21, g22, g23) which is highly conserved among *Myoviridae* from hosts as divergent as proteobacteria and cyanobacteria (Hambly et al., 2001). The evolution of g20 is so constrained because its protein product (gp20) initiates capsid assembly in T4, a process involving geometric precision (Coombs and Eiserling, 1977; Hsiao and Black, 1978; van Driel and Couture, 1978) through the formation of a proximal vertex (van Driel and Couture, 1978) used for DNA packaging (Hsiao and Black, 1978) and binding the capsid to the tail junction (Coombs and Eiserling, 1977).

The high degree of conservation in the g20 gene has allowed the design of PCR primers to amplify g20 fragments from a range of myophage, facilitating diversity studies of as-yet uncultured naturally occurring myophage (Zhong et al., 2002). Myophage g20 diversity has been measured in a variety of aquatic environments, and offers a first glimpse at the diversity of these phage. Studies using non-degenerate PCR primers and evaluation of the amplicons using denaturing gradient gel electrophoresis (DGGE) banding patterns and terminal-restriction fragment length polymorphism (T-RFLP) have revealed variability in g20 diversity across space

and time. Along a transect from the Falkland Islands to the United Kingdom, 2-12 g20 DGGE bands were observed in one liter water samples (Wilson et al., 1999, 2000). A similar range in diversity was observed throughout a depth profile in waters off British Columbia (Frederickson et al., 2003) and a seasonal cycle in a freshwater lake in France (Dorigo et al., 2004) and in the estuarine waters of the Chesapeake Bay (Wang and Chen, 2004). These studies concluded that g20 diversity was as great within a sample as between oceans (Wilson et al., 1999), that phage g20 diversity increased as *Synechococcus* abundance increased (Wilson et al., 1999, 2000; Frederickson et al., 2003; Wang and Chen, 2004) and that some g20 types were ubiquitous (Wilson et al., 1999, 2000; Frederickson et al., 2003; Dorigo et al., 2004).

Cloning and sequencing of g20 PCR amplicons has allowed phylogenetic analyses of myophage diversity along a coastal-to-open Atlantic Ocean transect (Zhong et al., 2002) and over a 3-year period in coastal waters off Rhode Island (Marston and Sallee, 2003). Again, there was high variability in g20 diversity among sites, with between thirteen and twenty-nine different g20 sequences obtained at different sites along the transect (Zhong et al., 2002). While these authors conclude that there was some correlation between ocean habitat and g20 phylogenty (e.g., phylogenetic cluster II represents "oceanic" g20 sequences), further sampling suggested this was not the case, as seven g20 sequences from coastal *Synechococcus* myophage isolated from Rhode Island waters clustered with the putative "oceanic" sequences (Marston and Sallee, 2003). Clone libraries from the deep chlorophyll maximum had a higher diversity of g20 sequences than those from surface water samples in both the Gulf Stream and the Sargasso Sea (Zhong et al., 2002). Finally, g20 sequences observed in these field studies are not all represented in cultured isolates. The sequencing of 207 clones amplified from samples along the Atlantic Ocean transect revealed 114 unique g20 sequences (Zhong et al., 2002), which grouped into 9 phylogenetic clusters – only 3 of which had cultured respresentatives (Zhong et al., 2002). Subsequent culturing of phage isolated from Rhode Island coastal waters using *Synechococcus* hosts led to isolates whose g20 sequences also grouped with the 3 previous isolate-containing clusters (Marston and Sallee, 2003). Thus, 6 of the 9 environmental myophage g20 clusters lack cultured isolates.

The MIT phage collection (Sullivan et al., 2003) consists of phage isolated from seawater samples collected from various depths in the euphotic zone throughout the Atlantic Ocean and the Gulf of Aqaba. The phage were isolated using strains of *Prochlorococcus* and *Synechococcus* that represented the known genetic diversity of these host cells at the time this study began (Rocap et al., 2002) but see ref. (Fuller et al., 2003). Because the cultured myophage used in

previous g20 diversity studies were isolated using only *Synechococcus* strains, we wondered whether analysis of our collection would expand the database of g20 sequences for phage isolates, and possibly help explain the broad diversity observed in the field. However, as we report below, we found that g20 sequences from our isolates cluster only with the existing g20 sequences for cultured phage, and do not help identify the sequences from field samples for which there are no cultured counterparts.

MATERIALS AND METHODS

Myophage isolates. 45 cyanobacterial myophage were isolated (Table 1) as described previously (Waterbury and Valois, 1993; Wilson et al., 1993; Marston and Sallee, 2003; Sullivan et al., 2003). S-PM2 and S-WHM1 were provided by W. Wilson and all S-RIM phages were provided by M. Marston. The specificity of cyanomyophage g20 primers was tested using five marine *Pseudoalteromonas* spp. bacteriophage (HER320, HER321, HER322, HER327, HER328; (Wichels et al., 1998) that were purchased from the Felix d'Herelle Reference Center for Bacterial Viruses (contact H. Ackermann) as well as 7 heterotrophic bacteriophage (IH6-ϕ1, IH6-ϕ7, IH11-ϕ2, IH11-ϕ5, CB8-ϕ2, CB8-ϕ6, CB-ϕ8; (Zhong et al., 2002) kindly provided by F. Chen.

Testing of published primer sets: The published (Wilson et al., 1999; Zhong et al., 2002) g20 primer sets (CPS4/CPS5 used in DGGE studies, and CPS1/CPS8 used in sequencing studies) were designed without sequence information from *Prochlorococcus* myophage. We found that CPS4/CPS5 only amplified g20 sequences from 80% of these new *Prochlorococcus* and *Synechococcus* myophage in our collection, while CPS1/CPS8 only amplified g20 sequences from 44% of these isolates (Table 1). While the CPS4GC/CPS5 primer set amplifies g20 from most of our myophage isolates, the PCR product is too small (~165 bp) for subsequent phylogenetic analyses. In contrast, the CPS1/CPS8 primer set amplifies a larger PCR product, but from fewer isolates (25 of 45; Table 1).

To obtain g20 PCR amplicons from myophage that would not amplify using published primers, we added degeneracies to both CPS1 and CPS8, and shifted the CPS8 primer based upon genomic sequence data obtained for two of the *Prochlorococcus* myophage isolates (P-SSM2, P-SSM4; http://www.jgi.gov/JGI_microbial/html/index.html). CPS1.1 5'-GTAGWATWTTYTAYATTGAYGTWGG-3' and CPS8.1 5'-

ARTAYTTDCCDAYRWAWGGWTC-3'. This redesigned primer set (CPS1.1/CPS8.1) produced the expected size of PCR amplicons (~594 bp) from all 45 cyanomyophage isolates (Table 1) and when sequenced these amplicons proved to be from g20 homologues (Figure 1). Despite their degeneracy, the redesigned CPS1.1/CPS8.1 primer set was able to specifically amplify g20 sequences from all of our cyanobacterial myophage isolates as shown in specificity testing (Table 1) of the primers against 7 heterotrophic marine bacteriophage from Maryland, USA waters (Zhong et al., 2002), 5 heterotrophic marine bacteriophages from the North Sea (Wichels et al., 1998), as well as 16 *Podoviridae* and 2 *Siphoviridae* that infect *Synechococcus* and *Prochlorococcus* (Sullivan et al., 2003) .

PCR amplification and sequencing. Previous g20 PCR primer sets (non-degenerate CPS4GC/CPS5 (Wilson et al., 1999) and degenerate CPS1/CPS8 (Fuller et al., 1998; Zhong et al., 2002) were designed to amplify ~200bp and ~592 bp fragments, respectively, of the T4 g20 homologue in myophage.

PCR reactions for CPS4GC/CPS5 and CPS1/CPS8 were conducted as described previously (Wilson et al., 1999; Zhong et al., 2002). Briefly, 2 µl of cyanophage lysate was added as DNA template to a PCR reaction mixture (total volume 50 µl) containing the following: 20 pmol each of a forward and reverse primer, 1x PCR buffer (50mM Tris-HCl, 100 mM NaCl, 1.5 mM $MgCl_2$), 250 µM of each dNTP, and 0.75 U of Expand High Fidelity DNA polymerase (Roche, Indianapolis, IN). PCR amplification was carried out with a PTC-100 DNA Engine Thermocycler (MJ Research, San Francisco, CA). Optimized thermal cycling conditions varied slightly from those reported as follows: CPS4GC/CPS5 required an initial denaturation step of 94°C for 3 minutes, followed by 35 cycles of denaturation at 94°C for 1 minute, annealing at 50°C for 1 min, ramping at 0.3°C/s, and elongation at 73°C for 1 minute with a final elongation step at 73°C for 4 minutes, whereas both primer sets CPS1/CPS8 and CPS1.1/CPS8.1 required an initial denaturation step of 94°C for 3 minutes, followed by 35 cycles of denaturation at 94°C for 15s, annealing at 35°C for 1 min, ramping at 0.3°C/s, and elongation at 73°C for 1 minute with a final elongation step at 73°C for 4 minutes. Systematic PCR screening using various primer sets was conducted using the same PCR reaction conditions and amplification protocol, but replacing the High Fidelity DNA polymerase with the less expensive Taq DNA polymerase (Invitrogen, Carlsbad, CA) and only using 20 µl reactions since replicate (range 3-8) PCR reactions were

pooled before sequencing to decrease PCR bias (Polz and Cavanaugh, 1998). In all cases, a 5-10 µl aliquot of PCR product was analyzed in a 1.5% TAE gel stained with EtBr. The gel image was captured and analyzed with an Eagle Eye II gel documentation system (Stratagene, La Jolla, CA). For purification and sequencing, replicate PCR reactions were combined, run out on a 1.5% TAE gel and purified using the QIAGEN QIAquick gel extraction kit (Qiagen, Valencia, CA). The purified PCR products were sequenced directly on both strands using the degenerate PCR primers used to obtain the product (CPS1, CPS8, CPS1.1, CPS8.1) with best results at primer concentrations ~10-fold those suggested by the sequencing facility (40 pmol per reaction). To have greater confidence in negative PCR results, templates that did not produce amplified product were tested against optimized primer sets multiple times (data not shown).

Where identical g20 sequences were observed in our study, we confirmed the match was real and not the result of PCR contamination by re-amplifying and sequencing directly from fresh phage isolates (e.g., for P-SSM4, P-RSM3, S-SSM2, and "Syn" phages Syn2, Syn9, Syn10, Syn26, Syn30, Syn33, Syn1, Syn19) – many of which were obtained from stocks kept at a separate institution.

Phylogenetic analysis. Paired sequence data were aligned using ClustalW (Thompson et al., 1997) and corrected manually using the sequence chromatograms. Consensus sequences for each cyanophage isolate were then translated in-frame into amino acids. Multiple sequence alignments of translated amino acid consensus sequences were done with ClustalW using the Gonnet protein weight matrix, a gap opening penalty of 15 and gap extension penalty of 0.30 (although changing these penalties did not significantly alter the alignments). Phylogenetic reconstruction was done using PAUP 4.0 (Swofford, 2002) for parsimony and distance trees and Tree-Puzzle 5.0 (Schmidt et al., 2002) for maximum likelihood trees. Evolutionary distances for neighbor-joining trees were calculated based on mean character distances, while evolutionary distances for maximum likelihood trees were calculated using the JTT model of substitution assuming a gamma-distributed model of rate heterogeneities with 16 gamma-rate categories empirically estimated from the data. A heuristic search with 10 random addition replicates using the tree-bisection-reconnection branch swapping algorithm was used for parsimony trees. Bootstrap analysis was used to estimate node reproducibility and tree topology for neighbor-joining (1,000 replicates) and parsimony (100 replicates) trees, while quartet puzzling (10,000 replicates) indicates support

for the maximum likelihood tree. The g20 sequence from coliphage T4 was used as the outgroup taxon for all analyses.

Phylogenetic analyses of 183 amino acids from viral g20 sequence from 79 taxa yielded robust, similar trees using both algorithmic (neighbor-joining) and tree-searching (parsimony and maximum likelihood) methods. The translated g20 sequences contained phylogenetically informative regions (e.g., for parsimony analyses, 41 positions were constant, 25 were parsimony uninformative and 117 were parsimony informative). Differences between the parsimony, distance and maximum likelihood trees were limited to the branching order of the terminal nodes in a given cluster. To evaluate whether g20 sequence diversity correlated to a suite of phage isolation parameters, we empirically defined a "well supported node" as one where the average support across all three phylogenetic methods was 80% or greater.

Nucleotide sequence accession numbers. The nucleotide sequences determined in this study were submitted to GenBank and assigned accession numbers AYXXXXXX to AYXXXXXX.

RESULTS AND DISCUSSION

Using the g20 sequences obtained from our myophage collection and those in the database, we asked whether our myophage contained g20 sequences that were novel or that were clustered with environmental clades lacking cultured representatives. All 9 previously defined phylogenetic clusters (Zhong et al., 2002) were reproduced in each of our phylogenetic trees (Fig. 1). The g20 sequences of all 44 sequenced phage isolates (one isolate, S-RIM9, was screened but not sequenced) did not group with environmental clusters that lacked cultured representatives. The identical g20 sequences from phages P-SSM9, P-SSM11 and P-SSM12 along with that from S-RIM6 form a fourth monophyletic cluster within the clusters containing cultured representatives (I, II, III). Finally, while phylogenetic analyses grouped the g20 sequences from S-BnM1 and P-ShM1 with those from other cultured myophage, low bootstrap support made their placement within a particular cluster ambiguous.

These analyses suggest that while g20 sequences from our collection are often novel, they are found only in those g20 clusters containing cultured representatives. In other words, these new sequences do not help "identify" the environmental sequences in clusters that lack cultured representatives. We suggest that the 6 environmental g20 sequence groups lacking cultured representatives may be from two possible sources: phages that infect as yet uncultured

cyanobacteria, or phages that infect other surface-dwelling non-cyanobacterial hosts. Although the g20 primers used in the studies that yielded these sequences have been tested for non-specific amplification against a range of cultivated g20-containing bacteriophage, it is likely that in the high viral diversity of the oceans (Breitbart et al., 2002) the primers are not truly specific for cyanophage. If these g20 sequences are from bacteriophages, their relative abundance in the environmental g20 tree – 6 of the 9 clusters contain most of the environmental g20 sequences (Zhong et al., 2002) – qualitatively suggests that the hosts of the phages containing these g20 sequences may be common surface water microbes. Given this criterion, candidate hosts include *Pelagibacter* (formerly SAR11), *Roseobacter* (formerly SAR83), SAR86, and SAR116 (Giovannoni and Rappe, 2000).

As previously observed (Zhong et al., 2002; Marston and Sallee, 2003), the nucleotide and amino acid sequence divergence of the g20 regions amplified from our myophage isolates suggests the g20 portal protein gene is highly conserved. This region of the g20 homologue from coliphage T4 was 49.6-54.5% identical at the nucleotide and 39.8-45.3% identical at the amino acid level to the g20 sequences from our myophage isolates. Among the cyanomyophage g20 sequences, there was less than 50% divergence of both nucleotide and amino acid sequences, with ranges of pairwise identities from 59.8-100% nucleotide identity and 59.6-100% amino acid identity. Thirteen groups of g20 sequences from myophage isolates and environmental sequences contained identical amino acid sequences (numbered 1-13 in Fig. 1). Such high conservation of g20 sequences has been noted previously (Zhong et al., 2002; Marston and Sallee, 2003) and is not limited to the isolates of the MIT phage collection. Such striking g20 sequence conservation from cyanobacterial phages isolated from variable depths and/or geographical regions suggests these phages, due to the global distribution of their hosts, might be exchanging g20 genes throughout a global pool of phage genetic information (Hendrix et al., 1999).

With such a rich database of g20 sequence information, we wondered whether the g20 sequence clusters could be used to identify the host genus or host strain used to isolate a given phage isolate even though it is known that many of these phages can infect across a broad range of hosts (Sullivan et al., 2003). While none of the three culture-containing clusters (I, II, III) were comprised solely of g20 sequences from either *Prochlorococcus* or *Synechococcus* phages, all 10 g20 sequence clusters with identical amino acid sequences from cultures and 3 of the 6 well-supported sequence clusters (as defined in methods) were represented either by *Prochlorococcus* or *Synechococcus* phage, but not both. These latter observations suggest a non-random

distribution of g20 sequences within the clusters of the tree. However, it is important to bear in mind that the designation of *Prochlorococcus* or *Synechococcus* phage is artificially designated as the result of the genus of the cultured host originally used to isolate the phage. In fact, with one exception (cluster #10, Fig. 1), all phages represented in these clusters with identical g20 amino acid sequences, can infect both *Prochlorococcus* and *Synechococcus* (Sullivan et al., 2003).

An examination of the distribution of the original host strains used to isolate the phages showed that 6 of these 10 g20 sequence clusters with identical amino acid sequences (clusters #3, 4, 6, 7, 10, 11; Fig. 1) and none of the 6 well-supported clusters contained phages isolated using the same original host strains, while 4 of the identical g20 amino acid sequence clusters (clusters #5, 8, 9, 12; Fig. 1) and all 6 of the well-supported clusters did not. Further, there were some hosts (e.g., NATL2A, WH 7803) used to isolate many phages that contained g20 sequences that sometimes clustered together and sometimes were scattered throughout the tree (Fig. 1). Where complete host range information was available (i.e., tested against 11 *Prochlorococcus* and 10 *Synechococcus* strains), there also was no obvious relationship between g20 phylogeny and host range (Fig. 1). Of 9 clusters with identical g20 amino acid sequences from cultured phages where host range data were available, 6 had different host ranges (clusters # 3, 5, 6, 7, 9, 12; Fig. 1) while 3 had the same host ranges (clusters #4, 8, 10; Fig. 1).

Taken together, these data present a complicated scenario where sometimes g20 clustering appears non-random, but oftentimes lacks obvious correlations that might indicate which host strain or genus was used to isolate a phage. Further work to more rigorously address whether the clustering of these g20 sequences from similar phages (e.g., at the generic host level or at the single strain host level) is different from a random distribution of these g20 sequences should be done, perhaps as follows. There are 39 phage g20 sequences distributed within 6 well-supported (as defined in methods) clusters and 10 identical g20 amino acid clusters within this phylogeny. Using statistical randomization procedures (e.g., Monte Carlo simulations), one could evaluate whether the distribution of g20 sequences observed in our tree could occur randomly.

The lack of obvious relationships between g20 sequence clustering and the identification of original phage host or host range requires explanation. The host ranges of these myophages vary greatly, ranging from infecting a single host strain to infecting other ecotypes and even other genera (Sullivan et al., 2003). Therefore the "true" or optimal host, if such a construct exists for broad host range phages, might be any one of the host strains that it cross-infects (or an uncultured strain yet to be isolated). This muddies the waters considerably in trying to link the

evolutionary history (using g20 as a proxy) of myophages to that of their host of isolation or even to their apparent range of hosts, given the severe limitations in our understanding of the *in situ* reality for these phages in the environment. Further, the disconnect between g20 sequence clustering and host relationships is likely due to the function of g20. In coliphage T4, the g20 gene encodes a portal protein (Marusich and Mesyanzhinov, 1989) involved in functions quite removed from the direct interaction between phage and host. Thus, there is little reason to expect that the phylogenetic affiliation of g20 sequences would be related to host range. (Recall that in phage, where extensive horizontal gene transfer is known to occur (Hendrix et al., 1999; Hendrix, 2003), one gene might meaningfully correlate to a given property while another gene in the genome might represent a different mode of selection). Host range and the proteins mediating it are dynamic due to the ongoing phage-host 'arms race', while g20 remains highly conserved, through selective pressures that are unrelated to host identity. A better candidate gene whose phylogeny might map onto host range could be the distal tail fiber gene, which is known to be the direct determinants of host range in T-even coliphages (Henning and Hashemolhosseini, 1994).

The exploration of g20 diversity in this cyanophage collection has expanded the sequence database to include phage isolated using *Prochlorococcus* strains. Analyses presented here, (1) confirm that phage culture collections represent only a fraction of the g20 sequence diversity observed in the field, and (2) suggest that g20 sequence clusters do not identify the original host of isolation or host range of phage isolates. We are only beginning to explore ecological questions in phage biology and have sampled less than 0.0002% of the global phage metagenome (Rohwer, 2003). Current investigations of phage-host interactions are dramatically hampered by the lack of diversity markers for quantitatively and specifically tracking phages that infect particular hosts. New efforts to identify appropriate marker genes for tracking such diversity are critical to systematically understand phage-host dynamics in the oceans.

ACKNOWELEDGEMENTS

Funding was provided by a Biological Oceanography NSF Grant to SWC; MIT, the Seaver Insitute and an NIH Bioinformatics Training Grant supported MBS; MIT UROP supported GT and JER. Sequence data from the g20 region of cyanomyophage P-SSM2 and P-SSM4 was provided by the DOE JGI. F. Chen and M. Marston kindly provided phage isolate information as well as isolates used in testing of PCR primer sets. The comments of V. Rich greatly improved the manuscript.

Table 1: Detailed description of phage isolates used in optimization of g20 PCR primers and results of specificity testing of three PCR primer sets against various phage isolates.

Phages	Hosts	Site of Isolation	Depth (m)	Date isolated	Family[a]	CPS 4GC/5[b]	CPS 1/8[b]	CPS 1.1/8.1[b]	Ref[c]
Prochlorococcus cyanophage									
P-SSP1	MIT 9215	BATS / 31°48'N, 64°16'W	100	6 June 2000	P	–	–	–	1
P-RSP1	MIT 9215	Red Sea / 29°28'N, 34°53'E	0	15 July 2000	P	–	–	–	1
P-RSP2	MIT 9302	Red Sea / 29°28'N, 34°53'E	0	15 July 2000	P	–	–	–	1
P-SSP2	MIT 9312	BATS / 31°48'N, 64°16'W	120	29 Sep 1999	P	–	–	–	1
P-SSP3	MIT 9312	BATS / 31°48'N, 64°16'W	100	29 Sep 1999	P	–	–	–	1
P-SSP4	MIT 9312	BATS / 31°48'N, 64°16'W	70	26 Sep 1999	P	–	–	–	1
P-SSP5	MIT 9515	BATS / 31°48'N, 64°16'W	120	29 Sep 1999	P	–	–	–	1
P-SSP6	MIT 9515	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-SSP7	MED4	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-GSP1	MED4	Gulf Stream / 38°21'N, 66°49'W	40	6 Oct 1999	P	–	–	–	1
P-SSP8	NATL2A	BATS / 31°48'N, 64°16'W	100	26 Sep 1999	P	–	–	–	1
P-RSP3	NATL2A	Red Sea / 29°28'N, 34°55'E	50	13 Sep 2000	P	–	–	–	1
P-SP1	SS120	Slope / 37°40'N, 73°30'W	83	17 Sep 2001	P	–	–	–	1
P-SSM8	MIT 9211	W Sargasso Sea / 34°24'N, 72°03'W	30	22 Sept 2001	M	+	+	+	2
P-SSM1	MIT 9303	BATS / 31°48'N, 64°16'W	100	6 June 2000	M	+	–	+	1
P-RSM1	MIT 9303	Red Sea / 29°28'N, 34°53'E	0	15 July 2000	M	+	–	+	1
P-RSM4	MIT 9303	Red Sea / 29°28'N, 34°55'E	130	13 Sep 2000	M	+	+	+	2
P-ShM1	MIT 9313	Shelf / 39°60'N, 71°48'W	40	16 Sep 2001	M	+	–	+	1
P-ShM2	MIT 9313	Shelf / 39°60'N, 71°48'W	0	16 Sep 2001	M	–	–	+	1
P-SSM2	NATL1A	BATS / 31°48'N, 64°16'W	100	6 June 2000	M	+	+	+	1
P-RSM5	NATL1A	Red Sea / 29°28'N, 34°55'E	130	13 Sep 2000	M	+	+	+	2
P-SSM7	NATL1A	BATS / 31°48'N, 64°16'W	120	29 Sep 1999	M	–	–	+	2
P-SSM3	NATL2A	BATS / 31°48'N, 64°16'W	100	6 Jun 2000	M	–	–	+	1
P-SSM4	NATL2A	BATS / 31°48'N, 64°16'W	10	6 June 2000	M	–	–	+	1
P-SSM5	NATL2A	BATS / 31°48'N, 64°16'W	15	26 Sep 1999	M	+	–	+	1
P-SSM6	NATL2A	BATS / 31°48'N, 64°16'W	40	29 Sep 1999	M	–	–	+	1
P-RSM2	NATL2A	Red Sea / 29°28'N, 34°55'E	50	13 Sep 2000	M	+	–	+	1
P-RSM3	NATL2A	Red Sea / 29°28'N, 34°55'E	50	13 Sep 2000	M	–	–	+	1
P-SSM9	NATL2A	W Sargasso Sea / 34°24'N, 72°03'W	0	22 Sep 2001	M?	+	–	+	2
P-SSM10	NATL2A	W Sargasso Sea / 34°24'N, 72°03'W	0	22 Sep 2001	M?	+	–	+	2
P-SSM11	NATL2A	W Sargasso Sea / 34°24'N, 72°03'W	0	22 Sep 2001	M?	+	–	+	2
P-SSM12	NATL2A	W Sargasso Sea / 34°24'N, 72°03'W	95	22 Sep 2001	M?	+	–	+	2
Synechococcus cyanophage									
Syn5	WH 8109	Sargasso Sea / 36°58'N, 73°42'W	0	Dec 1990	P	–	–	–	1
Syn12	WH 8017	Gulf Stream / 34°06'N, 61°01'W	0	July 1990	P	–	–	–	1
S-SM1	WH 6501	Slope / 37°40'N, 73°30'W	0	17 Sep 2001	M	–	–	+	1
S-ShM1	WH 6501	Shelf / 39°60'N, 71°48'W	0	16 Sep 2001	M	+	+	+	1
S-SSM1	WH 6501	W Sargasso Sea / 34°24'N, 72°03'W	70	22 Sep 2001	M	+	+	+	1
Syn 2	WH 8012	Sargasso Sea / 34°06'N, 61°01'W	0	July 1990	M	–	+	+	3
Syn 9	WH 8012	Woods Hole / 41°31'N, 71°40'W	0	Oct 1990	M	+	+	+	3
Syn 10	WH 8017	Gulf Stream / 36°58'N, 73°42'W	0	Dec 1990	M	+	+	+	3
Syn 26	WH 8017	NE Providence Channel / 25°53'N, 77°34'W	0	Jan 1992	M	+	+	+	3
S-SM2	WH 8017	Slope / 37°40'N, 73°30'W	15	17 Sep 2001	M	+	–	+	2
Syn30	WH 8018	NE Providence Channel / 25°53'N, 77°34'W	0	Jan 1992	M	+	–	+	3
S-SSM3	WH 8018	W Sargasso Sea / 34°24'N, 72°03'W	0	22 Sep 2001	M	+	+	+	2
S-SSM4	WH 8018	W Sargasso Sea / 34°24'N, 72°03'W	110	22 Sep 2001	M	+	+	+	2
S-RIM3	WH 8018	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	Sept. 1999	M?	+	–	+	4
Syn 33	WH 7803	Gulf Stream / 25°51'N, 79°26'W	0	Jan 1995	M	+	+	+	3
S-PM2	WH 7803	English Channel / 50°18'N, 4°12'W	0	23 Sep 1992	M	+	+	+	5
S-WHM1	WH 7803	Woods Hole / 41°31'N, 71°40'W	0	11 Aug 1992	M	+	+	+	5
S-RIM9	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	May 2000	M?	+	–	+	4
S-RIM17	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	July 2001	M?	+	–	+	4
S-RIM24	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	Dec 2001	M?	+	–	+	4
S-RIM30	WH 7803	Mt. Hope Bay, RI / 41°39'N, 71°15'W	0	June 2002	M?	+	–	+	4

Phages	Hosts	Site of Isolation	Depth (m)	Date isolated	Family[a]	CPS 4GC/5[b]	CPS 1/8[b]	CPS 1.1/8.1[b]	Ref[c]
Syn 1	WH 8101	Woods Hole / 41°31'N, 71°40'W	0	Aug 1990	M	+	–	+	3
S-ShM2	WH 8102	Shelf / 39°60'N, 71°48'W	0	16 Sep 2001	M	+	+	+	1
S-SSM2	WH 8102	W Sargasso Sea / 34°24'N, 72°03'W	0	22 Sep 2001	M	+	+	+	1
S-SSM5	WH 8102	W Sargasso Sea / 34°24'N, 72°03'W	95	22 Sep 2001	M	+	+	+	2
Syn 19	WH 8109	Sargasso Sea / 34°06'N, 61°01'W	0	July 1990	M	–	–	+	3
S-SSM6	WH 8109	W Sargasso Sea / 34°24'N, 72°03'W	70	22 Sep 2001	M	+	+	+	2
S-SSM7	WH 8109	W Sargasso Sea / 34°24'N, 72°03'W	95	22 Sep 2001	M	+	+	+	2
Other phages									
IH6-ϕ1	IH6	Inner Harbor, Baltimore, MD	0	17 Nov 2000	M	–	–	–	6
IH6-ϕ7	IH6	Inner Harbor, Baltimore, MD	0	17 Nov 2000	P	–	–	–	6
IH11-ϕ2	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 Nov 2000	M	–	–	–	6
IH11-ϕ5	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 Nov 2000	P	–	–	–	6
CB8-ϕ2	CB8	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
CB8-ϕ6	CB8	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
CB-ϕ8	*Vibrio alginolyticus*	Chesapeake Bay, MD	0	17 Nov 2000	M	–	–	–	6
HER320	H7	Helgoland, North Sea	0	1976-1978	M	–	–	–	7
HER321	H100	Helgoland, North Sea	0	1976-1978	P	–	–	–	7
HER322	H100	Helgoland, North Sea	0	1976-1978	M	–	–	–	7
HER327	11-68	Helgoland, North Sea	0	1976-1978	S	–	–	–	7
HER328	H105	Helgoland, North Sea	0	1976-1978	S	–	–	–	7

(Table 1 continued)

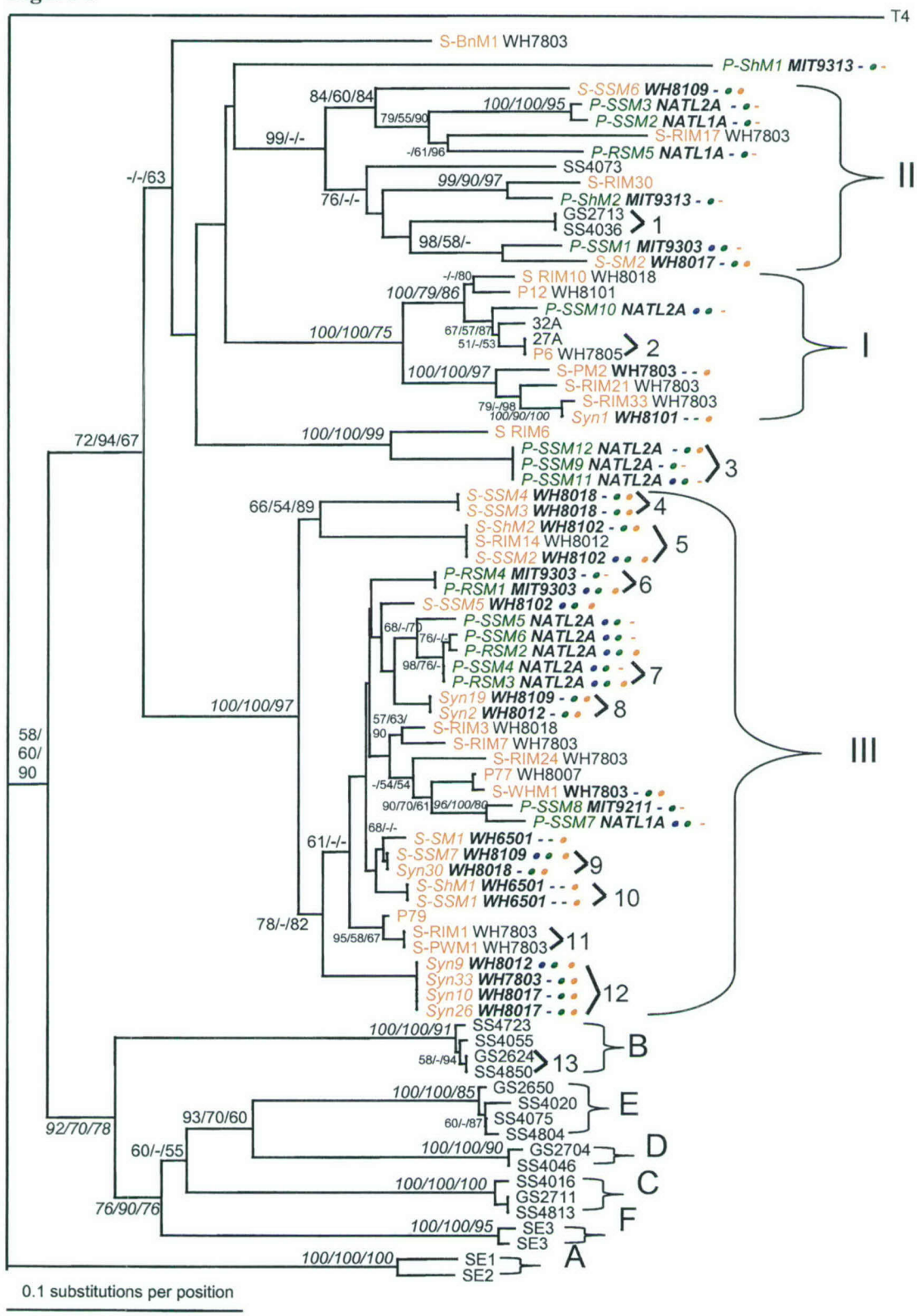
[a] M, P and S represent the virus families *Myoviridae, Podoviridae* and *Siphoviridae*, respectively. ? indicates the morphology of the phage particle has not been confirmed with electron microscopy, but is presumably a *Myoviridae* based upon amplification and sequencing of a g20 PCR product.

[b] +, positive PCR amplification; –, no desired PCR product.

[c] References code: 1 = Sullivan et al, 2003; 2 = This study; 3 = Waterbury & Valois; 1993; 4 = Marston & Salee, 2003; 5 = Wilson et al., 1993; 6 = Zhong et al., 2002; 7 = Wichels et al., 1998

Figure 1: Evolutionary relationships determined using 183 amino acids of the portal protein gene (g20) amplified from MIT myophage isolates (colored and *italicized*), previously characterized myophage isolates (colored), and environmental g20 sequences (Zhong et al., 2002; Marston and Sallee, 2003). The tree shown was inferred by neighbor-joining as described in the methods. Support values shown at the nodes are neighbor-joining bootstrap / maximum parsimony bootstrap / maximum likelihood quartet puzzling support (values less than 50 are designated with a dash). Well supported nodes (as defined in methods) are designated by italicized support values. Clusters were assigned as designated by Zhong et al. (2002); clusters I, II and III contain g20 sequences from cultured phage isolates, while clusters A-F represent environmental g20 sequences. Clusters containing g20 sequences that are identical are numbered with alphanumeric numbers (1-13). For cultured phage, colored isolate names indicate whether they were originally isolated using a *Synechococcus* (orange) or *Prochlorococcus* (green) host; black lettering following phage isolate names indicates the original host strain used for isolation, and colored dots indicates that the phage cross-infects at least one strain from the high-light adapted *Prochlorococcus* (blue), the low-light adapted *Prochlorococcus* (green) or *Synechococcus* (orange), whereas colored dashes indicates no cross-infection among those ecotypes. Isolates not available for host range testing have no indication of their host range.

Figure 1


T4
S-BnM1 WH7803
P-ShM1 MIT9313
S-SSM6 WH8109
P-SSM3 NATL2A
P-SSM2 NATL1A
S-RIM17 WH7803
P-RSM5 NATL1A
SS4073
S-RIM30
P-ShM2 MIT9313
GS2713
SS4036
P-SSM1 MIT9303
S-SM2 WH8017
S RIM10 WH8018
P12 WH8101
P-SSM10 NATL2A
32A
27A
P6 WH7805
S-PM2 WH7803
S-RIM21 WH7803
S-RIM33 WH7803
Syn1 WH8101
S RIM6
P-SSM12 NATL2A
P-SSM9 NATL2A
P-SSM11 NATL2A
S-SSM4 WH8018
S-SSM3 WH8018
S-ShM2 WH8102
S-RIM14 WH8012
S-SSM2 WH8102
P-RSM4 MIT9303
P-RSM1 MIT9303
S-SSM5 WH8102
P-SSM5 NATL2A
P-SSM6 NATL2A
P-RSM2 NATL2A
P-SSM4 NATL2A
P-RSM3 NATL2A
Syn19 WH8109
Syn2 WH8012
S-RIM3 WH8018
S-RIM7 WH7803
S-RIM24 WH7803
P77 WH8007
S-WHM1 WH7803
P-SSM8 MIT9211
P-SSM7 NATL1A
S-SM1 WH6501
S-SSM7 WH8109
Syn30 WH8018
S-ShM1 WH6501
S-SSM1 WH6501
P79
S-RIM1 WH7803
S-PWM1 WH7803
Syn9 WH8012
Syn33 WH7803
Syn10 WH8017
Syn26 WH8017
SS4723
SS4055
GS2624
SS4850
GS2650
SS4020
SS4075
SS4804
GS2704
SS4046
SS4016
GS2711
SS4813
SE3
SE3
SE1
SE2
I
II
III
A
B
C
D
E
F
0.1 substitutions per position

REFERENCES

Bergh, O. (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467-468.

Bratbak, G., Heldal, M., Norland, S., and Thingstad, T.F. (1990) Viruses as partners in spring bloom microbial trophodynamics. *Appl Environ Microbiol* **56**: 1400-1405.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D. et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Coombs, D., and Eiserling, F.A. (1977) Studies on the structure, protein composition and assembly of the neck of bacteriophage T4. *J. Mol. Biol.* **116**: 375-407.

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004) Cyanophage diversity, inferred from g20 gene analyses, in the Largest Natural Lake in France, Lake Bourget. *Appl Environ Microbiol* **70**: 1017-1022.

Frederickson, C.M., Short, S.M., and Suttle, C.A. (2003) The physical environment affects cyanophage communities in British Columbia Inlets. *Microb Ecol* **46**: 348-357.

Fuller, N.J., Wilson, W.H., Joint, I.R., and Mann, N.H. (1998) Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl Environ Microbiol* **64**: 2051-2060.

Fuller, N.J., Marie, D., Partensky, F., Vaulot, D., Post, A.F., and Scanlan, D.J. (2003) Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine Synechococcus clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69**: 2430-2443.

Giovannoni, S., and Rappe, M. (2000) Evolution, diversity, and molecular ecology of marine prokaryotes. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed). New York: Wiley-Liss, pp. 47-84.

Hambly, E., Tetart, F., Desplats, C., Wilson, W.H., Krisch, H.M., and Mann, N.H. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci U S A* **98**: 11411-11416.

Hendrix, R.W. (2003) Bacteriophage genomics. *Curr Opin Microbiol* **6**: 506-511.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* **96**: 2192-2197.

Henning, U., and Hashemolhosseini, S. (1994) Recepor recognition by T-even-type coliphages. In *Molecular biology of bacteriophage T4*. Karam, J. (ed). Washington D.C.: ASM Press, pp. 291-298.

Hsiao, C.L., and Black, L.W. (1978) Head morphogenesis of bacteriophage T4. III. The role of g20 in DNA packaging. *Virology* **91**: 26-38.

Lu, J., Chen, F., and Hodson, R.E. (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine Synechococcus spp. in river estuaries. *Appl Environ Microbiol* **67**: 3285-3290.

Marston, M.F., and Sallee, J.L. (2003) Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters. *Appl Environ Microbiol* **69**: 4639-4647.

Marusich, E.I., and Mesyanzhinov, V.V. (1989) Nucleotide and deduced amino acid sequence of bacteriophage T4 gene 20. *Nucleic Acids Res* **17**: 7514.

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

Paul, J.H., Sullivan, M.B., Segall, A.M., and Rohwer, F. (2002) Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**: 463-476.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64**: 3724-3730.

Proctor, L.M., and Fuhrman, J.A. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* **343**: 60-62.

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

Rohwer, F. (2003) Global phage diversity. *Cell* **113**: 141.

Rohwer, F., and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529-4535.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502-504.
Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophage infecting the oceanic cyanobacterium, *Prochlorococcus*. *Nature* **424**: 1047-1051.
Suttle, C.A. (2000) Cyanophages and their role in the ecology of cyanobacteria. In *The Ecology of Cyanobacteria*. Whitton, B.A., and Potts, M. (eds). Netherlands: Kluwer Academic Publishers, pp. 563-589.
Suttle, C.A., and Chan, A.M. (1993) Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92**: 99-109.
Suttle, C.A., and Chan, A.M. (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Applied. Environ. Microbiol.* **60**: 3167-3174.
Swofford, D.L. (2002) *PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.* Sunderland, MA: Sinauer Associates.
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.
van Driel, R., and Couture, E. (1978) Assembly of the scaffolding core of bacteriophage T4 proheads. *J. Mol. Biol.* **123**: 713-719.
Wang, K., and Chen, F. (2004) Genetic diversity and population dynamics of cyanophage communities in the Chesapeake Bay. *Aq. Microb. Ecol.* **34**: 105-116.
Waterbury, J.B., and Valois, F.W. (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**: 3393-3399.
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.* **214**: 71-120.
Wichels, A., Biel, S.S., Gelderblom, H.R., Brinkhoff, T., Muyzer, G., and Schutt, C. (1998) Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* **64**: 4128-4133.
Wilson, W.H., Joint, I.R., Carr, N.G., and Mann, N.H. (1993) Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl Environ Microbiol* **59**: 3736-3743.
Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (1999) Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. In *Marine Cyanobacteria*. Sharpy, L., and Larkum, A.W.D. (eds). Monaco: Bulletin de l'Institut oceanographique, pp. 209-216.
Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (2000) Analysis of cyanophage diversity in the marine environment using denaturing gradient gel electrophoresis. In *Microbial Biosystems: New Fronteirs. Proceedings of the 8th International Symposium on Microbial Ecology*. Bell, C.R., Brylinksky, M., and Johnson-Green, P. (eds). Halifax, Nova Scotia, Canada: Atlantic Canada Society for Microbial Ecology.
Zhong, Y., Chen, F., Wilhelm, S.W., Poorvin, L., and Hodson, R.E. (2002) Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl Environ Microbiol* **68**: 1576-1584.

Chapter IV

Transfer of photosynthetic genes to and from *Prochlorococcus* phage

Lindell, D.* & Sullivan, M. B.*, Johnson, Z. I., Tolonen, A., Rohwer, F., and Chishiolm, S. W. Transfer of photosynthesis genes to and from *Prochlorococcus* phage. *Proc Natl Acad Sci U S A*, in review.

* These authors contributed equally to this work

SUBMITTED TO PNAS (Track II) 3 MARCH, 2004.

Biological Sciences: Evolution

TRANSFER OF PHOTOSYNTHESIS GENES TO AND FROM *PROCHLOROCOCCUS* PHAGE

Debbie Lindell*[1], Matthew B. Sullivan*[2], Zackary I. Johnson[1], Andrew C. Tolonen[2], Forest Rohwer[3], Sallie W. Chisholm[1,4]

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2] Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution and Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[3] Department of Biology, San Diego State University, San Diego, California 92182, USA
[4] Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

*These authors contributed equally to this work.

Corresponding author:
Sallie W. Chisholm
Dept. Civil Environ. Eng.
MIT 48-425
77 Massachusetts Ave
Cambridge, MA 02139
Phone: 617 253 1771
Fax: 617 258 7009
Email: chisholm@mit.edu

Manuscript Information: 25 text pages, 3 Figures, 1 Table
Abstract contains 249 words. Total number of characters is 46,894.
Supplementary information includes 6 Figures and 3 Tables

Abbreviations: PSII, photosystem II; HLIP, high-light inducible protein

Data Deposition: The phage genome sequences and *Prochlorococcus psbA* sequences reported in this paper have been deposited in GenBank under accession numbers: (currently being processed).

Whole genome sequencing gives us a new window for viewing phage-host interactions and their evolutionary implications. Here we report the presence of genes central to oxygenic photosynthesis in the genomes of three phage from 2 families of viruses (*Myoviridae, Podoviridae*) that infect the marine cyanobacterium, *Prochlorococcus.* The gene that encodes the photosystem II (PSII) core reaction center protein D1 (*psbA*), and a gene (*hli*) that encodes one high light inducible protein (HLIP) type are present in all 3 phage genomes. The two myoviruses contain additional *hli* gene types, and one of them contains *psbD*, which encodes the second PSII core reaction center protein, D2, and the other contains photosynthetic electron transport genes coding for plastocyanin (*petE*) and ferredoxin (*petF*). These uninterrupted, full-length genes are conserved in their amino acid sequence, suggesting that they encode functional proteins that may help maintain photosynthetic activity during infection. Phylogenetic analyses show that phage D1, D2 and HLIP proteins cluster with those from *Prochlorococcus,* indicating that they are of cyanobacterial origin. Their distribution among several *Prochlorococcus* clades further suggests that the genes encoding these proteins were transferred from host to phage multiple times. Phage HLIPs cluster with multicopy types found exclusively in *Prochlorococococus,* suggesting that phage may be mediating the expansion of the *hli* gene family by transferring these genes back to their hosts after a period of evolution in the phage. These gene transfers are likely to play a role in the fitness landscape of host and phage in the surface oceans.

The genomes of viruses (phage) that infect bacteria contain a variety of genes homologous to those found in their bacterial hosts (1-5). Many encode functional proteins involved in processes of direct importance for the production of phage progeny. They include genes involved in DNA replication, nucleotide metabolism and RNA transcription and are found in the genomes of both lytic phage and prophage (3, 6). It is likely that many originated from their hosts (2, 4), and that some host genes that occur in multiple copies have been (re)acquired from phage (2, 7) — either after a period of evolution in the phage or after acquisition of the gene from a different host.

Host genes that are not directly related to the production of new phage, such as genes involved in phosphate sensing and metabolism (8, 9) and the scavenging of oxygen radicals (10) are also found in phage genomes, and may benefit phage by temporarily enhancing host functionality prior to lysis. In addition, prophage can provide their hosts with new functions by encoding genes not otherwise found in the host's genome, such as virulence factors, toxin production genes, and genes involved in the immune response to pathogenic infection (5, 6, 11).

Genes involved in photosynthesis have recently been found in a lytic phage isolated on *Synechococcus* WH7803 (12), a member of the marine cluster A unicellular cyanobacteria that is widespread in the oceans. A member of the *Myoviridae* family of dsDNA viruses, this phage contains 2 photosynthetic genes (*psbD* and an interrupted *psbA* gene) that code for the 2 photosystem II (PSII) core reaction center proteins found in all photosynthetic organisms. These genes were not found in another phage — a member of the *Podoviridae* family — isolated on the same strain of *Synechococcus* (13). This leads one to wonder whether the presence of photosynthetic genes in phage is a rare phenomenon and to what extent it is specific for a particular phage or host type. If these genes are widespread and diverse in cyanophage, what is their origin? Were they acquired through a single ancestral transfer event?

The phage-host system for *Prochlorococcus* and *Synechococcus* (14, 15), which form a monophyletic clade within the cyanobacteria (16-19), is well suited to begin to answer these questions. Members of each genus form distinct sub-genera clusters within this clade, which in *Prochlorococcus* also correspond to their efficiency of light utilization (17). Numerous phage have been isolated using this diverse group, including members of the *Myoviridae, Podoviridae* and *Siphoviridae* families, and the degree of cross infection — a mechanism for horizontal gene

transfer — among and between strains has been analyzed (14, 15). The genomes of 4 of the host strains have been published (20-22), and the genomes of three phage have been sequenced by the US Department of Energy Joint Genome Institute (DOE JGI; www.jgi.doe.gov), providing a database to begin an analysis of the distribution of host genes among hosts and the phage that infect them, and exploring their phylogenetic relationships.

Here we report that the genomes of 3 phage that infect *Prochlorococcus* collectively contain a number of host-like photosynthetic genes. We further infer from bioinformatic analyses that they are likely to play a functional role during infection, as well as impact the evolutionary trajectory of both phage and host in the surface oceans.

Materials and Methods

Selection and preparation of cyanophage for genome sequencing

Two myoviruses and 1 podovirus were chosen for sequencing based on their host range within *Prochlorococcus*, with no prior knowledge of their gene content. The podovirus P-SSP7 (43 kb genome), infects a single high-light adapted (HL) *Prochlorococcus* strain. The myovirus P-SSM2 (252 kb) infects 3 low-light adapted (LL) *Prochlorococcus* strains and the P-SSM4 myovirus (178 kb) infects 2 HL and 2 LL *Prochlorococcus* strains ((15), see Table I). None of these phage cross-infects any of the *Synechococcus* strains tested.

Phage were propagated on their respective *Prochlorococcus* hosts (P-SSP7 on MED4, P-SSM2 on NATL1A, P-SSM4 on NATL2A) and were purified for DNA extraction and the construction of clone libraries as described previously (8). Briefly, 1L of lysed culture was treated with DNase and RNase to degrade host nucleic acids. Cell debris were removed and the phage remaining in the supernatant were precipitated using PEG 8000. Concentrated phage were purified on a cesium chloride step gradient (steps were ρ=1.30, 1.40, 1.50, 1.65; the gradient was spun at 2 hour, 4°C, 104,000 Xg) and dialyzed against a buffer containing 100mM TrisCl (pH 7.5), 100 mM $MgSO_4$ and 30 mM NaCl. Purified phage were burst using SDS (0.5%) and proteinase K (50μg.ml^{-1}), DNA was extracted with phenol:chloroform and concentrated by ethanol precipitation. A custom LASL clone library was constructed by Lucigen Inc (Middleton, WI) as described previously (23). Inserts were sequenced and genomes assembled by the DOE JGI. All analyses were conducted on the phage genomes as provided on 17 Oct 03 (P-SSM2, P-SSM4) and 19 Nov 03

(P-SSP7). At that time, these genomes were in large high-quality contigs compiled from 26-fold (P-SSP7), 30-fold (P-SSM2) and 39-fold (P-SSM4) coverage, respectively. The P-SSP7 and P-SSM4 genomes were not yet closed.

PCR amplification of *psbA*

Genomic DNA was isolated from *Prochlorococcus* cultures using the DNEasy kit (Qiagen, Valencia, CA). Partial *psbA* sequences were amplified using primers from (19) or for *Prochlorococcus* MIT9211 using the following primers (5'-AACATCATYTCWGGTGCWGT-3') and (5'-TCGTGCATTACTTCCATACC-3'). Reactions (50μl) consisted of 4mM $MgCl_2$, 200μM each dNTP, 0.25μM each primer, 2.5 units *Taq* DNA polymerase (Invitrogen, Carslbad, CA), 4ng genomic DNA. Amplification conditions, run on a RoboCycler Gradient 96 thermocycler (Stratagene, La Jolla, CA), comprised steps at 92°C for 4 min, 35 cycles at 92°C for 1 min, 50°C for 1 min and 68°C for 1 min, followed by a final extension step at 68°C for 10 min. PCR products were gel purified and sequenced in both forward and reverse directions (Davis Sequencing).

Identification of genes and transcriptional regulatory elements

Assembled phage genomes were examined to identify open reading frames (ORFs) using GeneMark (24). Gene identifications were based on homology to known proteins using the blastp program (ftp://ftp.ncbi.nih.gov/blast) with an E-value cut-off of 10^{-5}. Further prerequisites were used for the identification of genes encoding ferredoxin and high light inducible proteins (HLIPs) due to their small size and sequence divergence. Ferredoxin encoding genes (*petF*) were included in our analyses if they encoded the 2Fe-2S iron-sulfur cluster binding domain (fer2) (with an E value $<10^{-10}$ as determined from the NCBI conserved domain database blast tool, rpsblast). HLIP encoding genes (*hli*) were identified as present for this study if they encoded at least 6 out of 10 of the amino acids in the motif AExxNGRxAMIGF (25). Bhaya et al. (26) report that many *Prochlorococcus hli* genes code for a conserved 9 amino acid C-terminal sequence, with the consensus sequence TGQIIPGI/FF. For this study, this sequence was defined as present when at least 6 out of 9 of the conserved amino acids were found at the C-terminus of the HLIP.

Rho-independent transcriptional terminators were identified using the TransTerm program (27), and all had an energy score of <-10 and a tail score of <-5. Potential bacterial sigma-70 promoters

were identified in intergenic regions using the program BPROM (www.softberry.com). Promoter sequences reported here had a linear discriminant function greater than 2.5. While identification of terminators is quite robust, detection of potential promoters in cyanophage is more precarious as the predictive ability of both cyanophage and cyanobacterial promoter elements is presently low.

Sequence manipulation and analyses

Sequences were initially aligned using Clustal X and edited manually when necessary. Amino acid alignments served as the basis for the manual alignment of nucleotide sequences. Regions that could not be confidently aligned were excluded from analyses, as were gaps. The divergence estimator program, K-estimator 6.0 (28), was used to estimate the frequency of synonymous and non-synonymous nucleotide substitutions and employs the Kimura 2p correction method for multiple-hits.

The PAUP V4.0b10 package was used for the construction of distance and maximum parsimony trees. Amino acid distance trees were inferred using minimum evolution as the objective function and mean distances. Heuristic searches were performed with 100 random addition sequence replicates and the tree-bisection and reconnection branch swapping algorithm. Starting trees were obtained by stepwise addition of sequences. Bootstrap analyses of 100 resamplings were carried out. Maximum likelihood trees were constructed using TREE-PUZZLE 5.0. Evolutionary distances were calculated using either the JTT model of substitution (for D1, D2 and ferredoxin) or the VT model of substitution (for the highly divergent HLIPs) assuming a gamma-distributed model of rate heterogeneities with 16 gamma-rate categories empirically estimated from the data. Quartet puzzling support was estimated from 10,000 replicates.

In cases where phylogenetic analyses of small genes received low bootstrap support we used the GeneRAGE clustering tool (29) which clusters protein sequences with significant relationships at user defined E-value thresholds. The input to GeneRage was an all-against-all table of blast comparisons of amino acid sequences made using blastp from NCBI. GeneRAGE uses a Smith-Waterman dynamic programming alignment algorithm to correct for false positive linkages whenever the pairwise relationships are not symmetrical. For HLIPs, an E-value cutoff of 10^{-14} was used. The clusters containing the phage HLIPs were preserved down to an E-value cutoff of

10^{-17}. For plastocyanin and ferredoxin respectively, E-value cutoffs of 10^{-26} and 10^{-34} linked the phage proteins with a large group of proteins, whereas at E-value cutoffs of 10^{-28} and 10^{-36} the respective phage proteins did not cluster with other sequences.

Results

A suite of host photosynthesis genes was found in the three phage genomes (Fig. 1). The *psbA* gene, encoding the PSII core reaction center protein, D1, and one *hli* gene type, encoding the HLIP cluster 14 type protein (*sensu* (26)) were present in all 3 phage. HLIPs are thought to be involved in protecting the photosynthetic apparatus from excess excitation energy during stressful conditions in cyanobacteria (30). Other photosynthesis related genes were present in the two myovirus genomes (Fig. 1B, C). Phage P-SSM4 contains the *psbD* gene encoding the second PSII core reaction center protein, D2, while phage P-SSM2 contains two photosynthetic electron transport genes coding for plastocyanin (*petE*) and ferredoxin (*petF*). Both phage contain additional gene types from the *hli* multigene family.

The deduced amino acid sequences of the phage photosynthesis genes are highly conserved and therefore have the potential to be functional proteins. The coding sequences of all of these genes are uninterrupted and show a high degree of identity to their host homologs (up to 85% and 95% nucleotide and amino acid identities respectively; Suppl. Table II, Suppl. Figs. 4-8). In the case of D1 and D2 from all three phage, the greatest amino acid divergence is in the N-terminal leader sequences that do not form part of the functional protein. Furthermore, divergence analyses based on estimates of the frequency of non-synonymous (Ka) and synonymous (Ks) nucleotide substitutions between phage- and host-encoded genes, revealed Ka/Ks ratios of less than 0.45 for all genes — with values of <0.1, similar to that among *Prochlorococcus* genes, for *psbA* and *psbD* (Suppl. Table III) — indicating that the majority of nucleotide substitutions did not cause a change in amino acid sequence. These findings suggest that the phage-encoded genes, particularly *psbA* and *psbD,* have been subjected to strong selective pressure to conserve their amino acid sequences, which is consistent with the hypothesis that they are functional.

All of the photosynthesis genes, with the exception of the plastocyanin gene, *petE*, are arranged together in the phage genomes, suggesting that they may be expressed at a similar stage of infection (3, 31). In addition, identification of potential promoter and terminator elements

suggests that distinct transcriptional units are found within these genomic regions. In the genome of P-SSP7, for example, *psbA* and the single *hli* gene may be co-transcribed with the adjacent phage structural genes in a single operon. Most of the genes in this region have overlapping start and stop codons and are flanked by a putative sigma-70 transcriptional promoter and rho-independent transcriptional terminator (Fig. 1A). This arrangement further suggests that the photosynthesis genes are expressed in the latter portion of the lytic cycle, as is known for structural proteins in other T7-like podoviruses (31). In contrast, the presence of transcriptional terminators flanking the regions containing photosynthetic genes in the myoviruses suggests that they may be transcribed as discrete transcriptional units largely independent of the surrounding phage genes.

The cyanobacterial origin of the phage *psbA* and *psbD* genes is suggested by the presence of certain features in both phage and host genes. Phage *psbA* genes code for a 7 amino acid indel close to the carboxy terminus of the D1 protein (Suppl. Fig. 4) which is almost exclusively found in cyanobacterial D1 proteins. Similarly, the phage *psbD* gene codes for a 7 amino acid indel in the center of the D2 protein that is also found in *Prochlorococcus* MED4 and SS120 (but not in other cyanobacteria or eukaryotic D2 proteins) (Suppl. Fig. 5). Interestingly, the *psbD* gene from neither *Synechococcus* WH8102 nor the *Synechococcus* phage S-PM2 codes for these additional amino acids (Suppl. Fig. 5). These findings suggest that *Prochlorococcus* phage acquired *psbD* from *Prochlorococcus* and *Synechococcus* phage acquired this gene from *Synechococcus.*

Phylogenetic analyses of the PSII core reaction center proteins further supports the cyanobacterial origin of the phage genes and, along with knowledge of phage host ranges (15), suggests that they were acquired multiple times from their hosts. The phage D1 and D2 proteins clustered with marine cyanobacteria in both distance (Fig. 2A & B) and maximum likelihood trees. Proteins encoded by phage that only infect *Prochlorococcus* clustered with *Prochlorococcus,* while those from a phage that infects only *Synechococcus* (sequences taken from (12)) clustered with *Synechococcus*, as did an environmental sequence (BAC9D04) encoding both D1 and phage structural genes (32). Moreover, D1 from two of the *Prochlorococcus* phage clustered within *Prochlorococcus* clades that match their host range (Fig. 2A). However, D1 from the third *Prochlorococcus* phage did not cluster within a specific *Prochlorococcus* clade suggesting that its *psbA* gene was either acquired from an as yet uncultured *Prochlorococcus* type, or has diverged

to an extent that prevents identification of the common ancestor. The fact that the phage D1 and D2 proteins are distributed in both the *Prochlorococcus* and *Synechococcus* clades, and are largely consistent with their host range, suggests that the genes were acquired in independent transfer events from their cyanobacterial hosts (*sensu* (2, 4)). This could have occurred *de novo* between distinct host and phage several times, or these genes may have been transferred from host to phage in a process akin to gene conversion subsequent to an ancestral transfer event (see Discussion section). If the host genes in phage resulted from a single ancestral event, followed by subsequent vertical or lateral transfers from phage to phage, the phage- and host-encoded genes would have formed monophyletic clades distinct from each other.

Phylogenetic analyses of the plastocyanin electron transport protein in host and phage also suggests that the phage *petE* gene is of cyanobacterial origin (Suppl. Fig. 9). However, the data is not conclusive as to the origin of the phage gene from within the cyanobacteria. The phage protein clusters with filamentous cyanobacteria, but contains a 10 amino acid indel found only in unicellular cyanobacteria (Suppl. Fig 6). GeneRAGE clustering analysis did not resolve the clustering of the phage plastocyanin protein. Both phylogenetic and GeneRAGE analyses of the ferredoxin electron transport protein encoded by the *petF* gene were inconclusive as to the origin of the phage gene. These results, together with the greater divergence estimates (Ka/Ks) for the phage- and *Prochlorococcus*-encoded *petE* and *petF* gene pairs (0.19-0.43) than among *Prochlorococcus* gene pairs (0.03-0.07) (Suppl. Table III), suggest that the photosynthetic electron transport genes encoded by P-SSM2 either originated from an organism for which a close relative does not currently exist in the database, or have diverged to an extent that prevents inference as to their origin. These may be new genes in the making.

Previous analyses of HLIPs in cyanobacterial genomes revealed the presence of genetically diverse types, with distinctly different clusters formed for single and multiple copy HLIPs (26). Genes found in a single copy in each of the 4 sequenced marine cyanobacterial genomes form 4 distinct clusters (GR C5, C6, C7 and C8 in Fig. 3) that are interspersed with HLIPs from freshwater cyanobacteria in a large cluster (Fig. 3), whereas multi-copy *Prochlorococcus* HLIPs are in a separate cluster (Fig. 3). While bootstrap support for these two broad clusters is not high, both distance and maximum likelihood analyses resulted in the same two broad groupings, lending some support to this tree architecture. When we add the phage HLIPs to this analysis

some interesting patterns appear. Ten out of eleven of the phage HLIPs cluster with those that are encoded by multiple gene copies in *Prochlorococcus* — some with more bootstrap support than others. That they do not group with those from freshwater cyanobacteria, nor with the HLIP types found in a single copy in marine unicellular cyanobacteria receives greater bootstrap support (Fig. 3). These results were obtained from three different analyses (distance and maximum likelihood phylogenetic analyses and GeneRAGE clustering). Indeed GeneRAGE clusters 7 out of 11 phage HLIPs with the four HLIP types encoded by multi-copy genes in *Prochlorococcus* genomes (GR 10, GR12, GR 14, and GR 15), with the remaining 4 of indeterminate affiliation using GeneRAGE. As for nearly all of the multi-copy HLIP sequences from *Prochlorococcus* (28 of 29), all but one of the phage HLIPs contain a 9 amino acid signature sequence at the carboxy terminus of the protein that is absent from other cyanobacterial HLIPs (26) — further supporting a connection between phage *hli* genes and multi-copy *hli* genes in the host.

While the lack of strong bootstrap support for most of the clustering patterns in Fig. 3 makes it impossible to draw definitive conclusions, the fact that both phage and *Prochlorococcus* HLIPs co-occur in four different clusters suggests that it is likely that *hli* genes have been transferred between host and phage multiple times. Moreover, the clustering of phage HLIPs with a subset of the HLIPs that are found exclusively in *Prochlorococcus* suggests that these distinct *hli* gene types may have been reacquired from phage after a period of evolution, leading to the expansion of the *hli* multigene family in this genus.

Discussion

Our findings, along with those in the companion paper (33) indicate that the presence of photosynthesis genes is widespread among phage that infect both *Prochlorococcus* and *Synechococcus*. Though they are not universal (13), they are found in representatives of both the *Myoviridae* and *Podoviridae*. The PSII core reaction center gene, *psbA,* has been found in all phage reported to have photosynthesis genes, suggesting that it plays a particularly significant role in these phage. Other photosynthesis genes were more sporadically distributed among the phage genomes that have been analyzed to date. Genes encoding HLIPs were found in all three *Prochlorococcus* phage we analyzed, but only one of five *Synechococcus* phage reported in Millard et al. (33). In contrast, the second PSII core reaction center gene, *psbD,* was found in all *Synechococcus* phage but only one *Prochlorococcus* phage. The small number of phage genomes

presently available for analysis precludes making robust conclusions from this asymmetry, but if the trend holds up as more genomes are sequenced, it is likely that there is a differential benefit of these two genes for the phage that is largely influenced by genera level attributes of their cyanobacterial host.

The photosynthetic electron transport genes were found in only one *Prochlorococcus* phage and no *Synechococcus* phage, whereas a transaldolase gene has been found in both *Prochlorococcus* myoviruses (MBS, FR, SWC unpubl data) and one *Synechococcus* phage (33). Until more phage genome sequences are available, it is not possible to determine the significance of what appears to be a sporadic distribution of these host genes among phage. However, assuming that the genes are functional, this scattered distribution may have arisen from differential gain and loss resulting from trade-offs between the burden of carrying such genes and their utility during infection. Alternatively, we may simply be observing the transient passage of host genes through the phage genome pool.

The arrangement of photosynthesis genes in both the *Prochlorococcus* and *Synechococcus* phage have some similar properties (compare Fig. 1 this study and (33)), yet are distinctly different from their organization in host genomes (20-22). In both phage we find adjacent *psbA* and *psbD* genes, adjacent *hli* and *psbA* genes, and *psbA* adjacent to a T4-like phage gene encoding gp49. Yet, phylogenetic analyses show that the proteins encoded by *psbA* and *psbD* from *Prochlorococcus* phage cluster with those from *Prochlorococcus*, and in at least the one *Synechococcus* phage available for analysis, these proteins cluster with those from *Synechococcus* (Fig. 2A, 2B). One likely explanation for these findings is that the genes were acquired from their respective hosts in separate transfer events, integrating at recombination hot-spots within the phage genome, and forming gene arrangements that may be advantageous. Alternatively, one early transfer event may have occurred, and the observed gene organization patterns formed prior to the divergence of these phage. In this latter case, for gene sequences to be similar to that from their respective hosts, they would have had to have been swapped between phage and host in a process similar to gene conversion, whereby one gene is replaced by another in a non-reciprocal fashion. The direction of this gene conversion is most likely with the host gene replacing the phage gene, as cyanobacterial phylogenies inferred from *psbA* and *psbD* gene products are congruent with those from other

genes (Fig. 2A, 2B, (16-19)). This latter scenario would suggest that encoding PSII reaction center genes similar to that from the host is advantageous.

The presence of highly conserved PSII reaction center and *hli* genes in the 3 *Prochlorococcus* phage suggests that strong evolutionary pressure has driven their acquisition and retention. This is liable to have important implications for phage-host interactions during infection. It has been known for some time that viral infection of many photosynthetic organisms leads to a decline in photosynthetic rates soon after infection (34, 35). This is attributed to damage to the PSII membrane-protein complexes (36, 37) and may be due to oxidative stress caused by an increase in destructive reactive oxygen species subsequent to infection (37). However, in many phage-infected unicellular cyanobacteria, the production of phage progeny is dependent upon photosynthetic activity continuing until just prior to lysis (38, 39). Phage PSII reaction center proteins may, if expressed, prevent photoinhibitory damage to PSII in *Synechococcus* (12). We further suggest that expression of phage PSII reaction center proteins, as well as the photoprotective HLIPs may help maintain photosynthetic activity during infection of *Prochlorococcus,* leading to increased phage fitness and resulting in selection for cyanophage that encode functional photosynthetic genes.

Our analysis of host genes in phage have implications not only for phage fitness, but also for the evolution of the hosts, as there is suggestive evidence that phage may have mediated horizontal gene transfer and hence expansion of the *hli* multigene family in the hosts. It has recently been suggested that widely distributed, single copy genes are resistant to horizontal transfer (40), while sporadically distributed multicopy genes are those most likely to have been dispersed by this method (40, 41). The clustering patterns displayed by the *hli* genes in our analyses, though not statistically robust, are consistent with this tenant. Each of the single copy *hli* gene types that are common to the four sequenced unicellular marine cyanobacteria (20-22), are likely to have been vertically transferred as is evident from the conserved gene arrangement surrounding these *hli* types (26) and from their clustering to those from the other marine unicellular cyanobacteria ((26), Fig. 3). In contrast, the *hli* gene types that are present in multiple copies per genome are found in only some *Prochlorococcus* genomes. It is these latter *hli* gene types that are found in the *Prochlorococcus* phage genomes, with at least one phage *hli* gene in each of the 4 clusters of multi-copy *Prochlorococcus hli* gene types (Fig. 3). We therefore suggest that phage may have

mediated the horizontal dispersal of these multi-copy genes among *Prochlorococcus*. Interestingly, none of the 3 *Prochlorococcus* phage carrying these *hli* genes infect marine *Synechococcus* (15), suggesting that if these genes were laterally transferred, the breadth of that transfer matches the limits of their host infection capabilities.

The presence of numerous *hli* genes in *Prochlorococcus* MED4, a high-light adapted ecotype, is likely to have facilitated its inhabitance of the surface waters of the open oceans (20, 26, 42). Indeed cyanobacteria with multiple *hli* genes have been shown to have a competitive advantage upon shifts to high light over mutants in which some of these gene copies have been inactivated (30). Our hypothesized phage-mediated expansion of the *hli* multigene family may have facilitated the specialization of *Prochlorococcus* to high-irradiance surface ocean waters leading to their dominance over other photosynthetic organisms in these environments. Other photosynthetic genes found in phage are also present in multiple copies in many cyanobacteria, including *psbA*, *psbD* and *petF* (Suppl. Table IV). The importance of gene duplication in the evolution of new gene functions is well recognized in other systems (43, 44) and thus it would not be surprising if it were playing a role in the evolution of physiological variants within the *Prochlorococcus* cluster.

The exchange of photosynthetic genes between *Prochlorococcus* and the phage that infect them could have significant implications for the evolutionary trajectory of both host and phage, and may represent a more general phenomenon of metabolic facilitation of key host processes. That is, the selection of host genes that are retained in a particular phage could reflect key selective forces in the host environment. Indeed, phosphate sensing and acquisition genes have been found in phage that infect organisms in low phosphate environments (8, 9). Might we also find salt tolerance genes in phage infecting halotolerant organisms, and thermal tolerance genes in phage that infect thermophilic organisms? Such coupled evolutionary processes in host and phage, if widespread, may play a role in defining host ranges for phage and niche space for hosts, leading to specialization and even speciation.

Acknowledgements

The sequencing and assembly of the phage genomes was performed by the production sequencing group at the DOE/Joint Genome Institute through the Sequence-For-Others Program under the

auspices of the USA Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36, and Stanford University under contract No. DE-FC02-99ER62873. This research was supported by the USA Department of Energy under grant Nos. DE-FG02-99ER62814 and DE-FG02-02ER63445, and the National Science Foundation OCE-9820035 (to SWC). We would like to thank D. Veneziano, G. Rocap, D. Mead, M. Ermolaeva and F. Chagnon for assistance and discussions with various aspects of this work.

References

1. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. (2000) *Trends Microbiol.* **8,** 504-508.
2. Filee, J., Forterre, P. & Laurent, J. (2003) *Res. Microbiol.* **154,** 237-43.
3. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Ruger, W. (2003) *Microb. Mol. Biol. Rev.* **67,** 86-156.
4. Moreira, D. (2000) *Mol. Microbiol.* **35,** 1-5.
5. Wagner, P. L. & Waldor, M. K. (2002) *Infect. Immun.* **70,** 3985-3993.
6. Casjens, S. (2003) *Mol. Microbiol.* **49,** 277-300.
7. Forterre, P. (1999) *Mol. Microbiol.* **33,** 457-65.
8. Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F. & Azam, F. (2000) *Limnol. Oceanogr.* **45,** 408-418.
9. Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., *et al.* (2003) *J. Bacteriol.* **185,** 5220-33.
10. Figueroa-Bossi, N. & Bossi, L. (1999) *Mol. Microbiol.* **33,** 167-76.
11. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., *et al.* (2003) *Cell* **113,** 171-82.
12. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. (2003) *Nature* **424,** 741.
13. Chen, F. & Lu, J. (2002) *Appl. Environ. Microbiol.* **68,** 2589-94.
14. Waterbury, J. B. & Valois, F. W. (1993) *Appl. Environ.l Microbiol.* **59,** 3393-3399.
15. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. (2003) *Nature* **424,** 1047-1051.

16. Ferris, M. J. & Palenik, B. (1998) *Nature* **396,** 226-228.

17. Moore, L. R., Rocap, G. & Chisholm, S. W. (1998) *Nature* **393,** 464-467.

18. Rocap, G., Distel, D., Waterbury, J. B. & Chisholm, S. W. (2002) *Appl. Environ. Microbiol.* **68,** 1180-1191.

19. Zeidner, G., Preston, C. M., Delong, E. F., Massana, R., Post, A. F., Scanlan, D. J. & Beja, O. (2003) *Environ. Microbiol.* **5,** 212-6.

20. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) *Nature* **424,** 1042-1047.

21. Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E. A., Partensky, F. & Larimer, F. W. (2003) *Nature* **424,** 1037-1041.

22. Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10020-5.

23. Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14250-5.

24. Besemer, J., Lomsadze, A. & Borodovsky, M. (2001) *Nucl. Acids Res.* **29,** 2607-18.

25. Jansson, S., Andersson, J., Kim, S. J. & Jackowski, G. (2000) *Plant Mol. Biol.* **42,** 345-51.

26. Bhaya, D., Dufresne, A., Vaulot, D. & Grossman, A. (2002) *FEMS Microbiol. Lett.* **215,** 209-19.

27. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. (2000) *J. Mol. Biol.* **301,** 27-33.

28. Comeron, J. M. (1999) *Bioinformatics* **15,** 763-764.

29. Enright, A. J. & Ouzounis, C. A. (2000) *Bioinformatics* **16,** 451-7.

30. He, Q., Dolganov, N., Bjorkman, O. & Grossman, A. R. (2001) *J. Biol. Chem.* **276,** 306-14.

31. Dunn, J. J. & Studier, F. W. (1983) *J. Mol. Biol.* **166,** 477-535.

32. Zeidner, G. & Beja, O. (2004) *Environ. Microbiol.* **in press**.

33. Millard, A., Clokie, M., Shub, D. A. & Mann, N. H. (2004) *submitted.*

34. Padan, E. & Shilo, M. (1973) *Bacteriol. Rev.* **37,** 343-370.

35. Zaitlin, M. & Hull, R. (1987) *Ann. Rev. Plant Physiol.* **38,** 291-315.

36. Rahoutei, J., Garcia-Luque, I. & Baron, M. (2000) *Physiologia Plantarum* **110,** 286-292.

37. Arias, M. C., Lenardon, S. & Taleisnik, E. (2003) *J. Phytopathology* **151,** 267-273.

38. Mackenzie, J. J. & Haselkorn, R. (1972) *Virology* **49,** 517-521.

39. Sherman, L. A. (1976) *Virology* **71,** 199-206.

40. Lerat, E., Daubin, V. & Moran, N. A. (2003) *PLoS Biol.* **1,** E19.

41. Kurland, C. G., Canback, B. & Berg, O. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9658-62.

42. Hess, W. R., Rocap, G., Ting, C. S., Larimer, F. W., Stilwagen, S., Lamerdin, J. & Chisholm, S. W. (2001) *Photosyn. Res.* **70,** 53-71.

43. Zhang, J. (2003) *Trends Ecol. Evol.* **18,** 292-298.

44. Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W. (2003) *Nature* **421,** 63-66.

Figure legends

Fig. 1. Arrangement of photosynthesis genes in three *Prochlorococcus* phage genomes. (A) Podovirus P-SSP7, (B) Myovirus P-SSM2, (C) Myovirus P-SSM4. Solid bars indicate genes related to photosynthesis, hatched bars indicate genes commonly found in phage, and open bars indicate predicted open reading frames of unknown function. Gene=protein designations: *psbA*=D1, *psbD*=D2, *hli*=HLIP, *petE*=plastocyanin, *petF*=ferredoxin, *8*=T7-like head-to-tail connector, *9*=T7-like portal protein, *10*=T7-like capsid protein, *nrdB*=T4-like ribonucleotide reductase beta subunit, *49*=T4-like restriction endonuclease VII, *td*=T4-like thymidylate synthetase.

Fig. 2. Distance trees of PSII core reaction center proteins (A) D1 (*psbA*), and (B) D2 (*psbD*). Phage sequences are shown in bold. The host strain(s) that each phage infects are indicated with black squares. Trees were generated from partial amino acid sequences with 244 and 336 amino acids included in the analyses for D1 and D2 respectively (see Suppl. Figs. 4 and 5). Bootstrap values for distance and maximum likelihood analyses, and quartet puzzling values for maximum likelihood analysis, greater than 50%, are shown to the left of the nodes (distance/maximum likelihood/maximum parsimony). Trees were rooted with genes from *Arabidopsis thaliana*. Essentially the same topology was obtained when nucleotide trees (with 3rd position excluded) were constructed except for *psbA* from P-SSP7, which clustered with HL *Prochlorococcus* in nucleotide trees, albeit with low bootstrap support. *Pro* = *Prochlorococcus, Syn* = *Synechococcus, Anab* = *Anabaena, Syncy* = *Synechocystis*.

Fig. 3. Distance tree of 80 HLIPs from cyanobacteria and phage. Phage HLIPs appear in bold. The tree was generated from partial amino acid sequences (the 36 amino acid region included in analyses is indicated in Suppl. Fig. 8) and gaps were treated as missing data. GeneRAGE clusters are indicated to the right of the tree (GR C#), with cluster designations following (26). Discrepancies between GeneRAGE and distance tree clustering were found for 3 HLIPs and are indicated by the dashed line and their GR cluster designations. Asterisks denote proteins encoding at least 6 out of 9 of the C-terminal 9 amino acid consensus sequence. Bootstrap and quartet puzzling values greater than 50% are shown to the left of the nodes for distance and maximum likelihood analyses respectively. The tree was rooted with the single HLIP from *Arabidopsis thaliana*. Abbreviations as for Fig. 2.

Table I: Phage used in this study and their photosynthesis-related genes. Phage family and host-range information as per (15). Bold indicates the host the phage was isolated on. *From Mann et al. (12).

Phage	**Family**	**Host Strains Infected**	**Genes**
P-SSP7	Podovirus	*Pro* **MED4** (HL)	*psbA*, 1 *hli*
P-SSM2	Myovirus	*Pro* **NATL1A**, NATL2A, MIT9211 (LL)	*psbA*, 6 *hli* genes, *petF, petE*
P-SSM4	Myovirus	*Pro* NATL1A, **NATL2A** (LL) *Pro* MED4, MIT9215 (HL)	*psbA*, *psbD*, 4 *hli* genes
S-PM2*	Myovirus	*Syn* **WH7803**, WH8109	*psbA*, *psbD*

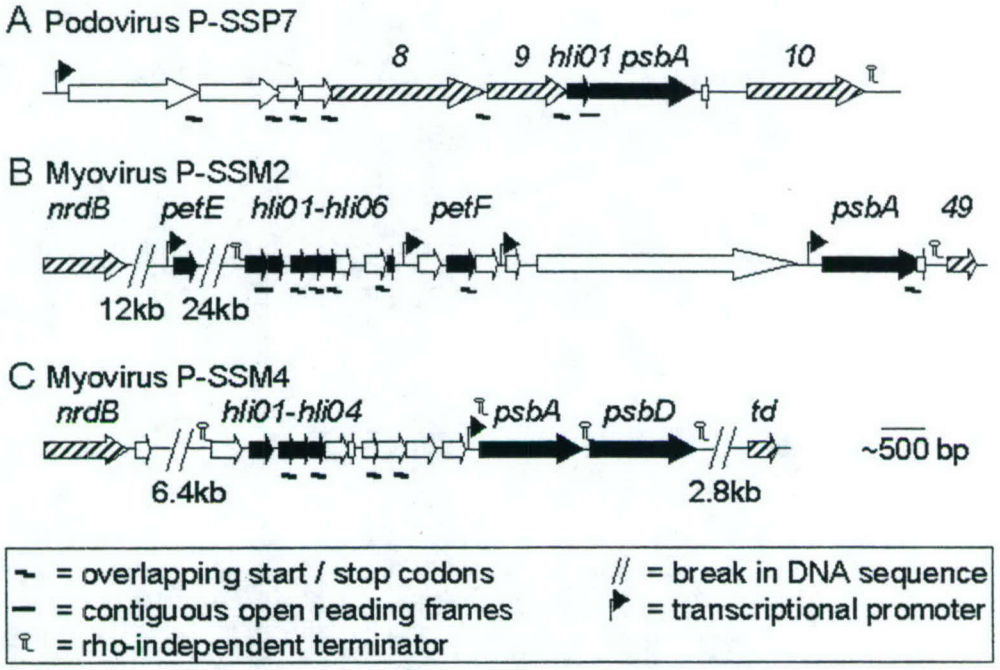

A Podovirus P-SSP7
8
9 hli01 psbA
10
B Myovirus P-SSM2
nrdB
petE
hli01-hli06
petF
psbA
49
12kb 24kb
C Myovirus P-SSM4
nrdB
hli01-hli04
psbA
psbD
td
~500 bp
6.4kb
2.8kb
~ = overlapping start / stop codons
— = contiguous open reading frames
= rho-independent terminator
// = break in DNA sequence
= transcriptional promoter

Fig. 1

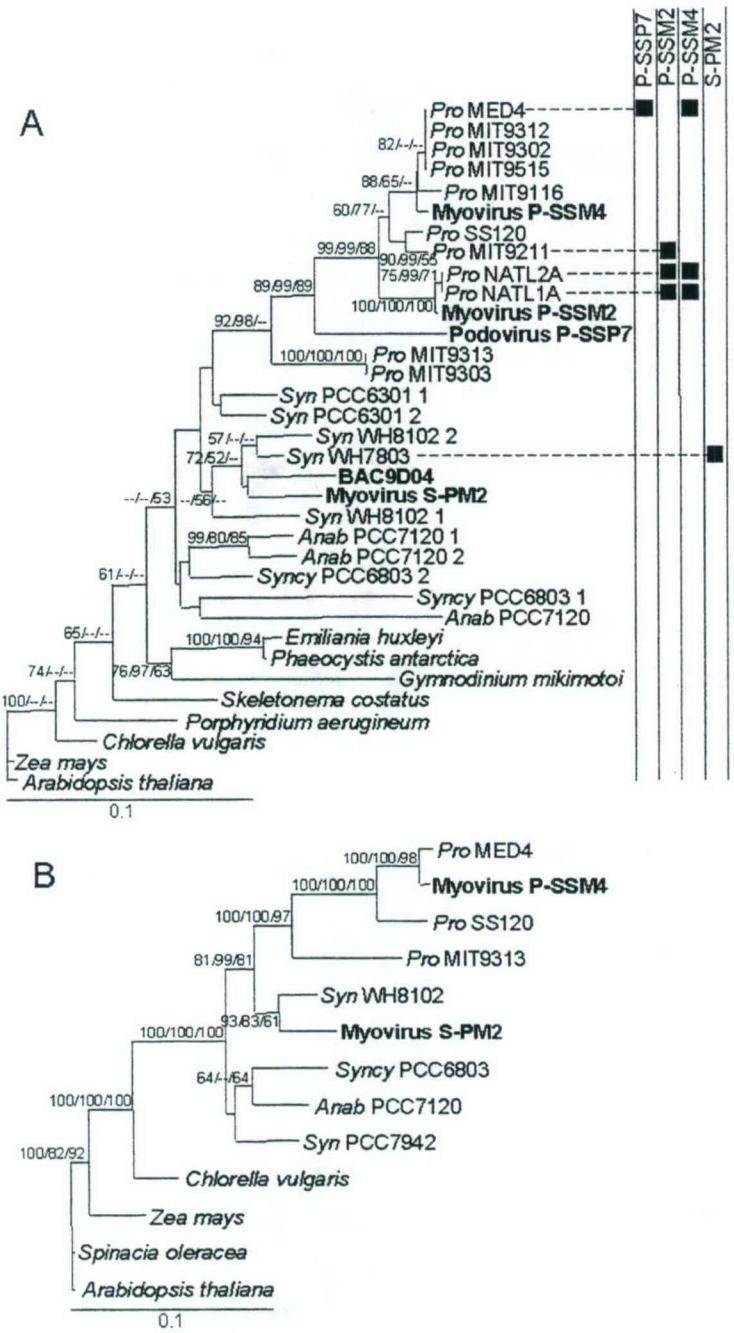

A
P-SSP7
P-SSM2
P-SSM4
S-PM2
Pro MED4
Pro MIT9312
Pro MIT9302
Pro MIT9515
Pro MIT9116
Myovirus P-SSM4
Pro SS120
Pro MIT9211
Pro NATL2A
Pro NATL1A
Myovirus P-SSM2
Podovirus P-SSP7
Pro MIT9313
Pro MIT9303
Syn PCC6301 1
Syn PCC6301 2
Syn WH8102 2
Syn WH7803
BAC9D04
Myovirus S-PM2
Syn WH8102 1
Anab PCC7120 1
Anab PCC7120 2
Syncy PCC6803 2
Syncy PCC6803 1
Anab PCC7120
Emiliania huxleyi
Phaeocystis antarctica
Gymnodinium mikimotoi
Skeletonema costatus
Porphyridium aerugineum
Chlorella vulgaris
Zea mays
Arabidopsis thaliana
0.1
B
Pro MED4
Myovirus P-SSM4
Pro SS120
Pro MIT9313
Syn WH8102
Myovirus S-PM2
Syncy PCC6803
Anab PCC7120
Syn PCC7942
Chlorella vulgaris
Zea mays
Spinacia oleracea
Arabidopsis thaliana
0.1

Fig. 2

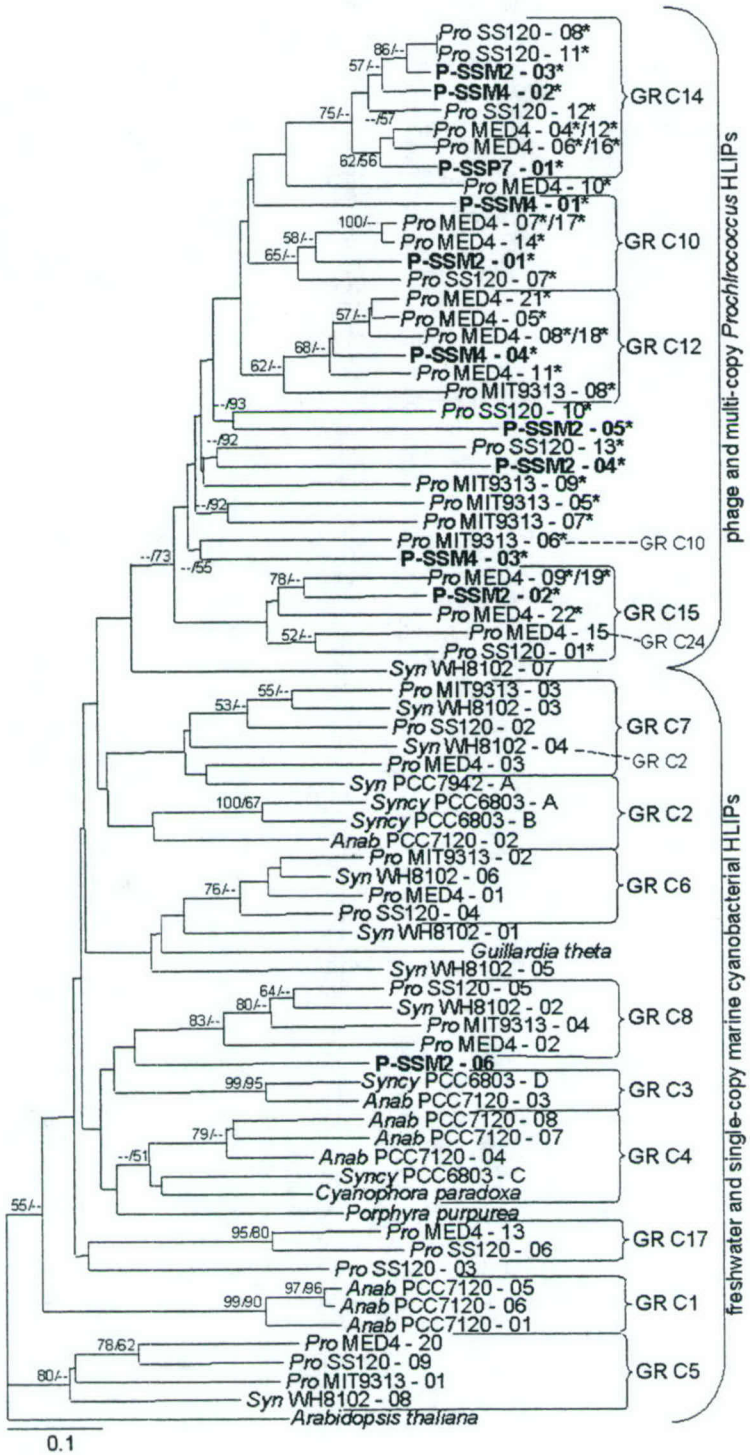

Pro SS120 - 08*
Pro SS120 - 11*
P-SSM2 - 03*
P-SSM4 - 02*
Pro SS120 - 12*
Pro MED4 - 04*/12*
Pro MED4 - 06*/16*
P-SSP7 - 01*
Pro MED4 - 10*
P-SSM4 - 01*
Pro MED4 - 07*/17*
Pro MED4 - 14*
P-SSM2 - 01*
Pro SS120 - 07*
Pro MED4 - 21*
Pro MED4 - 05*
Pro MED4 - 08*/18*
P-SSM4 - 04*
Pro MED4 - 11*
Pro MIT9313 - 08*
Pro SS120 - 10*
P-SSM2 - 05*
Pro SS120 - 13*
P-SSM2 - 04*
Pro MIT9313 - 09*
Pro MIT9313 - 05*
Pro MIT9313 - 07*
Pro MIT9313 - 06*
P-SSM4 - 03*
Pro MED4 - 09*/19*
P-SSM2 - 02*
Pro MED4 - 22*
Pro MED4 - 15
Pro SS120 - 01*
Syn WH8102 - 07
Pro MIT9313 - 03
Syn WH8102 - 03
Pro SS120 - 02
Syn WH8102 - 04
Pro MED4 - 03
Syn PCC7942 - A
Syncy PCC6803 - A
Syncy PCC6803 - B
Anab PCC7120 - 02
Pro MIT9313 - 02
Syn WH8102 - 06
Pro MED4 - 01
Pro SS120 - 04
Syn WH8102 - 01
Guillardia theta
Syn WH8102 - 05
Pro SS120 - 05
Syn WH8102 - 02
Pro MIT9313 - 04
Pro MED4 - 02
P-SSM2 - 06
Syncy PCC6803 - D
Anab PCC7120 - 03
Anab PCC7120 - 08
Anab PCC7120 - 07
Anab PCC7120 - 04
Syncy PCC6803 - C
Cyanophora paradoxa
Porphyra purpurea
Pro MED4 - 13
Pro SS120 - 06
Pro SS120 - 03
Anab PCC7120 - 05
Anab PCC7120 - 06
Anab PCC7120 - 01
Pro MED4 - 20
Pro SS120 - 09
Pro MIT9313 - 01
Syn WH8102 - 08
Arabidopsis thaliana
0.1
GR C14
GR C10
GR C12
GR C15
GR C24
GR C10
GR C7
GR C2
GR C2
GR C6
GR C8
GR C3
GR C4
GR C17
GR C1
GR C5
phage and multi-copy Prochlrococcus HLIPs
freshwater and single-copy marine cyanobacterial HLIPs

Fig. 3

Suppl. Table II: Range of nucleotide and amino acid percent identities for pairs of phage and *Prochlorococus* genes and among *Prochlorococcus* genes. *Prochlorococcus* strains MED4, MIT9313 and SS120 were used for these analyses with only one copy of identical genes within a single genome being included.

	Among *Prochlorococcus* genes		**Phage to *Prochlorococcus* genes**	
Gene	**nucleotide**	**amino acid**	**nucleotide**	**amino acid**
psbA	74 – 81 %	87 -96 %	69 – 83 %	85 – 93 %
psbD	71 – 80 %	85 – 93 %	70 – 78 %	83 – 95 %
petE	69 – 71 %	74 – 76 %	47 – 54 %	41 – 47 %
petF	73 – 78 %	91 – 94 %	48 – 57 %	51 – 53 %
***hli* cluster 10**	58 – 86 %	50 – 91 %	51 – 63 %	43 – 64 %
***hli* cluster 12**	53 – 73 %	56 – 75 %	58 – 70 %	63 – 71 %
***hli* cluster 14**	71 – 94 %	77 – 91 %	73 – 85 %	74 – 94 %
***hli* cluster 15**	64 – 70 %	59 – 66 %	56 – 66 %	54 – 64 %

Suppl. Table III: Divergence estimates for *Prochlorococcus* gene pairs (using strains MED4, MIT9313 and SS120) and phage- and *Prochlorococcus*-encodedgenes. Ks = the number of nucleotide substitutions that do not cause an amino acid change per synonymous site, and Ka = the number of nucleotide substitutions that cause an amino acid change per non-synonymous site. For the *petF* and *hli* genes, where multiple gene types are found, estimates are for genes clustered in the same group as determined from GeneRAGE analyses. Note that only one identical *hli* gene from the same genome was included in the analyses. Ranges are provided for gene pairs for which the Kimura 2p correction was applicable.

Gene	**Protein**	**Among *Prochlorococcus***		**Phage to *Prochlorococcus***	
		Ks	**Ka/Ks**	**Ks**	**Ka/Ks**
psbA	**D1**	1.47 – 2.14	0.03 – 0.05	0.93 – 3.11	0.03 – 0.07
psbD	**D2**	1.30 – 1.95	0.04 – 0.05	2.24	0.02
petE	**plastocyanin**	1.37 – 1.85	0.06 – 0.07	1.13 – 2.42	0.20 – 0.40
petF	**ferredoxin**	1.88	0.03	1.08 – 2.65	0.19 – 0.43
***hli* c10**	**HLIP C10**	0.59 – 1.59	0.09 – 0.31	0.99 – 1.80	0.14 – 0.39
***hli* c14**	**HLIP C14**	0.31 – 1.96	0.00 – 0.18	0.65 – 1.59	0.05 – 0.24
***hli* c15**	**HLIP C15**	0.82	0.21	1.64	0.13

Suppl. Table IV: Number of copies of the relevant photosynthesis genes found in cyanobacteria and *Prochlorococcus* phage.

	psbA	***psbD***	***petE***	***petF***	***hli***
***Prochlorococcus* MED4**	1	1	1	3	22
***Prochlorococcus* MIT9313**	2	1	1	2	9
***Prochlorococcus* SS120**	1	1	1	2	13
***Synechococcus* WH8102**	4	2	1	4	8
***Synechocystis* PCC6803**	3	2	1	4	4
***Anabaena* PCC7120**	5	2	1	3	8
Podovirus P-SSP7	1	0	0	0	1
Myovirus P-SSM2	1	0	1	1	6
Myovirus P-SSM4	1	1	0	0	4

Supplementary Figure Legends

Suppl. Fig. 4. Alignment of D1 amino acid sequences deduced from phage and cellular encoded *psbA* genes. Note the additional 7 amino acids towards the C-terminus of the protein from cyanobacterial and phage D1 proteins. The amino acids used in phylogenetic analyses correspond to the region extending from position 84 (VPSS) to position 327 (RANL) of the protein from *Prochlorococcus* MED4. *Pro = Prochlorococcus, Syn = Synechococcus, Ana = Anabaena, Syncy = Synechocystis.*

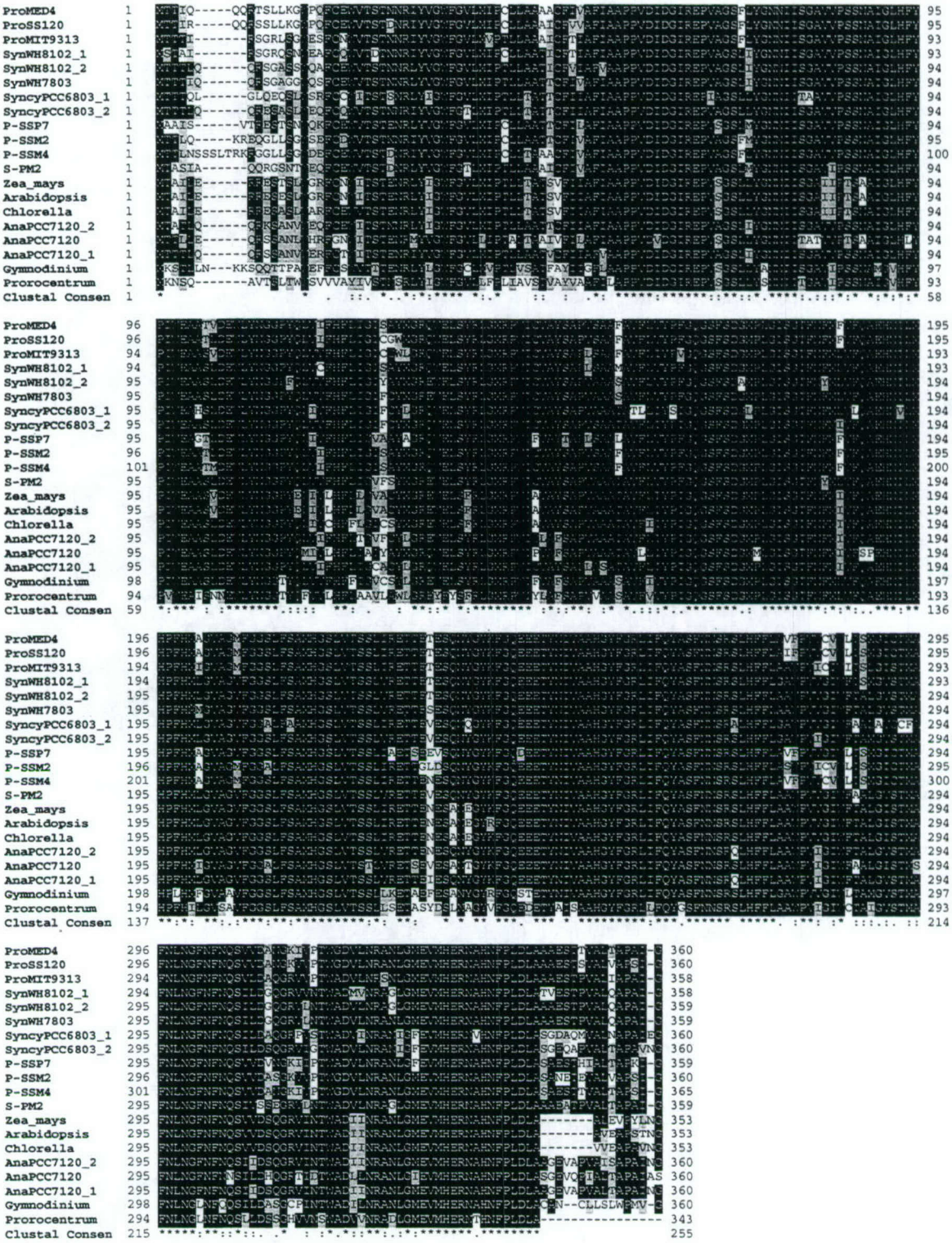
Suppl. Fig. 5. Alignment of D2 amino acid sequences deduced from phage and cellular encoded *psbD* gene sequences. Note the 7 amino acid indel in the D2 protein from *Prochlorococcus* MED4 and SS120 and *Prochlorococcus* phage P-SSM4. These additional 7 amino acids are not found in the D2 proteins from other cyanobacteria nor the *Synechococcus* phage S-PM2. The amino acids used in phylogenetic analyses correspond to the region extending from position 14 (FDVL) to position 358 (GNAL) of the protein from *Prochlorococcus* MED4. *Pro = Prochlorococcus, Syn = Synechococcus, Ana = Anabaena, Syncy = Synechocystis.*

Suppl. Fig. 6. Alignment of plastocyanin amino acid sequences deduced from phage and cellular encoded *petE* gene sequences. Note the 10 amino acid indel in the plastocyanin protein from unicellular cyanobacteria and *Prochlorococcus* phage P-SSM2. The amino acids used in phylogenetic analyses correspond to the region extending from position 36 (GMLA) to position 116 (VIVE) of the protein from *Prochlorococcus* MED4. *Pro = Prochlorococcus, Syn = Synechococcus, Ana = Anabaena, Syncy = Synechocystis, Tricho = Trichodesmium.*

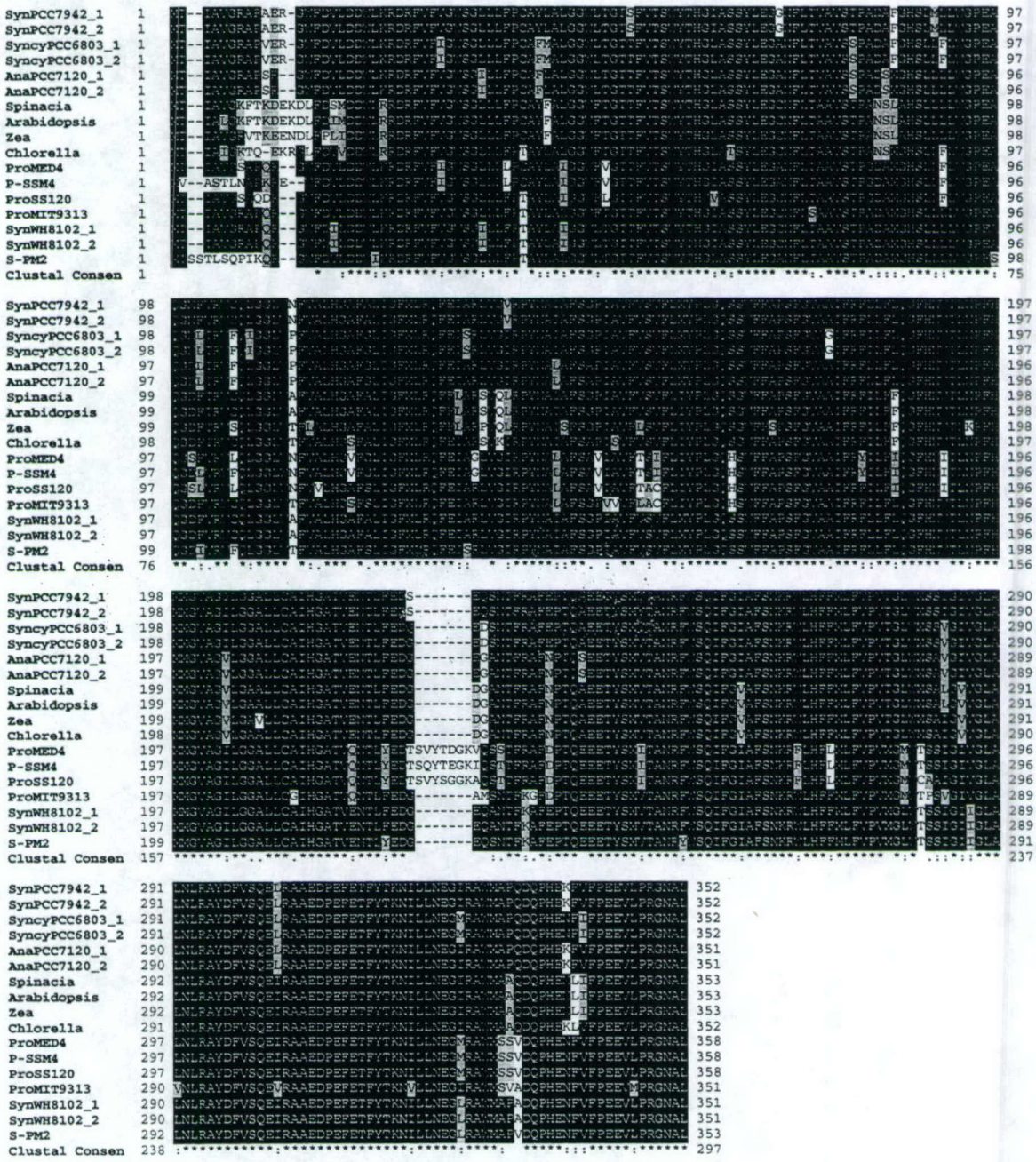
Suppl. Fig. 7. Alignment of ferredoxin amino acid sequences deduced from phage and cellular *petF* gene sequences. *Pro = Prochlorococcus, Syn = Synechococcus, Ana = Anabaena, Syncy = Synechocystis.*

Suppl. Fig. 8. Alignment of HLIP amino acid sequences deduced from cyanophage and cellular encoded *hli* genes. The motifs used for defining HLIPs (AExxNGRxAMIGF), as well as the region of the C-terminal consensus sequence (TGQIIPGI/FF) found in 28 out of 44 *Prochlorococcus* HLIPs and 10 out of 11 *Prochlorococcus* phage HLIPs, are indicated below the sequences. The *hli* gene number designation is shown after the strain ID. The amino acids used in phylogenetic analysis extend from 4 amino acids towards the N-terminus of the AExxNGRxAMIGF motif to the last amino acid of the C-terminal motif. *Pro = Prochlorococcus, Syn = Synechococcus, Ana = Anabaena, Syncy = Synechocystis.*

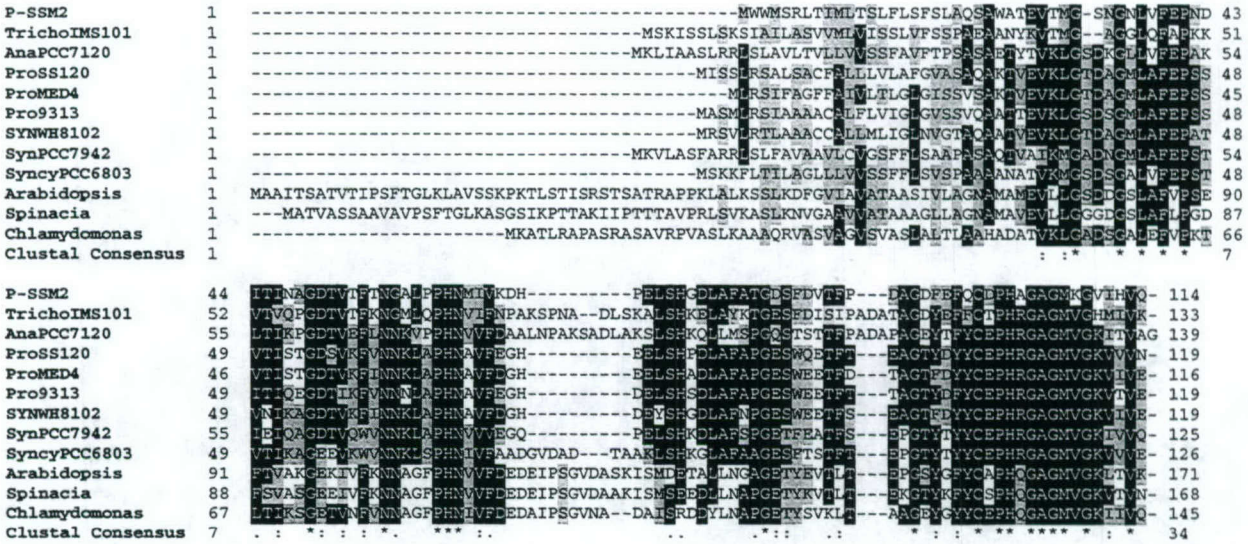
Suppl. Fig. 9. Distance tree of the plastocyanin photosynthetic electron transport protein encoded by the *petE* gene. The phage protein is shown in bold. The tree was generated from partial amino acid sequences. The 80 aa region included in analyses is shown in Suppl Fig. 6. Bootstrap values (for distance and maximum parsimony analyses) and quartet puzzling support (for maximum likelihood analysis) greater than 50% are shown to the left of the nodes (distance/maximum likelihood/maximum parsimony). The tree was rooted with the gene from *Arabidopsis thaliana.* Cyanobacterial genus abbreviations: *Pro = Prochlorococcus, Syn = Synechococcus, Anab = Anabaena, Syncy = Synechocystis, Tricho = Trichodesmium.*


ProMED4
ProSS120
ProMIT9313
SynWH8102_1
SynWH8102_2
SynWH7803
SyncyPCC6803_1
SyncyPCC6803_2
P-SSP7
P-SSM2
P-SSM4
S-PM2
Zea_mays
Arabidopsis
Chlorella
AnaPCC7120_2
AnaPCC7120
AnaPCC7120_1
Gymnodinium
Prorocentrum
Clustal Consen

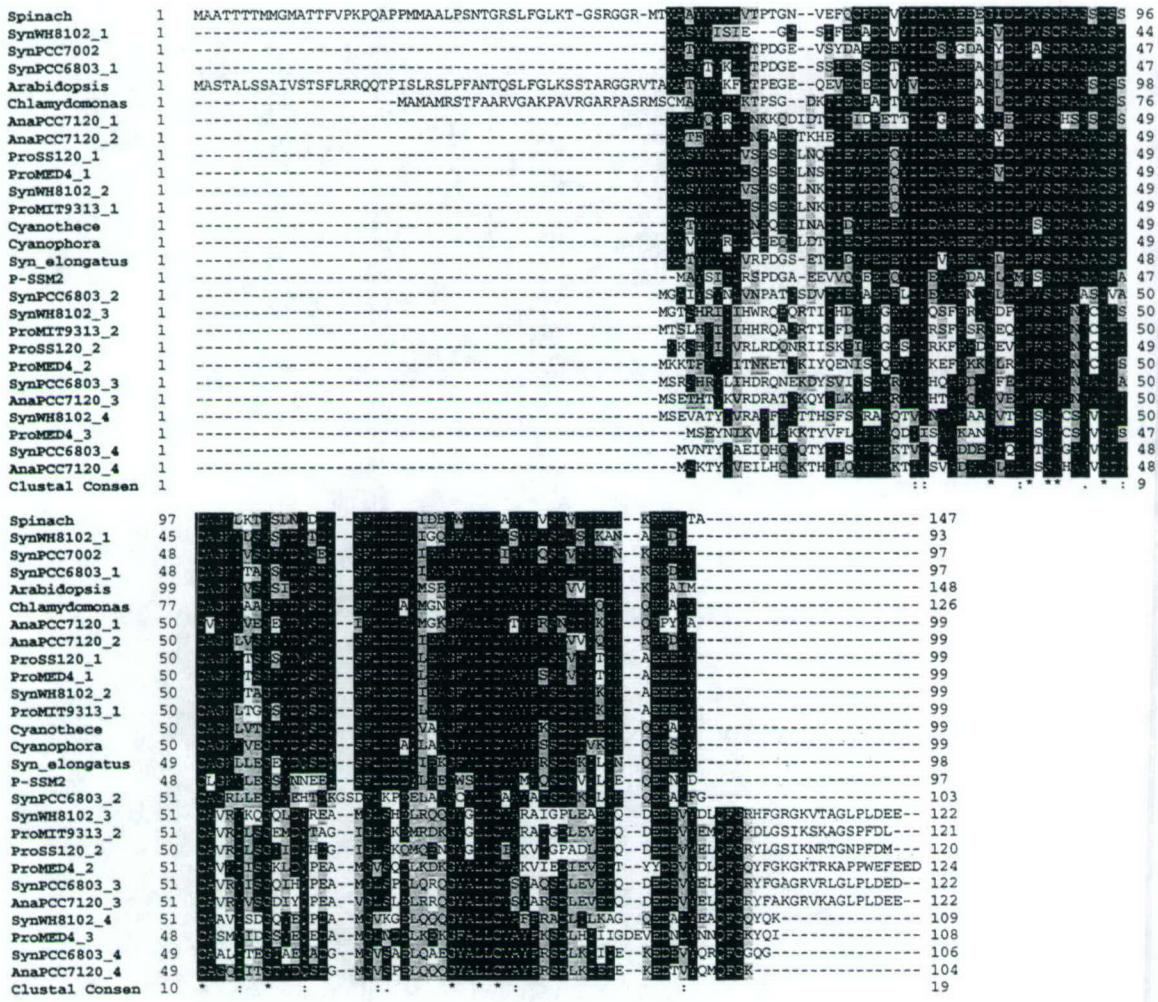
Suppl. Fig. 4


SynPCC7942_1
SynPCC7942_2
SyncyPCC6803_1
SyncyPCC6803_2
AnaPCC7120_1
AnaPCC7120_2
Spinacia
Arabidopsis
Zea
Chlorella
ProMED4
P-SSM4
ProSS120
ProMIT9313
SynWH8102_1
SynWH8102_2
S-PM2
Clustal Consen

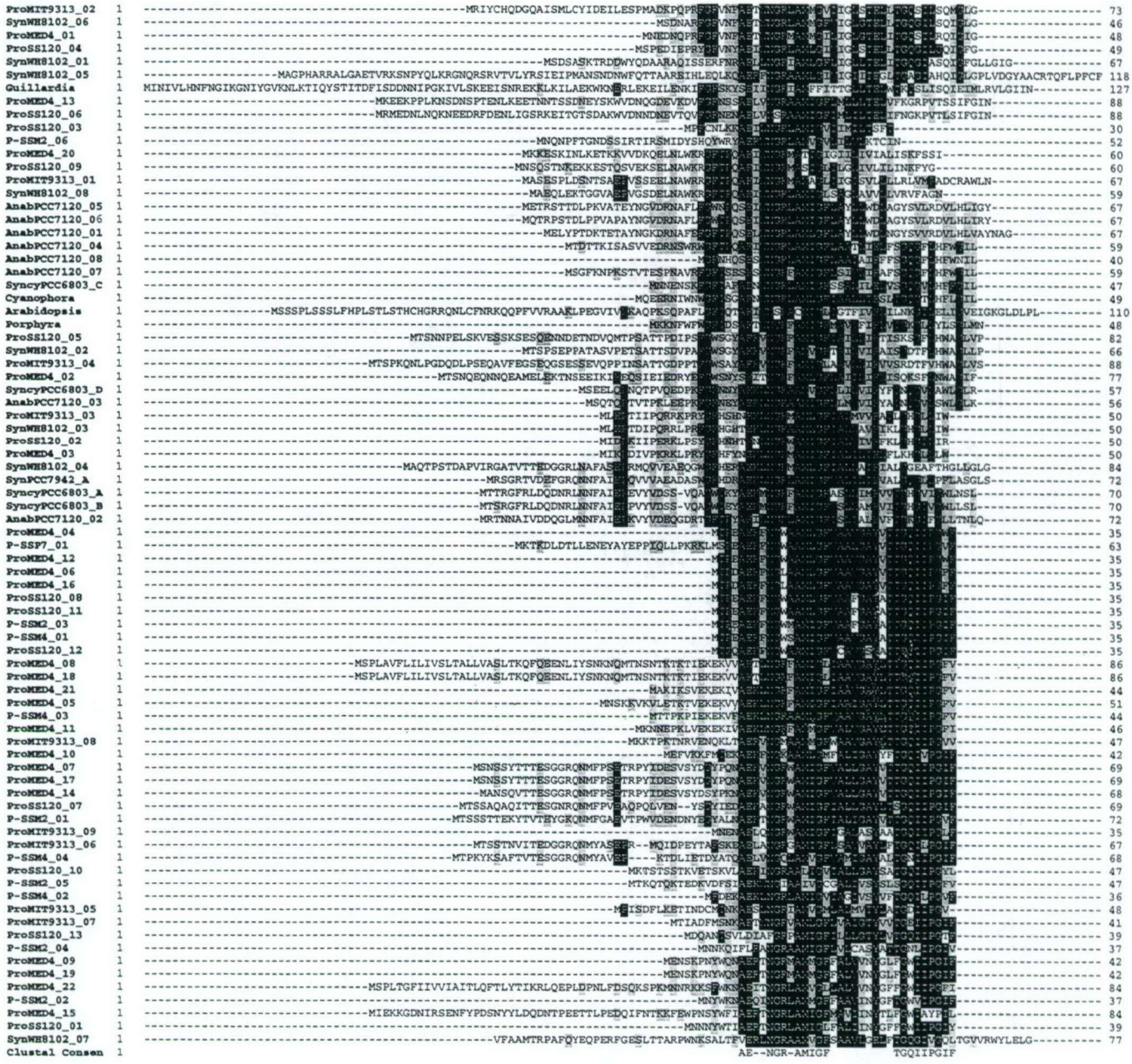
Suppl Fig. 5


P-SSM2
TrichoIMS101
AnaPCC7120
ProSS120
ProMED4
Pro9313
SYNWH8102
SynPCC7942
SyncyPCC6803
Arabidopsis
Spinacia
Chlamydomonas
Clustal Consensus

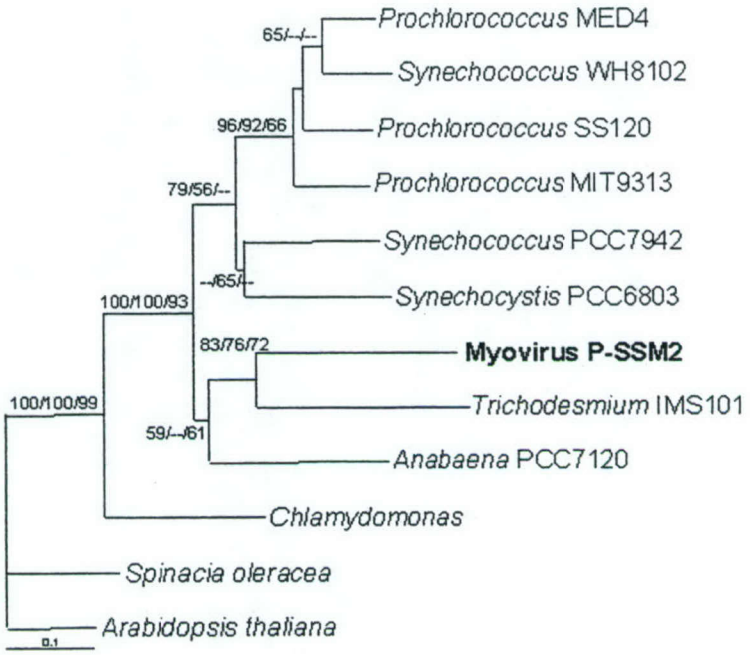
Suppl Fig. 6


Spinach
SynWH8102_1
SynPCC7002
SynPCC6803_1
Arabidopsis
Chlamydomonas
AnaPCC7120_1
AnaPCC7120_2
ProSS120_1
ProMED4_1
SynWH8102_2
ProMIT9313_1
Cyanothece
Cyanophora
Syn_elongatus
P-SSM2
SynPCC6803_2
SynWH8102_3
ProMIT9313_2
ProSS120_2
ProMED4_2
SynPCC6803_3
AnaPCC7120_3
SynWH8102_4
ProMED4_3
SynPCC6803_4
AnaPCC7120_4
Clustal Consen

Suppl. Fig. 7



Suppl. Fig. 8


Prochlorococcus MED4
65/--/--
Synechococcus WH8102
96/92/66
Prochlorococcus SS120
79/56/--
Prochlorococcus MIT9313
Synechococcus PCC7942
--/65/--
Synechocystis PCC6803
100/100/93
83/76/72
Myovirus P-SSM2
100/100/99
Trichodesmium IMS101
59/--/61
Anabaena PCC7120
Chlamydomonas
Spinacia oleracea
Arabidopsis thaliana
0.1

Suppl. Fig. 9

Chapter V

Prochlorococcus cyanophage genomes: Phages adapted for infection of open-ocean photosynthetic cells

Prochlorococcus cyanophage genomes: Phages adapted for infection of open-ocean photosynthetic cells

Matthew B. Sullivan et al.[1]

ABSTRACT

Phage genomic analyses suggest that dsDNA phage evolve through the exchange of modular cassettes of genes, perhaps forming a limited number of phage types which contain unique genes that are specialized for interactions with particular hosts. Here we present an analysis of three genomes from phages that infect *Prochlorococcus,* the numerically dominant primary producer in the vast oligotrophic surface oceans. These phage genomes contain core genes that suggest that podovirus P-SSP7 is a T7-like phage, while both myoviruses P-SSM2 and P-SSM4 are T4-like phages. They thus appear to be variations of two well-known phages, but have been modified for the infection of photosynthetic hosts living in oligotrophic environments. All three phage genomes encode core photosynthetic proteins that are full-length, conserved, and clustered in the genome, features which suggest they are being maintained by selection. The P-SSP7 genome also contains an integrase gene that suggests this phage is capable of integrating into its host. If this gene is functional, this would not only have evolutionary and ecological implications for both phage and host, but would also represent the first temperate T7-like phage and the first temperate marine cyanophage in culture. In addition, the P-SSM2 and P-SSM4 genomes collectively encode proteins that may be important for allowing the phage and/or host to respond to phosphate stress, as well as mobilize carbon stores and synthesize nucleotides, lipopolysaccharides, cobalamin and accessory pigments. We hypothesize that many of these phage-encoded genes are functional and likely aid in utilizing host metabolic capabilities during infection.

[1] This chapter is the draft of a manuscript that includes the following co-authors (in alphabetical order): Jan-Fang Chang, Sallie W. Chisholm, Maureen Coleman, Jane Grimwood, David Mead, Forest Rohwer, Peter Weigele

INTRODUCTION

The taxonomy of bacterial viruses (phages) is highly controversial and actively debated (Lawrence, Hatfull, and Hendrix, 2002; Rohwer and Edwards, 2002). This controversy arises as the result of observations among the 'lambdoid' phages (those phages that are able to recombine productively with coliphage lambda) that their DNA exhibits regions of high sequence similarity adjacent to regions of low sequence similarity which suggests that these genomes might evolve through the horizontal exchange of genetic material (Simon, Davis, and Davidson, 1971). Genome sequencing, primarily of lambdoid phages, has confirmed these observations and it has been hypothesized that perhaps all dsDNA phages primarily evolve through the horizontal exchange of genetic material, in the form of modular functional cassettes, through a global phage genome pool (termed the mosaic theory of phage evolution)(Hendrix et al., 1999). While any given phage has access to all the sequences in the global phage genome pool, it is suggested that access to this global phage genome pool must be limited in some manner (Hendrix et al., 1999), perhaps by host range (i.e., the range of hosts a given phage can infect)(Sullivan, Waterbury, and Chisholm, 2003).

If phage had access to a global phage genome pool, then one would predict that some phages might even contain genes considered to be characteristic of a different domain of life (i.e., a phage that infects exclusively bacteria might contain a eukaryotic or archaeal gene, not present in its hosts)(Hendrix, 2002). This has indeed been observed among the genomes of the mycobacteriophage (phages that infect *Mycobacteria*) (Pedulla et al., 2003). Further, examples of other potential gene exchanges between phages are commonplace in the current literature, particularly among the lambdoid phages which are relatively well represented in phage genome databases, but also among two other relatively well represented groups of phage genomes: the mycobacteriophage (Pedulla et al., 2003) and the "dairy phages" (phages that infect lactic acid bacteria) (Brussow, 2001). However, between these groups, there is a quantitative difference in horizontal gene exchange with the lambdoid phages being highly mosaic while the dairy phages are the least mosaic (Brussow and Hendrix, 2002).

Mechanistically, the setting for this horizontal exchange implied in the mosaic theory of evolution requires phage DNA to be in close proximity with DNA from a foreign source to allow genetic exchange. This occurs within multiply-infected hosts, where either two lytic (incapable of integration into the host chromosome) phages co-infect the same host or where a lytic phage infection occurs when a temperate (capable of stable integration of its genome into the host

chromosome) phage has already integrated into the host genome (Hendrix, 2003). In either of these cases, genes could be shared through homologous or non-homologous recombination (through an unknown mechanism) (Hendrix, 2003). Further, in the case of the temperate phage, another opportunity to horizontally obtain genes (in this case from the host) can occur through imprecise excision of the phage genome upon induction (Calendar, 1988). Thus through recombination with other phages and hosts, any given phage has the potential to access all the sequences in the global phage genome pool. This gene pool includes variants of not only the structural and developmental genes directly necessary for phage life cycles, but also those genes that have been copied from hosts and are being maintained in the phage gene pool (Hendrix et al., 2000). If lytic phage genomes have lower opportunity for recombination than the temperate phages, then these phage genomes might be expected to be less prone to horizontal exchange of genes and maybe even form biologically cohesive units that could allow a taxonomic hierarchy to apply due to evolution largely by speciation, genetic adaptation and accumulation of point mutations (Kovalyova and Kropinski, 2003).

While the genomes of lytic phage types (e.g., phages with genomes like those of the coliphages T7 and T4) are underrepresented in the phage genome databases, analyses of these genomes are beginning to reveal a large number of fixed, essential genes, interspersed with highly variable, non-essential genes which do *not* appear to be shared across groups (Desplats and Krisch, 2003; Miller et al., 2003a; Molineux, in press). However, the hosts of these phages are primarily alpha, gamma, delta proteobacteria (only one cyanobacterial host) so our understanding of core genes within these phage types could be broadened with sequence information from phages infecting a broader diversity of hosts that would more generally represent variations upon the basic lytic life style.

A group of phages that are ideal for such focused sequencing efforts are the marine phages. They are abundant (Bergh, 1989; Bratbak et al., 1990; Proctor and Fuhrman, 1990), are significant contributors to global biogeochemistry (Fuhrman, 1999), and are likely to include a wide diversity of types for comparison to the nonmarine phages already in the database (Paul et al., 2002). Although there are over 170 phage genomes in GenBank, only seven of these sequenced phages infect marine hosts (cyanophage P60; vibriophages VpV262, KVP40, VP16T, VP16C; roseophage SIO1; *Pseudoalteromonas* phage PM2), and only one infects cyanobacteria (cyanophage P60). The genomes of cyanophage P60 (Chen and Lu, 2002), roseophage SIO1 (Rohwer et al., 2000) and vibriophage VpV262 (Hardies et al., 2003) contain genes with high

homology to nonmarine phages, but also open reading frames (ORFs) with little homology to any known phage or host genes (Chen and Lu, 2002; Hardies et al., 2003; Rohwer et al., 2000), suggesting some support of the mosaic theory but also significant ambiguities, which will only be resolved as more marine phage genomes enter the databases. Marine phage genomes will permit a comparison of the evolutionary and ecological relationships of phages infecting marine versus non-marine hosts where different pressures may shape their genomes (e.g., differences in dispersal strategies, acclimation to environmental stressors).

The marine cyanobacteria *Prochlorococcus* are abundant (to 10^5 cells ml^{-1}) in the oligotrophic surface oceans (Partensky, Hess, and Vaulot, 1999) where nutrients, in particular phosphate, are often limiting (Karl, 1999; Wu et al., 2000). Genomic analyses of two marine phages, roseophage SIO1 (Rohwer et al., 2000) and vibriophage KVP40 (Miller et al., 2003a), has revealed that these genomes encode a phosphate-inducible gene (*phoH*). Such an observation suggests that the genetic composition of these phages has been shaped by their host's environments. *Prochlorococcus* phages are abundant in these regions (~10^3 phage ml^{-1}) (Sullivan, Waterbury, and Chisholm, 2003). Work with phages that infect the closely related cyanobacteria *Synechococcus* suggests that these phages play a small but significant role in host mortality (Mann, 2003) while phylogenetic inference from *Prochlorococcus* phage and host genomes suggests that they are also likely to regulate population dynamics through the movement of genetic material through the host population (Lindell et al., submitted). Because phage genomics has yielded information that complements what is known about the ecology of marine viruses and their hosts, we hoped to use genomic sequencing to further our understanding of *Prochlorococcus* and their phages.

Here we present analyses of three genomes from *Prochlorococcus* phages that were classified using electron microscopy as belonging to two morphological families: *Podoviridae* for P-SSP7 and *Myoviridae* for P-SSM2 and P-SSM4 (Sullivan, Waterbury, and Chisholm, 2003). The podovirus P-SSP7 infects a single high-light adapted (HL) *Prochlorococcus* strain, MED4, while the first myovirus P-SSM2 infects 3 low-light adapted (LL) *Prochlorococcus* strains (MIT 9211, NATL1A, NATL2A) and the other myovirus P-SSM4 infects 2 HL and 2 LL *Prochlorococcus* strains (MED4, MIT 9215, NATL2A, NATL1A)(Sullivan, Waterbury, and Chisholm, 2003). None of these phages cross-infects any of ten *Synechococcus* strains tested thus far. We had no prior knowledge of the gene content of these phage, thus with regards to their genomes, these phage were selected randomly for this study. At the genome level, these phages

appear in some ways to be T7- and T4-like phages, but are genetically adapted for infection of photosynthetic hosts in oligotrophic environments.

MATERIALS AND METHODS

Preparation of cyanophage for genome sequencing

Phages were prepared for genomic sequencing as previously described (Chapter 4)(Lindell et al., submitted). Briefly, phage particles were concentrated from phage lysates using PEG, then DNA-containing phage particles were purified from other material in phage lysates using a density cesium chloride gradient. These purified phage particles were broken open (SDS/proteinase K) and DNA was extracted (phenol:chloroform) and precipitated (ethanol) to provide small amounts of DNA (<1 μg). A custom LASL clone library was constructed by Lucigen Inc (Middletown, WI) as described previously (Breitbart et al., 2002). Subsequent clone libraries were constructed by DOE JGI using a similar protocol to provide further coverage of the phage genomes. Inserts were sequenced by the DOE JGI from all of these clone libraries and used for initial assembly of these phage genomes. The Stanford Finishing Group closed the genomes using primer walking.

Gene identification and characterization

The genomes were examined to identify open reading frames (ORFs) using GeneMark (Besemer, Lomsadze, and Borodovsky, 2001). Translated ORFs were analyzed for homology to known proteins in the non-redundant GenBank database and in the KEGG database by using the BLASTp program. Closest homologues were also compared to the query sequences to examine for the presence of domains and length similarity of the genes. Translated ORFs were also analyzed for signal sequences and transmembrane regions using the web-based software SignalP and TMHMM respectively (available at the CBS prediction servers http://www.cbs.dtu.dk/services/). In some cases, ORF annotation was also aided by the gene size, domain conservation and/or synteny (gene order), the latter as suggested for highly divergent genes encountered during phage genome annotation (Brussow and Hendrix, 2002).

RESULTS AND DISCUSSION

General characteristics

While the *Prochlorococcus* P-SSP7 podovirus genome is not to finished quality at time of this writing, draft sequence data suggests this genome is ~43 kb, 38.7%GC and contains 53 ORFs, terminal repeats and encodes its own RNA polymerase gene (RNAP), DNA polymerase gene (DNAP) and an integrase gene (for comparison to other *Podoviridae* see Table 1). The P-SSM2 genome is 252,401 bp, 35.5%GC and contains 327 ORFs, while the P-SSM4 genome is 178,249 bp, 36.7%GC and contains 198 ORFs (for comparison to other *Myoviridae* see Table 2). These myoviruses do not contain an integrase gene, and the fact that both of these genomes assembled and closed suggests their DNA is circularly permuted. These characteristics suggest that, generally speaking, P-SSP7 is a T7-like phage (though unique in containing an integrase gene) within the *Podoviridae*, while both *Myoviridae* P-SSM2 and P-SSM4 are T4-like phages. We test this hypothesis by looking in more detail at the complement of genes found within each of these genomes.

Core genes of T7-like phages

Comparative genomic analyses of T7-like phages suggests that they are characterized by a core of 24 genes that are widespread among them (Table 3) and are essential for T7 lytic growth (Molineux, in press). We have identified 16 of these genes in P-SSP7 (Table 3) and, strikingly, their gene order is also conserved (Fig. 2). By analogy to T7, these 16 genes should allow for the majority of host interactions and phage production including the following (T7-like gene designations for genes involved are shown in parentheses): shutdown of host transcription (*0.7*), degradation of host DNA (*3, 6*), DNA replication (*1, 2.5, 4, 5,* host thioredoxin), formation of a channel across the cell envelope via an extensible tail (*14, 15, 16*), DNA packaging (*18, 19*), and the formation of the virion structure (*8, 9, 10, 11, 12*)(Molineux, in press). Notable genes that are absent will be discussed below.

What defines a T7-like phage? The lytic T7-like phages have long been characterized by a *Podoviridae* morphology as well as a genome size of approximately 40 kb and the presence of terminal repeat sequences at the ends of the genome and a rifampicin resistant RNA polymerase (RNAP) gene (Hausmann, 1988). However, cases exist where the T7-like designation is unclear. For example, both vibriophage VpV262 and roseophage SIO1 have approximately 40 kb genomes containing terminal repeats and DNA replication genes with significant homology to

T7-like phages, but lack an RNAP (Hardies et al., 2003; Rohwer et al., 2000). In contrast, *Xanthomonas*-infecting phage Xp10 contains a rifampicin RNAP but has lambda morphology (Yuzenkova et al., 2003). Are these T7-like phages? A recent review chapter suggests that as the known diversity of T7-like phage increases, there are three increasingly distinct genome types emerging within the T7 group (Molineux, in press). The first type is exhibited by coliphages T3 and T7, yersiniaphages ϕA1122 and ϕYe03-12, and *Pseudomonas putida* phage gh-1, which exhibit a highly conserved genome organization that differs primarily in the presence or absence of several non-essential genes. The second T7-genome type includes coliphage K1-5, the salmonellaphage SP6, the *Pseudomonas aeruginosa* phage ϕKMV, the *Xanthomonas* phage Xp10, and the marine cyanophage P60, which have more dynamic genomes and different genome architecture from other T7 phages, but do contain a small subset of genes homologous to other T7-like phages, including an RNAP. Finally, the third T7-genome type contains genes homologous to the DNA replication machinery of T7-like phages, but has limited similarity of either overall genome organization or homology to any other T7-like genes and lacks an RNAP. This type is represented by several distantly related T7 phages, including two marine phage genomes from roseophage SIO1 and the vibriophage VpV262. Based upon the conserved overall genome organization and significant similarity of genes between the *Prochlorococcus* P-SSP7 phage genome and other T7-like phage genomes (Table 3, Fig. 2), we suggest this phage represents a T7-like phage.

Core genes of T4-like phages

Comparative genomics has suggested that T4-like phages contain a core of genes that are required during phage infection. Among the six T4-like phage genomes currently available, a comparison of the presence and absence of various genes suggests a core of 18 genes involved in DNA replication, recombination and repair, 7 genes involved in transcriptional and translational regulation, 10 genes involved in nucleotide metabolism and 34 genes involved in the virion structure (Table 4). We identified many of these genes in P-SSM2 and P-SSM4 and also found that as is the case of other T4-like phage genomes, these genes are often grouped into functional clusters within the genomes (Figs. 3, 4).

What defines a T4-like phage? The T4-like phages are differentiated from other *Myoviridae* by genome size, host type, presence or absence of an integrase gene and the conformation of the DNA found in the particle (Table 2)(van Regenmortel et al., 2000). The

marine vibriophage KVP40 is a *Myoviridae* phage that is most similar to T4-like phages (Miller et al., 2003a). Vibriophage KVP40 shows extensive regions of genome sequence and organization were conserved with phage T4, perhaps defining 'core' genes, while 65% of the ORFs were unique to KVP40 with no known function and while missing genes involved in host DNA degradation and early and middle transcription (Miller et al., 2003a). Such core genes are suggested to be common to other T4-like phages (Desplats and Krisch, 2003), but these analyses are constrained to genomes from phages that infect gamma and delta proteobacteria including this marine vibriophage. Based upon the high number of genes homologous to T4-like phage genes (Table 4), the lack of genes homologous to other known phage genes, and the distributed synteny of these genes into T4-like functional groups (Figs. 3,4), we suggest that these *Prochlorococcus* myoviruses are T4-like phages.

It is striking that the genomes from cyanobacterial phages contain a significant portion of the core genes found in the well-studied T7-like (podovirus P-SSP7) and T4-like (myoviruses P-SSM2 and P-SSM4) phages that infect only proteobacteria (alpha, delta, gamma) and in one case cyanobacteria (Tables 3, 4). For nearly 60 years (Demerec and Fano, 1945), these coliphages have been model systems for understanding the molecular underpinnings of phage infection. T7 and T4 are optimized for lytic infection of fast-growing *E. coli* hosts (which have 20 minute doubling times) that commonly dominate niches characterized by episodic feast-famine nutrient conditions and both anaerobic and aerobic conditions. Within minutes of either phage attaching to their host cell surface, host metabolism is compromised and diverted towards the transcription and translation of the phage genome for making phage particles (Kruger and Schroeder, 1981; Kutter, Guttman, and Carlson, 1994). In contrast, the marine cyanobacteria *Prochlorococcus* are slow growing (24-hour doubling time), oxygenic phototrophs that thrive in the nutrient-poor, aerobic surface waters of the tropical and sub-tropical oceans (Partensky, Hess, and Vaulot, 1999). We hypothesize that these fundamentally different host generation times and environmental conditions act as selective agents upon the sorts of genes carried by phages for infection of their respective hosts.

Photosynthetic genes in cyanophage

All three *Prochlorococcus* phage genomes contain core photosynthetic genes previously reviewed in Chapter 4 (Lindell et al., submitted). Because infection of *E. coli* by T4 and T7

phages results in shutdown of host gene transcription (Kruger and Schroeder, 1981; Miller et al., 2003b) and turnover of core photosynthetic proteins (e.g., D1) is rapid, if photosynthesis is to continue, it would be important for a phage to insure production of such critical proteins during infection. Thus, such phage-encoded photosynthetic genes are hypothesized to be important for maintaining an active PSII reaction center to allow continued photosynthesis during phage infection (Lindell et al., submitted; Mann et al., 2003; Millard et al., submitted).

In addition to these core photosynthetic genes found in each *Prochlorococcus* phage genome, the myoviruses contain additional photosynthesis-related genes – some of which are found in both phages and some of which are found in only one of the two myoviruses. The P-SSM2 genome also contains two genes (*ho1*, *pebA*) that encode proteins likely involved in biosynthesis of the accessory photosynthetic pigment, phycoerythrobilin, from a heme group (Frankenberg et al., 2001) and a *nifU*-like gene that is thought to be important for biosynthesis of iron-sulfur clusters in cyanobacteria (e.g., hemes) (Nishio and Nakai, 2000). The myovirus P-SSM4 genome also contains a gene (*pcyA*) that encodes a protein involved in the biosynthesis of the accessory photosynthetic pigments, phycocyanobilin (Frankenberg and Lagarias, 2003) as well as a gene (*speD*) which encodes a protein involved in the synthesis of polyamines. Many of these genes (exception = *speD*) occur in the marine unicellular cyanobacteria sequenced to date. Thus, it is unclear why the production of such accessory pigments may be useful to a phage during infection. One possibility might be that such pigments could further mitigate incoming photon fluxes to the PSII reaction center, again providing some level of protection for PSII to allow photosynthesis to continue throughout phage infection. S*peD* is not found in marine unicellular cyanobacteria, but its usefulness in generating polyamines, which are known in higher plants to affect the structure and oxygen evolution rate of the PSII reaction center (Bograh et al., 1997) suggests that perhaps the *speD* gene product (polyamines) is also useful for maintenance of the PSII reaction center, again allowing photosynthesis to continue.

Photosynthetic organisms only fix carbon through photosynthesis during daylight hours. The transaldolase protein (encoded by the *tal* gene) is a key enzyme in the oxidative pentose phosphate pathway that is not active in cyanobacteria during daylight hours. Thus, in cyanobacteria this pathway is utilized for maintenance of energy during dark metabolism and converting the fixed carbon products of photosynthesis to sugars/nucleotides/amino acids via a ribulose-5-phosphate intermediate (Schmetter, 1994). It has been previously hypothesized that a

phage-encoded *tal* gene may allow access to stored carbon pools during non-photosynthetic periods for continued phage production (Millard et al., submitted).

Genes involved in the specialization of podovirus P-SSP7 to its host metabolism

In addition to core T7 genes and photosynthesis-related genes, the podovirus P-SSP7 genome also contains an *int* gene, which encodes a site-specific recombinase protein (Table 3). The *int* gene contains conserved amino acid motifs previously identified for site-specific recombinases, which suggests a functional role (Nunes-Duby et al., 1998)(Appendix B). The presence of a putative site-specific recombinase in the podovirus P-SSP7 genome is consistent with a previous hypothesis (Chapter 2) that the integrating phase (as a prophage) of the temperate phage life cycle may be selected for in the oligotrophic, open oceans where these isolates were obtained (Sullivan, Waterbury, and Chisholm, 2003). However, the genomes of currently available freshwater cyanobacterial genomes (Canchaya et al., 2003; Casjens, 2003) and marine cyanobacterial genomes lack intact prophage (Dufresne et al., 2003; Palenik et al., 2003; Rocap et al., 2003) and despite repeated attempts, no temperate phage has been induced from marine cyanobacterial cultures (Waterbury and Valois, 1993). Indirect measures from induction experiments in the field are suggestive that temperate cyanophage can be induced from *Synechococcus* (McDaniel et al., 2002; Ortmann, Lawrence, and Suttle, 2002), but no integrated prophage has been demonstrated and temperate cyanophage are not yet in culture. If this integrase gene in podovirus P-SSP7 is demonstrated to be functional and allows this phage to integrate into the host genome, then this would be the first example of a temperate T7-like phage and of a temperate marine cyanophage in culture.

The *Prochlorococcus* P-SSP7 genome also contains a putative gene for ribonucleotide reductase, which catalyzes the thioredoxin mediated reduction of diphosphates (e.g., ADP, GDP, TDP, CDP). Thus far, ribonucleotide reductase genes among T7-like podovirus genomes have only been observed in those from marine environments (Table 3). The roseophage SIO1 and cyanophage P60 phage genomes each contain 1 and 2 copies, respectively, of ribonucleotide reductases (Chen and Lu, 2002; Rohwer et al., 2000). Those found in cyanophage P60 are most similar to cyanobacterial ribonucleotide reducases (Chen and Lu, 2002) which are thought to be vitamin B12-dependent (Gleason and Olszewski, 2002). In both of these phages and the *Prochlorococcus* phage P-SSP7, the location of the putative ribonucleotide reductase gene is within a region homologous to the class II gene region of coliphage T7 that is involved in

nucleotide metabolism (Molineux, in press). The presence and location of putative ribonucleotide reductase genes in all three marine T7-like podovirus genomes sequenced to date suggests that the incorporation of nucleotides from degraded host DNA that has been well characterized for many phage host-systems (Calendar, 1988) including Roseophage SIO1 (Wikner et al., 1993) is particularly important for efficient nutrient utilization during T7-like phage infection of nutrient-deprived marine hosts.

Host metabolic genes present in both myoviruses – required for myovirus infection of *Prochlorococcus*?

In addition to core T4-like phage genes and the photosynthesis-related genes, both myovirus genomes contain genes encoding proteins that are induced during phosphate stress (*pstS, phoH*) (Kim et al., 1993; Scanlan et al., 1997; Torriani, 1990), involved in carbon metabolism (*tal*) (Sprenger, 1995), the biosynthesis of cobalamin (*cobS*) (Lawrence and Roth, 1995; Maggio-Hall and Escalante-Semerena, 1999) and lipopolysaccharides, as well as the modification of phage DNA (Miller et al., 2003b).

Phosphate is a scarce resource in the oligotrophic oceans (Karl, 1999; Wu et al., 2000) where *Prochlorococcus* are abundant. Interestingly, both *Prochlorococcus* myoviruses contain a *phoH* gene, which in *E. coli* is induced under phosphate stress and thought to encode an ATPase (Kim et al., 1993). The *phoH* gene family is widely distributed among prokaryotes (Kazakov et al., 2003) and has been identified in two other marine phage genomes (Miller et al., 2003a; Rohwer et al., 2000). The presence of the *phoH* gene in four marine phages suggests it has an important role under conditions where phosphate may be limiting. However, because the role of PhoH is unknown, we can only speculate that it is likely involved in mobilizing phosphate, perhaps through an anabolic phospholipids metabolism (Kazakov et al., 2003).

Both myovirus genomes also contain the gene, *pstS*, which encodes a periplasmic phosphate binding protein involved in phosphate uptake (Wanner, 1996). In marine *Synechococcus* endogenous *pstS* is induced under limiting phosphate conditions (<50 nM inorganic phosphate) (Scanlan et al., 1997). Hence, expression of a phage-encoded *pstS* gene might provide greater access to phosphate pools during infection of phosphate-stressed cells. Although multiple Pho-Boxes (transcriptional promoter sequences bound by the protein regulators of the pho-regulon) were identified in roseophage SIO1, none were identified in either *Prochlorococcus* myovirus genome using the following consensus sequence

(CTGTCATA[AT]A[AT]CTGT[CA]A[CT]) (as suggested by (Wanner, 1996). This may suggest that the regulation machinery of the phosphate-inducible genes in cyanobacteria may recognize a different Pho-Box than in other bacteria. Alternatively, it might suggest that during phage production phosphate is always limiting, thus the phages might not want to regulate expression of a gene that encodes a phosphate uptake protein.

Finally, the presence of *mazG* in both *Prochlorococcus* myoviruses further suggests that manipulation of nucleotides is important to infection by myoviruses of marine hosts. This gene is widely distributed among prokaryotes (Zhang and Inouye, 2002; Zhang, Zhang, and Inouye, 2003) and is a member of a newly described gene family encoding nucleoside triphosphate pyrophosphohydrolase / pyrophosphatases which catalyze the conversion of GTP to and from GMP.

Cobalt is also found in extremely low concentrations (pM) in seawater (Saito and Moffett, 2002) and is required for growth of *Prochlorococcus* (Saito et al., 2002). Both *Prochlorococcus* myoviruses contain the gene, *cobS*, which in bacteria encodes a protein that catalyzes the final step in cobalamin (vitamin B12) biosynthesis. The marine cyanobacterial genomes contain an intact cobalamin pathway including *cobS* (E. Webb, pers. comm.), which supports the anecdotal evidence that they can make their own vitamin B12 as it is not required in their growth medium. Both myoviruses encode ribonucleotide reductases which are most similar to those found in the marine cyanobacteria, which in the cyanobacteria are thought to require cobalamin (B12) as a cofactor (Gleason and Olszewski, 2002). Thus, the production of cobalamin during *Prochlorococcus* phage infection could be important for nucleotide metabolism provided by the activity of ribonucleotide reductase. Alternatively, the overexpression of the *cobS* gene has recently been shown to induce the phage shock protein *PspA*, known to destabilize membranes in *Salmonella* (Escalante-Semerena, pers. comm.). Could expression of phage-encoded *cobS* also provide a novel mechanism for cell lysis upon completion of lytic phage production?

Both *Prochlorococcus* myovirus genomes contain putative N6-adenine methyl transferases and N4-cytosine methyltransferases (Table 4). Such enzymes are used in T4 to modify the DNA to prevent degradation of phage DNA by host restriction enzymes upon infection and can allow specific recognition of phage DNA by host RNA polymerase during early transcription (Miller et al., 2003b). Thus, it is likely that both *Prochlorococcus* myovirus genomes contain modified bases for similar reasons. Notably, these modification enzymes are

sporadically distributed among the T4-like phages (Table 4) and, in particular, are absent from vibriophage KVP40 where it is unknown how the phage DNA is differentiated from host DNA in early transcription (Miller et al., 2003a).

Taken together, the presence of these many metabolic genes in both *Prochlorococcus* myovirus genomes suggests numerous links between phage and host metabolism, which may be critical to successful phage production.

Metabolic genes present in one but not both myoviruses – genes transiently passing through or necessary genes for specialized phage-host interactions?

There are many genes in these *Prochlorococcus* myovirus genomes that are found only in one of the two and are typically absent from any other T4-like phages. The P-SSM2 genome contains significant number of genes unique to this myovirus (Table 4). The presence of five genes involved in purine (*purH, purL, purM, purN*) and pyrimidine (*pyrE*) biosynthesis suggests that this phage infects hosts where the *de novo* synthesis of nucleotides are limiting to phage production. Indeed, this is one of the largest T4-like phage genomes sequenced to date (252kb) and it infects hosts that have small (~1.7MB) genomes (as opposed to the large 255kb vibriophage KVP40 genome phage that infects the large ~3-4MB *Vibrio* host genomes) and live in highly oligotrophic conditions.

Further, three phage-encoded clusters of a total of 35 LPS genes (e.g., epimerases, transferases, phospholipases) were found in this phage genome. Phage-encoded LPS genes have previously been limited to temperate phages that use these genes upon infection and establishment of the prophage state to alter the cell-surface composition of the host (serotype conversion) to prevent superinfecting phage from attaching (thus conferring homoimmunity) (Calendar, 1988). While it is unlikely that this T4-like *Prochlorococcus* myovirus is temperate (lack of an integrase gene, no known temperate T4-like phage), there is reason to believe that this LPS gene cluster could still be functional and important to the phage. For example, the *Prochlorococcus* myovirus is likely to have a significantly long lytic cycle (extrapolating from *Synechococcus* phage lytic cycle lengths) (Suttle and Chan, 1993) relative to other T4-like lytic phages, thus it is plausible that the phage LPS genes might be expressed early during infection to alter the cell surface to provide similar exclusion of super-infecting phages. The presence of some phospholipases also suggests the possibility that LPS is degraded during infection for metabolic nutrients for phage production.

Finally, this phage genome contains a gene (*prnA*) that encodes a tryptophan halogenase which catalyzes the NADH consuming first of four steps of converting tryptophan to the antibiotic pyrrolnitrin (Hammer et al., 1999; Hammer et al., 1997; Kirner et al., 1998). The fact that this gene is full-length suggests that it might encode a functional protein. However, the other steps in pyrrolnitrin biosynthesis require enzymes produced by the *prnBCD* gene cluster (Hammer et al., 1999), but none of these genes, including *prnA*, are found in marine cyanobacterial genomes. If we assume that the intermediate product produced by the *PrnA* catalyzed reaction is of no use to the phage, then the lack of these other genes suggests that the *prnA* gene encoded by the phage is not being used during infection of marine cyanobacterial hosts and was likely acquired directly from a non-cyanobacterial host or through the fluid phage genome pool (Hendrix et al., 1999).

The other *Prochlorococcus* myovirus P-SSM4 genome also contains some unique genes, although these have perhaps less obvious roles in facilitating phage production. Two proteins with signaling domains, carboxylesterase and a cAMP class II phosphodiesterase, are encoded by this genome. Based upon their function in other organisms, it is likely they play a role in the modification of carbon compounds and nucleotide metabolism, both of which are important for phage production during infection. It is also noteworthy that this phage genome contains a group of some genes with little homology to known bacterial proteins, but with homology to eukaryotic prion-like proteins (e-5), an archaeal protease (e-6) and a hypothetical protein from a eukaryotic slime mold (e-4) all grouped near each other (Fig. 4). Similar observations of eukaryotic and prion-like genes have been made in the genomes of mycobacteriophages (Pedulla et al., 2003). The presence of these genes in a *Prochlorococcus* myovirus where they are unlikely to have a role suggests support for lateral transfer of genes to and from a common phage genome pool (Hendrix et al., 1999) – in this case gene transfer has managed to cross the domains of life. Alternatively, such genes might be misidentified due to lack of homology to genes with known function and may actually be important to particular phage-host interactions in their prokaryotic hosts.

Finally, one ORF in this phage has homology (e-33) to a possible hemagglutinin neuraminidase in *Prochlorococcus*. Such genes in the viral world have previously been limited to the ssRNA viruses commonly infecting humans. In these ssRNA viruses, hemagglutinin neuraminidase allows phage to attach to cellular receptors by cleaving sialic acid from glycoproteins. Glycoproteins can be formed from polysaccharides and proteins are commonly

important as cellular surface receptors residing on the external surface of the membrane where they are in contact with the environment (Madigan, Martinko, and Parker, 2003). Such a gene could be playing a similar role allowing attachment to a receptor for this *Prochlorococcus* phage.

The sporadic distribution of these genes suggests either that we are observing the transient passage of these genes through the phage genomic pool or that these genes are functional and being maintained by selection. If they are transiently passing through the genome, one might expect the gene to be degraded over time due to neutral mutation. However, if genes do not appear to be degraded, as is the case for the genes presented here, one might assume either that the genes are either relatively recent acquisitions from hosts where the gene was functional, or must be important to this particular phage and functional during some stage of the phage life cycle.

'Missing' genes

The fact that some genes required for T7 and T4 phage life styles are absent in our *Prochlorococococcus* phage genomes suggests two distinct possibilities. First, these genes may have significantly diverged from known phage genes thus preventing detection through homology searches. It is plausible that many of these missing genes are present in our phages but more difficult to identify due to the limited sampling of cyanophage genomes to date (Chen and Lu, 2002). If these genes are present, they may have lost sequence homology through significant sequence divergence of an ancient homologue within this under-sampled phage lineage or due to the replacement of a gene with a functional equivalent from a foreign source (termed non-orthologous displacement) (Forterre, 1999). Alternatively, these genes may not be required for infection of *Prochlorococcus* hosts and may be degraded by neutral mutations without selection for sequence conservation.

The podovirus P-SSP7 genome is missing a suite of genes found in other T7-like phages (Table 3). The protein products of these genes are likely host-specific as they are involved in protecting phage DNA against host restriction enzymes (gene *0.3*), specific interactions between the phage and host cell wall during formation of the virion structure (e.g., genes *6.7, 7.3, 13*) and encoding the lysis-mediating amidase (gene *3.5*) that is also responsible for regulation of T7 RNAP activity (Molineux, in press). These genes have yet to be identified in the other marine podovirus genomes, cyanophage P60 and roseophage SIO1 (Table 3). Thus, these genes are either significantly diverged from well-characterized genes (e.g., structural genes) or are not

required for infection of these marine hosts (e.g., anti-restriction proteins specific for particular hosts).

Analyses of both P-SSM2 and P-SSM4 myovirus genomes also reveal some missing genes whose loss may be a consequence of the specific marine host's lifestyle. First, both *Prochlorococcus* myoviruses lack the genes required for anaerobic nucleotide synthesis pathways (*nrdD*, *nrdG*, *nrdH*) that are commonly found in T4-like phages (including the marine vibriophage KVP40). However, while *Prochlorococcus* have been observed in oxygen minimum zones in the Arabian Sea (Johnson et al., 1999), it is unlikely that *Prochlorococcus* cells in the surface oceans of the Sargasso Sea are exposed to anaerobic conditions, thus there is no need for anaerobic nucleotide synthesis genes in the phage genomes. Second, the lack of tRNA genes encoded by the *Prochlorococcus* myoviruses (there is only 1 tRNA gene between them, Table 4) is striking because the other T4-like phages encode an average of ~16 tRNA genes per phage genome (range: 5-32, Table 4). In T4, eight tRNA genes supplement host tRNA species that are present in minor amounts but that recognize commonly occurring phage codons (Kunisawa, 1992). In contrast, vibriophage KVP40 which encodes 30 tRNAs has similar codon usage to its host, *Vibrio cholerae*, so it was suggested that the presence of so many tRNAs might reflect a broader host range by KVP40 (Miller et al., 2003a). Such a lack of tRNA genes in *Prochlorococcus* phages suggests that the codon usage of these phage genomes more closely matches that of their range of host genomes.

Finally, it should be noted that there are many virion structural genes that have yet to be identified in these *Prochlorococcus* myovirus genomes. It is clear from electron micrographs (Fig. 1 B, C) that the genes encoding functionally homologous proteins to those found in T4-like phages must be encoded in these genomes. In fact, there are many candidate ORFs, based upon size and location in the genome, that may eventually be identified as encoding the proteins for these structures. It is likely that, just as has been found for the other marine T4-like phage (vibriophage KVP40)(Miller et al., 2003a), many of these structural genes are significantly diverged and thus standard homology based searches will not identify these genes. More sophisticated approaches targeting regions of homology, coupled with laboratory work will greatly aid in this process.

Evolutionary fit enzymes?

The few 'host' genes that are found in these *Prochlorococcus* phages are predominately genes from the cyanobacteria that are also found in the host genomes as part of much larger pathways. If we assume that some (or all) of these phage-encoded genes are functional, then one must wonder why only particular genes in a much larger metabolic pathway are of value. In the oligotrophic environment where *Prochlorococcus* and their phage are abundant, it is likely that metabolic pathways which determine the use of key nutrients such as nitrogen, phosphorus and carbon are strictly controlled (Scanlan and West, 2002). It is plausible that one way for the host cells to control such pathways might be through modulating the activity of key pathway enzymes through transcriptional regulation over small time scales. For example, the transaldolase gene is not expressed in cyanobacteria while photosynthesis is occurring (Schmetter, 1994). However, if the phage-encoded transaldolase gene were resistant to regulation, thus constituitively expressed, this could allow the phage constant access to important metabolites during infection. Alternatively, over evolutionary time scales, there could be selection in the host for "control point" enzymes whose activity was decreased so that these enzymes would act as bottlenecks in a pathway (Watt and Dean, 2000). If there were selection for such control point enzymes with reduced activity, could phage-encoded versions of these same enzymes offer a more efficient form of the enzyme to maximize access to key metabolites during infection? This remains an open question.

CONCLUSIONS

Phages populations are genetically dynamic, and must not only maintain their ability to infect their hosts, they must also be able to delicately manipulate their hosts to enable optimal phage replication. The three *Prochlorococcus* phage genomes analyzed here provide an extensive demonstration of this intimate connection. First, all three *Prochlorococcus* phage genomes representing two types contain integral photosynthetic genes that are likely required to maintain photosynthetic activity of the hosts during infection (Lindell et al., submitted; Mann et al., 2003; Millard et al., submitted) which is critical to successful phage production (Sherman, 1976). Second, both myoviruses contain genes involved in key metabolic pathways that suggest that phosphate stress, cobalamin biosynthesis, carbon-store mobilization and modification of nucleotides are important for infection of *Prochlorococcus* by myoviruses. Third, presence of other metabolic genes in one but not both of the myoviruses may indicate genes that are important to the specialized interactions between the particular phage and host. Fourth, the integrase gene in the podovirus offers an adaptive strategy demonstrated in many other phage types, but not in the classically lytic T7-like and T4-like phages. If the *Prochlorococcus* P-SSP7 integrase is functional, this has significant evolutionary and ecological implications for the phage. Finally, these genomes provide support for the model that phages evolve by gene exchange through differential access to a global phage gene pool and with their hosts. The gene complements in these genomes suggest that these phages obtained genes from this global phage genome pool as well as from their host genomes, while maintaining core group-specific genes of the T7-like and T4-like phages. Understanding the mechanisms for this complex genetic shuffling will be a vital next step in our understanding of phage (and hosts) in the biosphere, as will be an understanding of the *functional consequences* of this gene diversity. The proposed roles played by many of the phage and host genes documented in these phage genomes awaits laboratory experiments designed to examine their expression and, if expressed, eventual protein localization / interactions to determine whether these genes are transiently passing through or play important functional roles during infection.

Figure 1: Electron micrographs of *Prochlorococcus* phages: (A) podovirus P-SSP7, (B) myovirus P-SSM2, (C) myovirus P-SSM4. Concentrated phage stocks were stained with 2% uranyl acetate. Note the distinct T4-like baseplate and tail tube, sheath and fibers in both myoviruses. Scale bars indicate 100 nm. Photo credit: Peter Weigele.

Figure 2: Genome arrangement of the *Prochlorococcus* podovirus phage P-SSP7. ORFs are sequentially numbered within the boxes and gene names are designated above the boxes. Gene designations are as per T7 nomenclature for T7-like genes (Molineux, in press) or as per microbial gene designation for non-phage genes. ORFs are all oriented the same direction in this genome. While the phage genome is one molecule of DNA, for viewing purposes the genome representation is broken to fit on a single page. Colors indicate the class I (light green), II (yellow) and III (light blue) genes as inferred from T7 phage. Note that the many T7 genes and genome order are conserved.

Figure 3: Genome arrangement of the *Prochlorococcus* myovirus phage P-SSM2. Gene names are designated above the box representing the ORF where genes were identified. If the ORF is located above the centering line then it is on the forward strand of DNA, while an ORF located below the line indicates reverse strand DNA. While the phage genome is one molecule of DNA, for viewing purposes the genome representation is broken to fit on a single page. Colors indicate the putative role for the identified genes as inferred from T4 phage. Unknown genes = White, 'Non-phage' genes = Pink, Transcription / Translation = red, nucleotide metabolism = yellow, DNA replication, recombination, repair = orange, head = dark blue, tail = light blue, purple = presumed lipopolysaccharide gene clusters, other phage genes = green. Gene designations are as per T4 gene nomenclature for T4-like genes (Miller et al., 2003b) or as per microbial gene designations for non-phage genes. Note the large gene, gp7, is predicted to be over 7000 amino acids long.

Figure 4: Genome arrangement of the *Prochlorococcus* myovirus phage P-SSM4. Color coding, ORF coding and gene nomenclature as in figure 3.

Figure 1

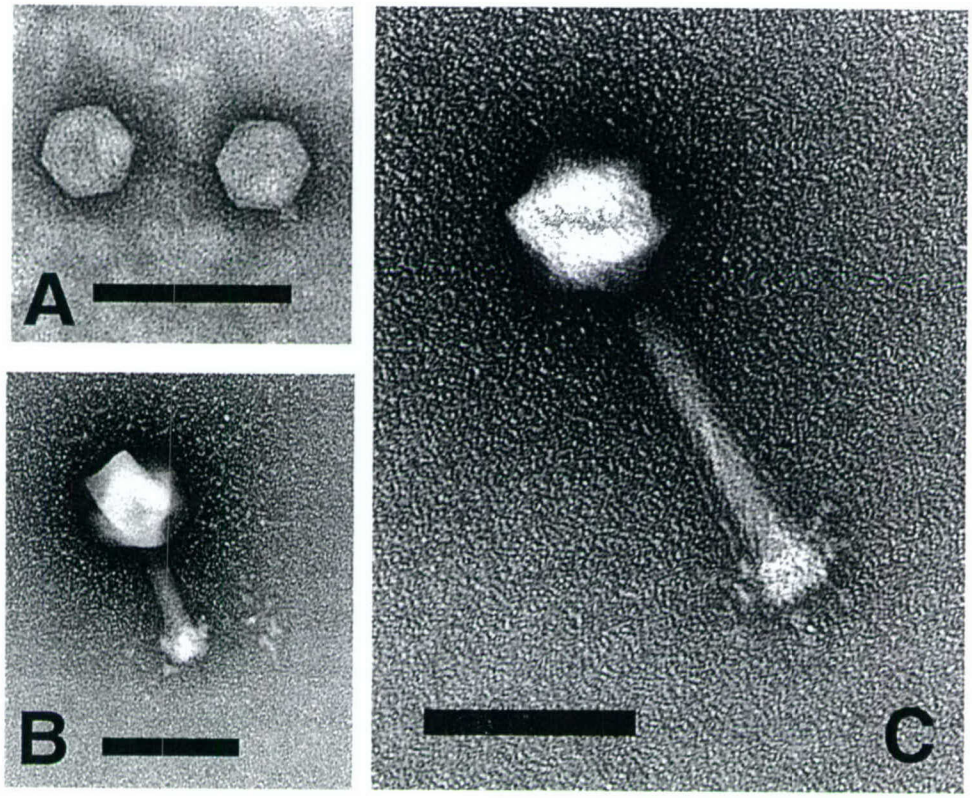


Figure 2

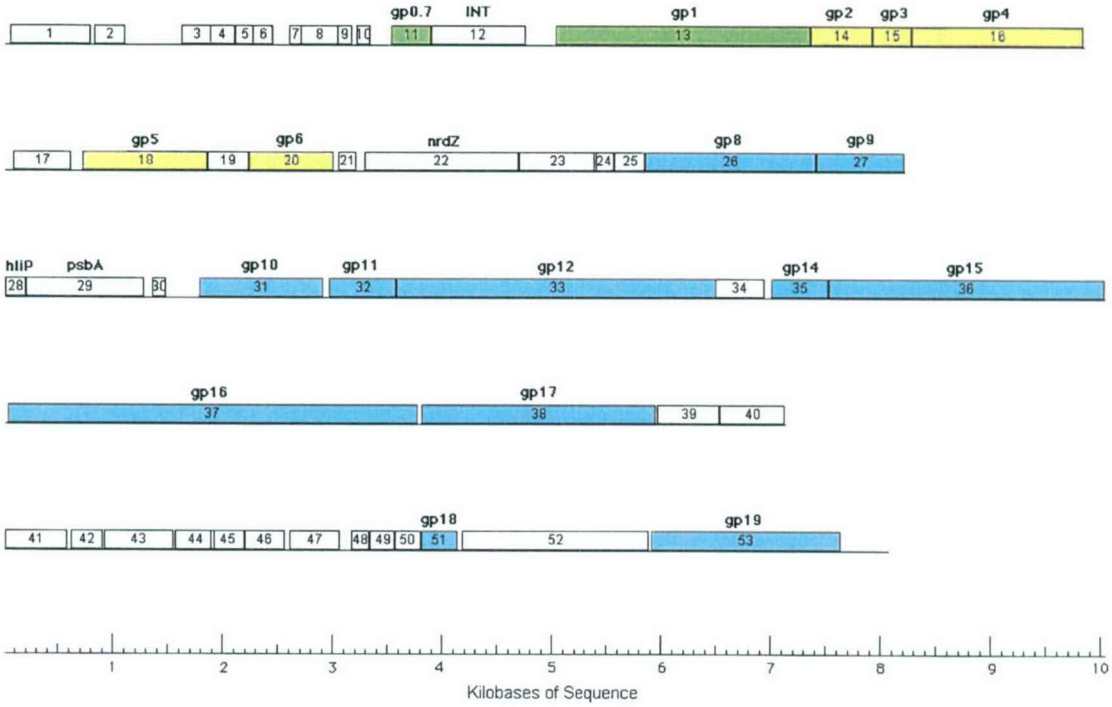

gp0.7
INT
gp1
gp2
gp3
gp4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
gp5
gp6
nrdZ
gp8
gp9
17
18
19
20
21
22
23
24
25
26
27
hliP
psbA
gp10
gp11
gp12
gp14
gp15
28
29
30
31
32
33
34
35
36
gp16
gp17
37
38
39
40
gp18
gp19
41
42
43
44
45
46
47
48
49
50
51
52
53
1
2
3
4
5
6
7
8
9
10
Kilobases of Sequence

Figure 3

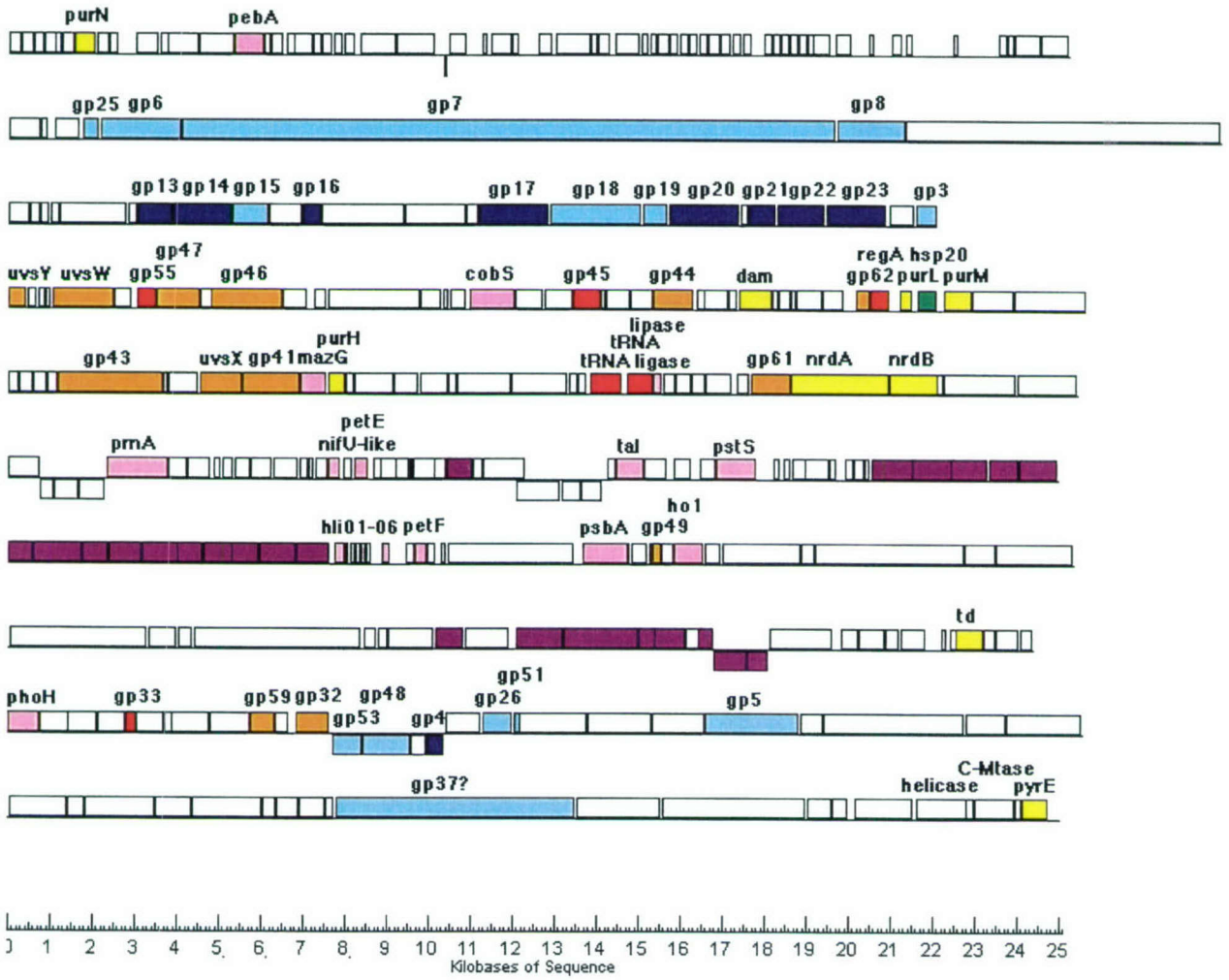

purN
pebA
gp25 gp6
gp7
gp8
gp13 gp14 gp15 gp16
gp17
gp18 gp19 gp20 gp21 gp22 gp23
gp3
gp47
uvsY uvsW gp55
gp46
cobS
gp45
gp44
dam
regA hsp20
gp62 purL purM
lipase
tRNA
purH
gp43
uvsX gp41 mazG
tRNA ligase
gp61 nrdA
nrdB
petE
prnA
nifU-like
tal
pstS
ho1
hli01-06 petF
psbA gp49
td
gp51
phoH
gp33
gp59 gp32 gp48
gp26
gp5
gp53
gp4
C-Mtase
gp37?
helicase
pyrE
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Kilobases of Sequence

Figure 4

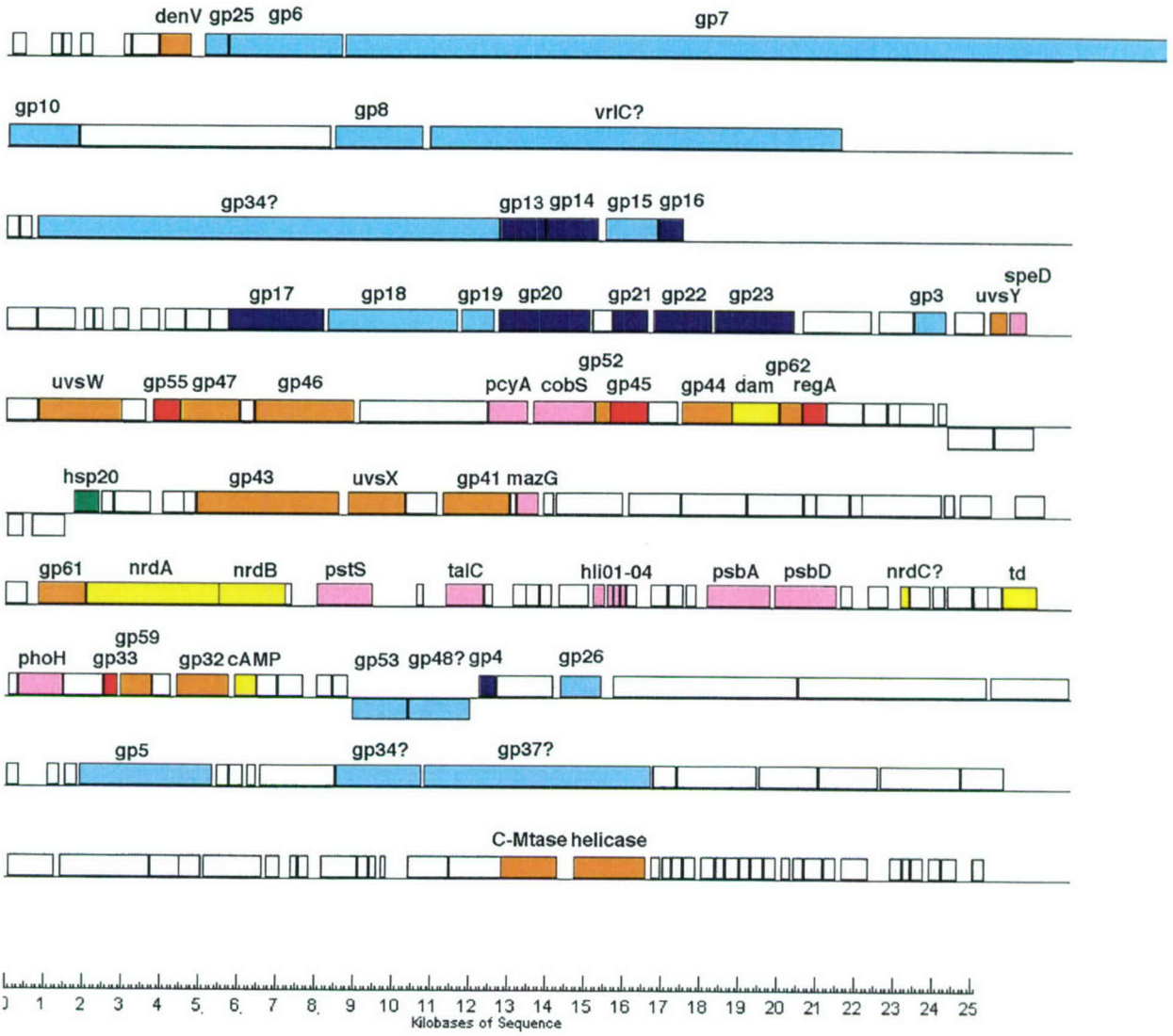

denV gp25 gp6
gp7
gp10
gp8
vrlC?
gp34?
gp13 gp14 gp15 gp16
gp17
gp18
gp19 gp20
gp21 gp22 gp23
gp3
uvsY
speD
uvsW
gp55 gp47
gp46
pcyA cobS
gp52
gp45
gp44 dam
gp62
regA
hsp20
gp43
uvsX
gp41 mazG
gp61
nrdA
nrdB
pstS
talC
hli01-04
psbA
psbD
nrdC?
td
phoH
gp33
gp59
gp32 cAMP
gp53 gp48? gp4
gp26
gp5
gp34?
gp37?
C-Mtase helicase
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Kilobases of Sequence

Table 1: Genome-wide characteristics of the *Prochlorococcus* cyanophage P-SSP7 relative to the other recognized phage groups within the *Podoviridae* (van Regenmortel et al., 2000). Categories include: type of host infected by the phages, genome size (kb), number of open reading frames (ORFs), terminal repeats (TRs), the presence of absence of an RNA polymerase gene (RNAP), the presence or absence of a putative site-specific integrase gene (INT). Y indicates that the feature is present, N indicates that the feature is absent.

Phage	*Host(s)*	*Size (kb)*	*# ORFs*	*TRs (bp)*	*RNAP*	*INT*
T7-like	*Gram negatives*	38-43	43-56	Y	Y	N
Φ29-like	*Gram positives*	18-22	17-35	N	N	N
P22-like	*Gram negatives*	38-50	60-65	N	N	Y
P-SSP7	***Prochlorococcus***	**43**	**53**	**Y**	**Y**	**Y**

Table 2: Genome-wide characteristics of of the *Prochlorococcus* cyanomyophages P-SSM2 and P-SSM4 relative to the other recognized phage groups within the *Myoviridae* (van Regenmortel et al., 2000). Categories include: the type of hosts infected by the phages, genome size (kb), number of open reading frames (ORFs), the presence or absence of an integrating phage element in the life cycle (INT). Y indicates that the feature is present, N indicates that the feature is absent, Y* indicates the phage integrates using a transposase rather than a site-specific integrase. ? indicates where no representative phage genomes have been completely sequences, so the presence or absence of the character is unknown.

Phage	*Host(s)*	*Size (kb)*	*# ORFs*	*INT*	*DNA conformation*
T4-like	Gram negatives	164-255	252-384	N	circularly permuted
P1-like	Gram negatives	<50 kb	40	?	circularly permuted
P2-like	Gram negatives	30-34	40-44	N	circularly permuted
Mu-like	Gram negatives	37	55	Y*	linear
Spo1-like	Gram positives	<50	40	?	linear
ϕH-like	Archaea	58-78	98-121	Y	circularly permuted
P-SSM2	***Prochlorococcus***	**252**	**327**	**N**	circularly permuted
P-SSM4	***Prochlorococcus***	**178**	**198**	**N**	circularly permuted

Table 3: 'Core' genes in T7-like phages (modified from (Molineux, in press). The size (amino acids) of each translated gene is presented for the genes classified using gene numbers according to T7 terminology. For P-SSP7, the size (amino acids) of each gene is followed by the best T7-like or microbe-related e-value in parentheses, with no e-value given for ORFs that were assigned using domain homology and synteny. The T7 supergroup is divided into T7-like phages (T7, T3, gh-1, ϕYe03-12, ϕA1122), the distant T7-like phages (e.g., P60) and the very distant T7-like phages (e.g., SIO1) (Molineux, in press). Representative phage(s) are included from each grouping for comparison. A '–' indicates the lack of a particular gene using standard seaches, * indicates a split gene split, and ** indicates a putative frameshift in the gene.

Gene	*P-SSP7 (e-value)*	*T7*	*T3*	*gh-1*	*ϕYe03-12*	*ϕA1122*	*P60*	*SIO1*	*notes*
Class I									
0.3	–	117	153	–	153	141	–	–	B-DNA mimic; anti-type I restriction
0.7	125 (e-18)	359	370	–	370	–	–	–	protein kinase to shut-off host transcription
1	779 (e-91)	883	885	886	885	884	574	–	RNA Polymerase
1.1	–	42	46	–	47	44	–	–	conserved; not essential
1.2	–	85	92	–	92	86	–	–	host dGTPase inhibitor; F-exclusion
1.3	–	359	347	–	347	341	–	–	DNA Ligase
Class II									
1.7	–	196	164	–	157	–	–	–	full length gene not conserved; beneficial for growth
2	–	64	55	56	79	65	–	–	host RNA polymerase inhibitor
2.5	190 (e=0.004)	232	233	234	233	233	–	–	SSB
3	117 (e-19)	149	153	148	154	152	–	135	Endonuclease I. Holliday junction resolvase
3.5	–	151	152	147	152	152	–	–	Amidase (lysozyme); regulates T7 RNAP activity
4A	521 (e-132)	566	567	563	567	567	531	523	primase-helicase; gp4B helicase has internal in-frame start
4.3	–	70	70	–	71	71	–	–	conserved, non-essential
4.5	–	89	94	–	95	90	–	–	conserved, non-essential
5	589 **	704	705	710	705	705	587	581	DNAP
5.7	–	69	69	70	70	70	–	–	conserved, non-essential
6	260 (e-44)	300	303	315	304	301	243	–	5'-3' dsDNA exonuclease; RNase H
Class III									
6.5	–	84	81	81	82	85	–	–	conserved, non-essential
6.7	–	88	83	91	84	89	–	–	essential virion ejected protein
7	–	133	107	–	–	134	–	–	non-essential, not conserved; host-range
7.3	–	99	–	–	107	80	–	–	essential virion ejected protein
8	523 (e-171)	536	536	544	536	537	555	–	head-tail connector protein
9	266 (e-50)	307	311	292	311	305		–	scaffolding protein
10A	376 (e-40)	345	348	348	348	345	221	–	major capsid protein; -1 frame-shift yields minor capsid protein gp10B, F-exclusion
11	205 (e-23)	196	197	196	197	197	192 *	–	tail protein
12	977 (e-58)	794	802	809	802	795	680	–	tail protein
13	–	138	137	145	139	139	–	–	essential; required for gp6.7 incorporation into virion
14	–	196	198	194	198	197	–	–	internal core protein; ejected into infected cell
15	838	747	749	739	748	748	–	–	internal core protein; ejected into infected cell
16	1246	1318	1319	1393	1321	1319	–	–	internal core protein; ejected into infected cell
17	716 (e-10)	553	559	619	646	559	–	–	tail fiber protein
17.5	–	67	67	72	68	68	–	–	class II holin
18	111 (e-11)	89	89	86	89	90	–	–	small terminase subunit
18.5	–	143	148	150	151	148	–	–	conserved; lambda Rz-RzI homologs
18.7	–	83	83	–	85	84	–	–	conserved; lambda Rz-RzI homologs
19	578 (e-121)	586	587	583	588	587	566 *	–	large terminase subunit
19.2	–	85	147	–	78	78	–	–	overlappon, conserved
19.3	–	57	57	–	43	58	–	–	overlappon, conserved
19.5	–	49	49	–	50	50	–	–	non-essential, conserved
Others									
phoH	–	–	–	–	–	–	–	385	phosphate-stress inducible; putative RNA helicase
nrd	469 (e-11)	–	–	–	–	–	406 / 446	672	ribonucleotide reductase domain
hli	64 (e-11)	–	–	–	–	–	–	–	high-light inducible
psbA	360 (e=0)	–	–	–	–	–	–	–	gene encoding D1 protein of PSII reaction center

Table 4: 'Core' genes in T4-like phages (modified from (Miller et al., 2003a; Miller et al., 2003b)). The T4 supergroup is divided into T-evens (e.g., T4, RB69), pseudo T-evens (e.g., RB49, 44RR2.8t), Schizo T-evens (e.g., Aeh1) and the Exo T-evens (e.g., S-PM2) (Desplats and Krisch, 2003; Tetart et al., 2001). The size of each translated gene (amino acids) is presented for the genes classified and named according to T4 terminology. For P-SSM2 and P-SSM4, the size of each translated gene and the e-value of the the best phage-T4-like-related or microbe-related e-value is presented; where no e-value is given these ORFs were assigned based upon size, domain homology and synteny. . A '–' indicates the lack of a particular gene, whereas a '?' suggests this gene may exist in the phage but is not readily identifiable.

Gene	*P-SSM2 (e-value)*	*e-value*	*P-SSM4*	*e-value*	*T4*	*RB69*	*RB49*	*44RR2.8t*	*KVP40*	*Aeh1*	*Description of protein product*
Transcription											
dsbA	–		–		89	95	91	95	90	95	DNA binding
33	87	e-7	87	e-4	112	112	89	85	98	76	Late activator
55	160	e-17	159	e-17	185	185	177	172	170	173	Late sigma
45	222	e-26	223	e-28	228	228	228	224	221	225	Late activator
asiA	–	–	–	–	90	90	–	90	99	97	Anti-sigma 70
rbpA	–	–	–	–	129	138	125	117	–	98	RNAP binding protein
alt	–	–	–	–	682	695	683	642	–	646	Adenylribosylation of RNAP; packaged in head
modA	–	–	–	–	200	200	–	–	–	–	ADP ribosylates alpha RNAP subunits
modB	–	–	–	–	207	193	167	–	–	–	ADP ribosylates ribosomal proteins S1, EF-Tu
motA	–	–	–	–	211	210	–	217	–	–	binds mot boxes at T4 middle promoters, transcriptional activator with *asiA*
Translation, RNA											
regA	148	e-34	140	e-34	122	125	120	118	132	122	Repressor
rnlA	–	–	–	–	374	374	389	383	381	389	RNA ligase 1
pseT	–	–	–	–	301	299	292	295	305	305	Polynucleotide kinase
rnlB	–	–	–	–	334	332	–	–	335	338[a]	RNA ligase 2
tRNAs	x1		–	–	x8	x2		x17	x30	x24	tRNA
regB	–	–	–	–	152	152	140	149	–	–	RNA RNase
Nucleotide metabolism											
frd	–	–	–	–	193	195	189	184	181	172	Dihydrofolate reductase; product used by *td*
nrdD	–	–	–	–	606	605	620	608	611	704	Anaerobic NTP reductase
nrdG	–	–	–	–	156	156	167	–	158	159	Anaerobic NTP reductase
td	212	e-84	211	e-78	286	238	410	279	300	277	Thymidylate synthase
nrdH	–	–	–	–	102	90	89	91	79	94	Anaerobic glutaredoxin

Table 4 continued

Gene	*P-SSM2 (e-value)*	*e-value*	*P-SSM4*	*e-value*	*T4*	*RB69*	*RB49*	*44RR2.8t*	*KVP40*	*Aeh1*	*Description of protein product*
cd	–	–	–	–	193	169	168	172	150	182	dCMP deaminase
tk	–	–	–	–	193	193	198	191	194	199	Thymidine kinase
nrdA	776	e=0	769	e=0	754	751	747	570	742	796[a]	Aerobic NDP reductase (alpha)
nrdB	385	e-104	388	e-104	388	390	386	323	374	376	Aerobic NDP reductase (beta)
nrdC			59	e=0.002	87	92	93	75	99	90	Aerobic thioredoxin
1	–	–	–	–	241	244	218	224	212	229	dNMP kinase
denA	–	–	–	–	144	136	153	138	–	145	Endonuclease II; degradation of non-modified cytosines
denB	–	–	–	–	185	158	–	–	–	–	ENase targets gaps / ssDNA; degrade host DNA
56	–	–	–	–	171	173	172	172	–	–	Removes dCTP,dCDP / dUTP,dUDP from precursor pool
alc	–	–	–	–	167	164	–	178	–	–	attenuates transcription elongation on non-modified cytosines
dam	265	e-24	276	e-43	260	–	270	–	–	–	N6-methyladenine
α-gt	–	–	–	–	401	–	–	–	–	–	Glucosylation of T4 DNA
β-gt	–	–	–	–	352	–	–	–	–	–	Glucosylation of T4 DNA
42	–	–	–	–	246	–	–	231	–	–	produces hydroxymethyl cytosine in T4
ndd	–	–	–	–	151	148	–	145	–	–	binds host HU protein and host DNA to membrane
	323	e-75	327	e-79	–	–	–	–	–	–	N4-methylcytosine
pyrE	215	e-44	–		–	–	–	–	–	–	Orotate phosphoribosyltransferase
purL	108	e-7	–		–	–	–	–	–	–	Phosphoribosylformyl glycinamide synthase
purM	223	e-80	–		–	–	–	–	–	–	Phosphoribosyl formyl glycinamide cyclo-ligase
purH	314	e-97	–		–	–	–	–	–	–	AICARFT/IMPCHase bienzyme
purN	175	e-33	–		–	–	–	–	–	–	phosphoribosyl glycinamide formyltransferase
	–	–	133	e-10	–	–	–	–	–	–	cAMP class II phosphodiesterase
Replication, recombination, repair											
rnh	–	–	–	–	305	290	315	307	335	306	RNase H
59	199	e-16	192	e-15	217	175	215	219	216	215	Helicase loader
32	273	e-42	308	e-45	301	299	322	295	304	302	SSB
uvsX	336	e-29	337	e-26	390	390	356	–	366	411	RecA-like protein
41	461	e-93	393	e-88	475	458	470	469	466	485	DNA helicase
61	324	e-48	278	e-36	342	340	342	334	352	344	Primase
dexA	–	–	–	–	227	225	222	221	230	226	Exonuclease A
39	–	–	–	–	516	606	607	607	601	613	Topoisomerase II
30	–	–	–	–	487	497	498	501	447	495	DNA ligase
47	347	e-41	344	e-41	339	339	341	355	346	342	Recombination nuclease
46	571	e-84	573	e-84	560	562	560	570	745	772	Recombination nuclease
45	222	e-26	223	e-28	228	228	228	223	221	225	Clamp
44	314	e-51	293	e-37	319	320	324	319	318	321	Clamp loader
62	111	e-8	135	e-43	187	187	192	192	163	193	Clamp loader
43	832	e-142	827	e-149	898	903	892	889[a]	850	919	DNA polymerase

Table 4 continued

Gene	*P-SSM2 (e-value)*	*e-value*	*P-SSM4*	*e-value*	*T4*	*RB69*	*RB49*	*44RR2.8t*	*KVP40*	*Aeh1*	*Description of protein product*
49	122	e-4	–		157	157	157	157	151	161	Recombination endonuclease VII
dda	397	e-8	424	e-9	439	437	463	439	421	454	DNA helicase
denV	–	–	185	e=0.46	138	137	–	152	138	–	N-glycosidase; pyrimidine dimer repair
52	–	–	–		442	441	454	555	428	438	Topoisomerase II
uvsW	488	e-82	488	e-75	587	504	500	493	507	503	RNA-DNA helicase
uvsY	143	e-9	99	e-4	137	164	135	–	144	136	UvsX assistant; SSB
Head											
40	–	–	–	–	114	114	105	113	–	105	Membrane assoc. initiator of head vertex
24	–	–	–	–	427	427	413	410	298	392	Vertex precursor
2	–	–	–	–	274	273	273	268	198	332	DNA end binding
4	146	e-33	109	e-23	150	149	157	150	151	156	Head completion
13	282	e-10	269	e-14	309	308	310	307	307	306	Head completion
14	471	e-23	309	e-26	256	254	246	252	278	262	Head completion
16	144	e-5	145	e-5	495	164	165	154	219	172	Terminase subunit
17	548	e-96	551	e-97	610	611	607	613	601	633	Terminase subunit
20	559	e=0	538	e=0	624	523	521	516	515	521	Portal vertex protein
67	–	–	–	–	80	78	76	69	55	53	Prohead core protein
68	–	–	–	–	141	141	135	140	163	157	Prohead core protein
21	217	e-68	215	e-77	212	213	231	210	213	209	Prohead protease
22	367	e-54	334	e-46	269	270	264	274	283	264	Prohead core protein
23	471	e-139	463	e-149	521	522	528	529	514	534	Head protein
inh	–	–	–	–	226	222	242	244	163	232	Head protease inhibitor
hoc	–	–	–	–	376	471	404	180	–	–	Highly immunogenic capsid protein
soc	–	–	–	–	80	78	–	–	–	–	Small outer capsid head protein
ip1	–	–	–	–	96	99	–	–	–	–	Ip1 internal head protein
Tail, tail fiber genes											
36	–	–	–	–	221	218	496	1420[b]	–	–	
37	1909		–		1026	1103	979	1420[b]	1085 / 1094	1358 / 1303	Distal long tail fiber
38	–	–	–	–	183	257	258	–	–	–	T2-like gp38 adhesin
3	178	e-8	184	e-5	176	194	196	175	177	189	Tail sheath stabilizer
27	–	–	–	–	391	390	369	321	–	–	Baseplate hub subunit
28	–	–	–	–	177	156	172	175	–	–	Baseplate hub assembly
29	–	–	–	–	590	590	577	566	–	–	Baseplate hub subunit
51	–	–	–	–	249	251	256	248	–	–	Baseplate hub assembly
54	–	–	–	–	320	320	310	285	–	331	Baseplate subunit
26	240	e-13	237	e-13	208	208	209	204	282	258	Baseplate hub

Table 4 continued

Gene	*P-SSM2 (e-value)*	*e-value*	*P-SSM4*	*e-value*	*T4*	*RB69*	*RB49*	*44RR2.8t*	*KVP40*	*Aeh1*	*Description of protein product*
48	387		–	–	364	369	352	342	379	358	Baseplate tube cap
53	242	e-4	324	e=0.003	196	191	184	179	192	188	Baseplate wedge
5	753	e=0.025	770	e-8	575	577	600	600	421	604	Baseplate hub
25	134	e-8	140	e-8	132	132	128	127	139	140	Baseplate wedge
6	648	e-46	663	e-34	660	656	634	627	646	651	Baseplate wedge
7	5196		7313	e=0.005	1032	1032	3087	1019	1165	1163	Baseplate wedge
8	534	e-6	511	e-6	334	334	331	328	343	328	Baseplate wedge
9	–	–	–	–	288	287	284	285	327	308	Baseplate; socket
10	–	–	410	e-4	602	604	600	604	748	718	Baseplate; pin
11	–	–	–	–	219	219	214	220	234	324	Baseplate; pin
12	–	–	–	–	527	516	466	466	512 / 473	436	Short tail fiber
wac	–	–	–	–	487	480	589	587	559	1035	Whiskers
15	281	e-26	311	e-23	272	258	277	272	450	275	Tail sheath stabilizer
18	730	e-95	750	e-101	659	660	666	663	671	679	Tail sheath
19	196	e-32	197	e-35	163	163	164	162	166	162	Tail tube
35	–	–	–	–	372	374	379	377	894	1312	Fiber hinge
34	–	–	–	–	1289	1277	1246	1222	1290	1236	Proximal long tail fiber
Chaperonins, catalysis genes											
rnlA	–	–	–	–	374	374	389	383	381	389	Fiber attachment
31	–	–	–	–	111	110	107	136	112	136	Cochaperonin for head assembly
57A	–	–	–	–	80	76	86	78	89 ?	77	Chaperone for long and short tail fibers
hsp20	153	e-15	150	e-22	–	–	–	–	–	–	hsp20 family chaperone proteins
Lysis exclusion											
rI	–	–	–	–	97	90	100	94	–	–	Acts with *t* holin for lysis timing
rIIA	–	–	–	–	725	737	702	705	689	732	Lysis inhibition
rIIB	–	–	–	–	312	311	330	377	345	438	Lysis inhibition
rIII	–	–	–	–	82	82	–	–	–	94	rapid lysis phenotype
segD	–	–	–	–	223	–	–	–	231	–	Endonuclease
e	–	–	–	–	164	157	–	164	–	165	intracellular N-acetyl muramidase; lysis from within
t	–	–	–	–	218	223	218	221	–	–	holing lysis mediator; pore protein for T4 endolysozyme
	–	–	158	e-33	–	–	–	–	–	–	hemaglutinin neuraminidase
Others											
57B	–	–	–	–	152	151	154	142	151	164	Unknown function, conserved

Table 4 continued

Gene	*P-SSM2 (e-value)*	*e-value*	*P-SSM4*	*e-value*	*T4*	*RB69*	*RB49*	*44RR2.8t*	*KVP40*	*Aeh1*	*Description of protein product*
mrh	–	–	–	–	161	69	–	–	–	–	allows growth on *rpoH* mutants
srd	–	–	–	–	248	248	251	–	–	–	decoy of sigma factors rpoS,
motB	–	–	–	–	162	151	–	–	–	–	unknown function
pin	–	–	–	–	161	137	–	–	–	–	stabilizes proteins during T4 infection
dmd	–	–	–	–	60	63	67	–	–	–	mutants show destabilization of T4 mRNAs
imm	–	–	–	–	84	82	–	–	–	–	alters membrane so superinfecting phages are left in periplasm to be degraded
sp	–	–	–	–	97	97	–	–	–	–	mutants give rapid lysis
nudE	–	–	–	–	151	152	–	352	–	183 / 351	nudix hydrolase activity
psbA	361	e=0	366	e=0	–	–	–	–	–	–	D1 protein of photosystem II
psbD	–	–	359	e=0	–	–	–	–	–	–	D2 protein of photosystem II
hli	x6 *hli*		x4 *hli*		–	–	–	–	–	–	High-light inducible protein
petE	115	e-21	–	–	–	–	–	–	–	–	Plastocyanin; photosynthetic electron transport
petF	98	e-29	–	–	–	–	–	–	–	–	Ferredoxin; photosynthetic electron transport
pcyA	–	–	230	e-26	–	–	–	–	–	–	Ferredoxin:phycocyanobilin oxidoreductase
pebA	234	e-12	–	–	–	–	–	–	–	–	Phycoerythrobilin : Ferredoxin oxidoreductase
ho1	234	e-63	–	–	–	–	–	–	–	–	Catalyzes conversion of heme to bilivderin IX
nifU-like	97	e-14	–	–	–	–	–	–	–	–	putative role in biosynthesis of Fe-S clusters
speD	102	e-17	–	–	–	–	–	–	–	–	S-adenosylmethionine decarboxylase
pstS	322	e-136	322	e-137	–	–	–	–	–	–	Periplasmic phosphate binding protein
phoH	251	e-24	258	e-20	–	–	–	–	235	–	Phoshate inducible protein, putative ATPase
tal	216	e-47	218	e-53	–	–	–	–	–	–	Transaldolase
prnA	486	e-52	–	–	–	–	–	–	–	–	Bacterial tryptophan halogenase
nol	601	e-54	–	–	–	–	–	–	–	–	Carbomyltransferase / nodulation protein
cobS	365	e-21	365	e-23	–	–	–	–	–	–	Cobalamin biosynthesis
mazG	139	e-11	134	e-27	–	–	–	–	–	–	putative pyrophosphohydrolase / pyrophosphatase
	–	–	97	e-29	–	–	–	–	–	–	Carboxylesterase
LPS cluster	x35 LPS genes	–	–	–	–	–	–	–	–	–	Epimerases, transferases, phospholipases

[a] gene is split into two segments, often by an intron or homing endonuclease
[b] genes are fused together

ACKNOWLEDGEMENTS

This work is the result of significant collaboration with the following eventual co-authors on the comparative genome manuscript. The clone libraries were constructed by David Mead (Lucigen) and Chris Detter (DOE JGI). The inserts were shotgun sequenced by Jan-Feng Chang and others at the DOE JGI. Assembly was done by Jan-Fang Chang (DOE JGI) and Jane Grimwood (Stanford) and finishing reactions were performed by Jane Grimwood. Peter Weigele and I have struggled with defining tail fiber and other structural genes that are likely part of the phage, but are non-homologous to known sequences. Peter also provided the TEM micrographs for Figure 1. Maureen Coleman has written scripts to make better reference files for compiling tables 3 and 4, as well as prepared the genome figures using DNA Master (provided by Dr. Jeffrey Lawrence). Forest Rohwer trained and guided me from day one on this project and Penny Chisholm offered support at all stages of the project. Virginia Rich, Debbie Lindell and Erik Zinser provided valuable comments on the written chapter.

REFERENCES

Bergh, O. (1989). High abundance of viruses found in aquatic environments. *Nature* **340,** 467-468.

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**(12), 2607-18.

Bograh, A., Gingras, Y., Tajmir-Riahi, H. A., and Carpentier, R. (1997). The effects of spermine and spermidine on the structure of photosystem II proteins in relation to inhibition of electron transport. *FEBS Lett* **402**(1), 41-4.

Bratbak, G., Heldal, M., Norland, S., and Thingstad, T. F. (1990). Viruses as partners in spring bloom microbial trophodynamics. *Appl Environ Microbiol* **56,** 1400-1405.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**(22), 14250-5.

Brussow, H. (2001). Phages of Dairy Bacteria. *Annu Rev Microbiol* **55,** 283-303.

Brussow, H., and Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell* **108**(1), 13-6.

Calendar, R. (1988). "The Bacteriophages." (R. Calendar, Ed.), I Plenum, New York.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev* **67**(2), 238-76.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* **49**(2), 277-300.

Chen, F., and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage p60: a new insight on lytic and lysogenic phages. *Appl Environ Microbiol* **68**(5), 2589-94.

Demerec, M., and Fano, U. (1945). Bacteriophage-resistant mutants in *Escherichia coli*. *Genetics* **30,** 119.

Desplats, C., and Krisch, H. M. (2003). The diversity and evolution of the T4-type bacteriophages. *Res Microbiol* **154,** 259-267.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I., and Hess, W. R. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**(17), 10020-5.

Forterre, P. (1999). Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Microbiol* **33**(3), 457-65.

Frankenberg, N., and Lagarias, J. C. (2003). Phycocyanobilin:ferredoxin oxidoreductase of Anabaena sp. PCC 7120. Biochemical and spectroscopic. *J Biol Chem* **278**(11), 9219-26.

Frankenberg, N., Mukougawa, K., Kohchi, T., and Lagarias, J. C. (2001). Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *Plant Cell* **13**(4), 965-78.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548.

Gleason, F. K., and Olszewski, N. E. (2002). Isolation of the gene for the B12-dependent ribonucleotide reductase from Anabaena sp. strain PCC 7120 and expression in Escherichia coli. *J Bacteriol* **184**(23), 6544-50.

Hammer, P. E., Burd, W., Hill, D. S., Ligon, J. M., and van Pee, K. (1999). Conservation of the pyrrolnitrin biosynthetic gene cluster among six pyrrolnitrin-producing strains. *FEMS Microbiol Lett* **180**(1), 39-44.

Hammer, P. E., Hill, D. S., Lam, S. T., Van Pee, K. H., and Ligon, J. M. (1997). Four genes from Pseudomonas fluorescens that encode the biosynthesis of pyrrolnitrin. *Appl Environ Microbiol* **63**(6), 2147-54.

Hardies, S. C., Comeau, A. M., Serwer, P., and Suttle, C. A. (2003). The complete sequence of marine bacteriophage VpV262 infecting vibrio parahaemolyticus indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. *Virology* **310**(2), 359-71.

Hausmann, R. (1988). The T7 Group. *In* "The Bacteriophage" (R. Calendar, Ed.), pp. 259-289. I Plenum, New York.

Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor Popul Biol* **61**(4), 471-80.

Hendrix, R. W. (2003). Bacteriophage genomics. *Curr Opin Microbiol* **6**(5), 506-11.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends in Microbiology* **8**, 504-508.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* **96**(5), 2192-7.

Johnson, Z. I., Landry, M. L., Bidigare, R. R., Brown, S. L., Campbell, L., Gunderson, J., Marra, J., and Trees, C. (1999). Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. *Deep Sea Research II* **46**, 1719-1743.

Karl, D. M. (1999). A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems* **2**, 181-214.

Kazakov, A. E., Vassieva, O., Gelfand, M. S., Osterman, A., and Overbeek, R. (2003). Bioinformatics classification and functional analysis of PhoH homologs. *In Silico Biol* **3**(1-2), 3-15.

Kim, S. K., Makino, K., Amemura, M., Shinagawa, H., and Nakata, A. (1993). Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli. *J Bacteriol* **175**(5), 1316-24.

Kirner, S., Hammer, P. E., Hill, D. S., Altmann, A., Fischer, I., Weislo, L. J., Lanahan, M., van Pee, K. H., and Ligon, J. M. (1998). Functions encoded by pyrrolnitrin biosynthetic genes from Pseudomonas fluorescens. *J Bacteriol* **180**(7), 1939-43.

Kovalyova, I. V., and Kropinski, A. M. (2003). The complete genomic sequence of lytic bacteriophage gh-1 infecting Pseudomonas putida--evidence for close relationship to the T7 group. *Virology* **311**(2), 305-15.

Kruger, D. H., and Schroeder, C. (1981). Bacteriophage T3 and bacteriophage T7 virus-host cell interactions. *Microbiol Rev* **45**(1), 9-51.

Kunisawa, T. (1992). Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and its host *Escherichia coli*. *J. Theor. Biol.* **159**, 287-298.

Kutter, E., Guttman, B., and Carlson, K. (1994). The transition from host to phage metabolism after T4 infection. *In* "Molecular biology of bacteriophage T4" (J. Karam, J. W. Drake, K. N. Kreuzer, G. Mosig, D. H. Hall, F. A. Eiserling, L. W. Black, E. K. Spicer, E. Kutter, K. Carlson, and E. S. Miller, Eds.), pp. 343-346. American Society for Microbiology, Washington DC.

Lawrence, J. G., Hatfull, G. F., and Hendrix, R. W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J Bacteriol* **184**(17), 4891-905.

Lawrence, J. G., and Roth, J. R. (1995). The cobalamin (coenzyme B12) biosynthetic genes of Escherichia coli. *J Bacteriol* **177**(22), 6371-80.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A., Rohwer, F., and Chishiolm, S. W. (submitted). Transfer of photosynthesis genes to and from *Prochlorococcus* phage.

Madigan, M. T., Martinko, J. M., and Parker, J. (2003). "Brock Biology of Microorganisms." 10 ed. Prentice Hall, Upper Saddle River.

Maggio-Hall, L. A., and Escalante-Semerena, J. C. (1999). In vitro synthesis of the nucleotide loop of cobalamin by Salmonella typhimurium enzymes. *Proc Natl Acad Sci U S A* **96**(21), 11798-803.

Mann, N. H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev* **27**(1), 17-34.

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature* **424**, 741.

McDaniel, L., Houchin, L. A., Williamson, S. J., and Paul, J. H. (2002). Lysogeny in marine *Synechococcus*. *Nature* **415**, 496.

Millard, A., Clokie, M., Shub, D. A., and Mann, N. H. (submitted). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains, submitted.

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., Lee, J., Szczypinski, B., and Fraser, C. M. (2003a). Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185**(17), 5220-33.

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003b). Bacteriophage T4 genome. *Microb. Mol. Biol. Rev.* **67**, 86-156.

Molineux, I. (in press). The T7 group. *In* "The Bacteriophages" (R. Calendar, Ed.).

Nishio, K., and Nakai, M. (2000). Transfer of iron-sulfur cluster from NifU to apoferredoxin. *J Biol Chem* **275**(30), 22615-8.

Nunes-Duby, S. E., Kwon, H. J., Tirumalai, R. S., Ellenberger, T., and Landy, A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* **26**(2), 391-406.

Ortmann, A. C., Lawrence, J. E., and Suttle, C. A. (2002). Lysogeny and Lytic Viral Production during a Bloom of the Cyanobacterium *Synechococcus* spp. *Microb. Ecol.* **in press**.

Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E. A., Partensky, F., and Larimer, F. W. (2003). The genome of a motile marine *Synechococcus*. *Nature* **424,** 1037-1041.

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

Paul, J. H., Sullivan, M. B., Segall, A. M., and Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**(4), 463-76.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Hendrix, R. W., and Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**(2), 171-82.

Proctor, L. M., and Fuhrman, J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature* **343,** 60-62.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., 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., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042-1047.

Rohwer, F., and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**(16), 4529-35.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45,** 408-418.

Saito, M. A., and Moffett, J. W. (2002). Temporal and spatial variability of cobalt in the Atlantic ocean. *Geochemica et Cosmochimica Acta* **66,** 1943-53.

Saito, M. A., Moffett, J. W., Chisholm, S. W., and Waterbury, J. B. (2002). Cobalt limitation and uptake in *Prochlorococcus*. *Limnol. Oceanogr.* **47,** 1629-1636.

Scanlan, D. J., Silman, N. J., Donald, K. M., Wilson, W. H., Carr, N. G., Joint, I., and Mann, N. H. (1997). An immunological approach to detect phosphate stress in populations and single cells of photosynthetic picoplankton. *Appl. Env. Microbiol.* **63**(6), 2411-2420.

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

Schmetter, G. (1994). *In* "The molecular biology of cyanobacteria" (D. A. Bryant, Ed.), pp. 409-435. Kluwer Academic Publishers, Dordrecht.

Sherman, L. A. (1976). Infection of *Synechococcus cedrorum* by the cyanophage AS-1M. III. Cellular metabolism and phage development. *Virology* **71,** 199-206.

Simon, M. N., Davis, R. W., and Davidson, N. (1971). The Bacteriophage Lambda (A. D. e. Hershey, Ed.), pp. 313-328. Cold Spring Harbor Laboratory Press, Plainview, NY.

Sprenger, G. A. (1995). Genetics of pentose-phosphate pathway enzymes of Escherichia coli K-12. *Arch Microbiol* **164**(5), 324-30.

Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W. (2003). Cyanophage infecting the oceanic cyanobacterium, *Prochlorococcus*. *Nature* **424,** 1047-1051.

Suttle, C. A., and Chan, A. M. (1993). Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92,** 99-109.

Tetart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H. W., and Krisch, H. M. (2001). Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J Bacteriol* **183**(1), 358-66.

Torriani, A. (1990). From cell membrane to nucleotides: the phosphate regulon in Escherichia coli. *Bioessays* **12**(8), 371-6.

van Regenmortel, M. H. V., Fauquet, C. M., Bishop, D. H. L., Carstens, E. B., Estes, M. K., Lemon, S. M., Maniloff, J., Mayo, M. A., McGeoch, D. J., Pringle, C. R., and Wickner, R. B. (2000). "Virus taxonomy: The Classification and Nomenclature of Viruses." Academic Press, San Deigo.

Wanner, B. L. (1996). Phosphorus assimilation and control of the phosphate regulon. 2nd ed. *In "Escherichia coli* and *Salmonella:* Cellular and molecular biology" (F. C. Neidhardt, Ed.), pp. 1357-1381. ASM Press, Washington DC.

Waterbury, J. B., and Valois, F. W. (1993). Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**(10), 3393-3399.

Watt, W. B., and Dean, A. M. (2000). Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annu Rev Genet* **34,** 593-622.

Wikner, J., Vallino, J. J., Steward, G. F., Smith, D. C., and Azam, F. (1993). Nucleic acids from the host bacterium as a major source of nucleotides for three marine bacteriophages. *FEMS Microb Ecol* **12,** 237-248.

Wu, J., Sunda, W., Boyle, E. A., and Karl, D. M. (2000). Phosphate depletion in the Western North Atlantic Ocean. *Science* **289,** 759-762.

Yuzenkova, J., Nechaev, S., Berlin, J., Rogulja, D., Kuznedelov, K., Inman, R., Mushegian, A., and Severinov, K. (2003). Genome of Xanthomonas oryzae bacteriophage Xp10: an odd T-odd phage. *J Mol Biol* **330**(4), 735-48.

Zhang, J., and Inouye, M. (2002). MazG, a nucleoside triphosphate pyrophosphohydrolase, interacts with Era, an essential GTPase in Escherichia coli. *J Bacteriol* **184**(19), 5323-9.

Zhang, J., Zhang, Y., and Inouye, M. (2003). Thermotoga maritima MazG protein has both nucleoside triphosphate pyrophosphohydrolase and pyrophosphatase activities. *J Biol Chem* **278**(24), 21408-14.

CHAPTER VI

SUMMARY AND FUTURE DIRECTIONS

CHAPTER VI. SUMMARY AND FUTURE DIRECTIONS

SUMMARY

This thesis describes the first isolation of *Prochlorococcus* cyanophage and the establishment of a cyanophage culture collection, isolated using a phylogenetically diverse selection of *Prochlorococcus* and *Synechococcus* hosts (Sullivan, Waterbury, and Chisholm, 2003)(Chapter 2). This phage collection contains all three phage morphological families previously described for *Synechococcus* (Lu, Chen, and Hodson, 2001; Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993), and the breadth of the collection revealed a relationship between host of isolation and the resulting phage morphology. Specifically, HL *Prochlorococcus* strains yielded exclusively *Podoviridae* while LL *Prochlorococcus* and *Synechococcus* yielded primarily *Myoviridae*. Host-range analyses of 21 host strains and 45 cyanophage isolates demonstrated varying levels of specificity between phage morphological types; *Podoviridae* commonly infected a single strain, while *Myoviridae* commonly infected across ecotypes and even across the two cyanobacterial genera (Sullivan, Waterbury, and Chisholm, 2003). While such trends are difficult to interpret, they may reflect variable host properties, such as differences in cell surface characteristics and/or variability in anti-phage defense mechanisms (e.g., R-M systems, blocking proteins), as well as differences in phage properties such as tail fiber switching and/or copy number (see below). For example, host range may be controlled by the distribution among host ecotypes of cell surface molecules that act as phage receptors used to gain entry to their hosts. Using this logic, one possible explanation for the observed host range patterns is that that *Podoviridae* use the highly variable o-antigen of the lipopolysaccharide (LPS) as a receptor, while *Myoviridae* use the less variable core oligosaccharides of the LPS or outer membrane proteins (OMPs), thus allowing them to infect a broader range of hosts. Alternatively, though not mutually exclusive, the tail fibers of the phage that determine the host receptor a phage binds (Henning and Hashemolhosseini, 1994) may not recognize different receptor types. Instead, host range might be determined by swapping the specificity determinant from one phage to another (termed tail fiber switching) (Tetart, Desplats, and Krisch, 1998; Tetart et al., 1996) or by a given phage genome encoding multiple tail fiber genes (Miller et al., 2003). The mechanisms of host specificity remain one of the most pressing unknowns in marine phage-host ecology.

Next, we focused on one group of viruses in the collection, the *Myoviridae,* to compare the captured phage diversity in our collection with that found in other studies focusing on this

family. Diversity within the *Myoviridae* has been previously characterized using the g20 gene that encodes the portal protein (Sullivan et al.)(Chapter 3). g20 sequences from our phage collection grouped with those of previously cultured isolates, but not with environmental g20 sequences which lacked g20 sequences from cultured representatives. This suggests that the g20 sequence diversity in all current myophage isolates, including those of our collection, lack some of the g20 diversity observed in the environment. In addition, methodological complications of our work indicated that the previous studies of environmental g20 diversity may also have missed myophage groups. g20 PCR screening of our collection was hindered by the fact that the published g20 primers could not amplify sequences from all myophage in the collection. Therefore, we redesigned the existing primers to provide a new primer set that could successfully amplify g20 sequence from all myophage in the collection. Once able to adequately map the myophage g20 diversity of the isolates, we found that this index of diversity showed little relationship to various phage properties, such as host range and geographic isolation. Together this work suggests, not only have previous environmental g20 diversity studies likely missed g20 diversity from many myophage, but also suggests that the g20 gene may not be the most appropriate marker for tracking phages specific for particular hosts (e.g., in studies of coevolution of phage and host). Future research might focus on using the tial fiber genes as a more ecologically informative diversity proxy for this phage family, since tail fiber genes are known to be the host range determinant in these phages (Henning and Hashemolhosseini, 1994).

To examine the abundance and diversity of cyanophage in the field using culture-based assays, a broad selection of host strains were used to quantify lytic cyanophage titers along a coastal-to-open ocean transect. These assays demonstrated that cyanophage titers were lower in open ocean environments than coastal environments (Sullivan, Waterbury, and Chisholm, 2003)(Chapter 2). Because of the complexity of the phage-host interaction, definitve explanations for such trends remain elusive. However, there are some factors which could be involved. For example, if host strain microdiversity increased along the transect, and cross-infection ability did not increase concurrently, this would lead to lower phage titers (Thingstad, 2000). Another possible explanation is decreased nutrient availability in the oligotrophic oceans (Cavender-Bares, Karl, and Chisholm, 2001) resulting in sub-optimal growth of host cells in the Sargasso Sea (Mann et al., 2002) relative to *Synechococcus* at the coastal site (Waterbury et al., 1986). Viral production is correlated with host growth rates in chemostats (Bohannan and Lenski, 2000) and in the field (Steward, Smith, and Azam, 1996). Such a correlation might result

from nutrient limitation that causes physiological changes in the host that stall the lytic process of obligately lytic phage (Stent, 1963), or favours lysogeny in temperate phage (Rohwer et al., 2000; Suttle, 2000; Wilson, Carr, and Mann, 1996). Because we currently lack the ability to specifically quantify phages that infect particular hosts, it is not yet possible to clarify the mechanistic underpinnings explaining such observations.

To further explore the potential interactions between phage and host during cyanophage infection as inferred from genome sequences, we analyzed genome sequences provided through collaborations with the Department of Energy Joint Genome Institute of three hosts (Palenik et al., 2003; Rocap et al., 2003)(Appendix B) and three *Prochlorococcus* cyanophages (Chapter 4, 5). Analysis of three cyanobacterial genomes (*Prochlorococcus* MED4, MIT9313 and *Synechococcus* WH8102) revealed that, in spite of the fact that prophage are common to prokaryotic hosts (Canchaya et al., 2003; Casjens, 2003) and that field data are suggestive that prophage are inducible from field *Synechococcus* populations (McDaniel et al., 2002; Ortmann, Lawrence, and Suttle, 2002), these three genomes do not contain intact prophage. However, it is likely that such prophage, if they existed, may have been induced from our cultured host cells due to the stressful conditions encountered during 10+ years of culturing. While we cannot rule out the possibility that the opportunity for reintegration of these prophages exists, it is also plausible that induced phages could have become non-infective through adsorption to non-productive hosts or host particles in cell cultures left in prolonged stationary phase. Two lines of evidence from genomic sequencing suggest that prophage may have once existed in these genomes. First, the occurrence of many intact and fragmented site-specific integrase genes found in all three genomes (Appendix B) strongly suggests that prophage may have once existed in these genomes (Palenik et al., 2003). Further, the podovirus P-SSP7 contains an integrase gene (Chapter 5) which suggests that, if this integrase gene is functional, then temperate phage do occur among the marine cyanobacteria.

Analyses of the three *Prochlorococcus* cyanophage genomes (Chapter 4, 5) suggest these phages are related to well-characterized groups, the T7-like (for phage P-SSP7) and T4-like (for phages P-SSM2, P-SSM4) phages, but are optimized for infection of photosynthetic hosts, and specifically those which thrive in oligotrophic environments (Sullivan and al.). All three of these genomes contain genes that encode core photosynthetic proteins that are full-length, conserved and clustered in the genome (Lindell et al., submitted)(Chapter 4). We hypothesize that such conservation of sequence results from selection, and thus that these genes are expressed and

functional during infection and offer an advantage to the phage and/or host that has fixed these genes in the phage genomes. In addition, both *Myoviridae* P-SSM2 and P-SSM4 genomes contain conserved, full-length proteins involved in phosphate stress, the mobilization of carbon storesand the synthesis of cobalamin, purines, LPS and several accessory pigments. Together, these genomic observations suggest the unique ways that cyanophage may alter, enhance and co-opt the metabolic capabilities of their hosts. In addition, the integrase gene found in the *Podoviridae* P-SSP7 genome suggests this phage is capable of integrating into its host; if further experiments provide direct evidence for thes, then it will represent the first temperate T7-like phage and the first temperate marine cyanophage in culture.

FUTURE DIRECTIONS

In order to more broadly understand the ecology of cyanophage and their hosts, it will be important to pinpoint the mechanisms controlling their interactions. In fact, the need for mechanistic understanding of microscale interactions is important for developing our understanding of all microbial processes fundamental to the global biogeochemistry of the oceans (Azam and Worden, 2004). As mechanistic information is then incorporated into models of phage-host population dynamics, and model outputs are compared to real-world data, such as host range information and transect titers presented in this thesis, the validity of the model will be tested, the parameters refined and remaining holes in our understanding will be highlighted. In this way, genomics, genetics, biochemistry, field studies and computer modeling will all play important roles in furthering our understanding of marine cyanophage ecology.

Efficiency of cross-infection

The host range analysis conducted in this thesis offers a preliminary look at trends in phage-host cross infection (Chapter 2). However, more information about each individual host-phage interaction is necessary before adequate modeling efforts can be made. I have provided binary data (either a phage infects or does not infect a host strain of interest) for the entire matrix of the host range table (Chapter 2). To adequately model this system, we will need more details about the interactions that are occurring, including quantifying the *efficiency* of infection among different phage-host pairings. Do the broad host-ranges of the *Myoviridae* come at the expense of rapid and efficient infection of any single strain relative to other *Myoviridae*? Do the more specific host-ranges of the *Podoviridae* correspond to a higher infection efficiency? Thus

revisiting the host range table and repeating the observations, this time recording the time from infection to lysis, will be important for refining our understanding of phage-host dynamics.Such an experiment can now be done in a high-throughput manner using the 96 well plate format and the fluorescence plate reader. In this way, every phage-host combination can be tested to allow quantitative documentation of the timing, and extent, of host cell lysis. Clearly, controlled experiments using phage and host at similar concentrations is critical to accurate comparisons. Together, these data would better describe the the efficiency with which a given cyanophage infects a suite of host strains.

Physiological characterization of phage isolates

Only one marine cyanophage has been physiologically characterized to obtain important parameters necessary for modeling the significance of phage-host interactions (Suttle and Chan, 1993). Although it has been shown that the majority of phage that infect *Synechococcus* are of the family *Myoviridae* (Suttle and Chan, 1993; Suttle and Chan, 1994; Waterbury and Valois, 1993), the only cyanophage that has been characterized, the cyanophage S-BBS1, belongs to the family *Siphoviridae*, and is no longer in culture (Suttle and Chan, 1993). The average adsorption rate constant over a 60 minute incubation was 0.035 min^{-1} which is slow relative to those typically seen for other bacteriophage, but within the range reported for cyanophage infecting freshwater cyanobacteria (Amla, 1981; Cseke and Farkas, 1979; Samimi and Drews, 1978). A one-step growth curve experiment indicated that cell lysis occurred between 9 and 17 hours following infection, while the burst size was estimated to be 250 phage per infected cell (Suttle and Chan, 1993). The burst size is similar to those previously reported for *Siphoviridae* that infect freshwater cyanobacteria (Martin, Leach, and Kuo, 1978), but is considerably larger than burst sizes estimated for the *Myoviridae* (Safferman et al., 1972; Sherman and Connelley, 1976), the phage morphology most commonly infecting *Synechococcus* (Suttle and Chan, 1993; Suttle and Chan, 1994; Waterbury and Valois, 1993).

Such values are important for understanding and modeling phage-host interactions in the environment. For example, the length of the lytic cycle and the burst size produced during a phage life cycle are critical parameters for calculating mortality and phage production. What are the values of these parameters for the phages in the MIT cyanophage collection? How might they differ between HL and LL *Prochlorococcus* hosts? Between *Podoviridae, Myoviridae* and *Siphoviridae*? These types of experiments are not trivial in a system where the host cells double

at best once per day, but these data would provide valuable information about adsorbtion, the length of the latent period and burst size. Some work on this front is already underway in the Suttle lab with *Synechococcus* cyanophage (C.A.Suttle, pers. comm.) and it would be beneficial to confer with their group before beginning these studies with the MIT collection.

Towards a tool for genetic manipulations

The vast number of phages and phage lysates in our collection might include a phage(s) that could transduce (move) genes between their hosts. Such transducing phages are a vital tool for genetically altering host strains to produce mutants suitable for important hypothesis-testing, allowing a critical genetic investigation not only of many host properties but also of the cellular components potentially mediating phage-host interactions. One way to identify such transducing phage from the collection is to screen them for the ability to transfer a selectable marker gene between host strains. However, this approach requires the creation or identification of a selectable marker gene in the host. Although there have not been selectable markers defined for *Prochlorococcus* previously, a streptomycin resistant mutant (Str-R) has recently been created in the MED4 strain (E. Zinser, unpubl. data) which will hopefully provide an appropriate selectable marker. While the mechanism of resistance is not yet determined in *Prochlorococcus,* in *E. coli* streptomycin resistance occurs spontaneously at a frequency of about 10^{-9} to 10^{-10} per generation and is often due to a single point mutation in the ribosomal protein small subunit, *rpsL* (E. Zinser, pers. comm.).

Given a Str-R marker, a screen for transducing phages could be conducted as follows. The mutant Str-R strains would be infected with a phage isolate, the phage produced from this infection would be harvested and used to infect a streptomycin-susceptible strain which could then be subjected to selection using streptomycin. In known transducing phage, the series of molecular steps involved in the transfer of a gene has been described. It begins with the packaging of a portion of the first host's DNA, encompassing the selectable marker gene, into a phage capsid during assembly (Calendar, 1988; Stent, 1963). This mispackaged DNA can result in a phage particle that is still capable of infection but not of lysis (i.e., the required phage DNA has been replaced with host DNA). If this phage successfully infects a host cells, then the introduced DNA has an opportunity to undergo genetic exchage with the new host genome via homologous recombination. Thus in our proposed assay, to screen for such recombination events, the remaining cells that did not lyse would be tested for growth against streptomycin

media to show Str-R. PCR, cloning and sequencing could be used to confirm the sequence of the genetic marker in the putative streptomycin resistant strain was due to the integration of the donor DNA and not a spontaneous mutation. In parallel, it is critical to quantify the rate of transduction to be certain that this rate was greater than the rate at which spontaneous mutation to streptomycin resistance occurs in cells infected with the same phage isolate using aliquots that have not passed through the Str-R host first. Calculation of the transduction rate is also crucial for planning experiments for subsequent genetic work. I note that, in the case of a temperate phage, it is possible that packaging of a small portion of the host genome can occur (Calendar, 1988). Because the mechanism of packaging host DNA in a temperate phage is commonly by imprecise excision of the prophage genome from the host genome, this mispackaged host DNA would be limited to DNA that is adjacent to the insertion site of the phage. Thus, unless the temperate phage inserted randomly into the host genome (as is the case for phage Mu), it is improbable that such an insertion site might be near the gene causing streptomycin resistance. However, to be certain that streptomycin resistance in the recipient strain were due to transduction rather than prophage insertion, one might confirm that no phage DNA were integrated into the host genome. This might be best achieved through a southern hybridization with phage genomic DNA against DNA extracted from host cells that were washed to remove adsorbed phages.

Towards a direct measure for quantifying viruses specific for a particular host

In spite of the fact that viruses are known to be abundant, dynamic members of the microbial community (Bergh, 1989; Bratbak, 1990; Proctor and Fuhrman, 1990), it has remained difficult to enumerate the phage that are specific to a particular host. Currently used methods either indiscriminately stain viruses without any indication for the host strain infected by the virus (e.g., SYBR, Yo-Pro) (Hennes and Suttle, 1995; Noble and Fuhrman, 1998) or underestimate the number of viruses because they are limited to organisms amenable to culture and to phage that can be propogated under laboratory conditions (e.g., MPN, plaque). Hence there is a need to develop a culture-independent approach, for example the use of Q-PCR to quantify phage of particular types. This, however, still relies on the identification and validation of appropriate marker genes in the phage. We have examined the sequence of the gene encoding the portal protein, g20, from the *Myoviridae* isolates in our diverse collection of cyanophage (Sullivan, Waterbury, and Chisholm, 2003) because it is a commonly used genetic diversity marker

(Frederickson and Suttle, 2001; Marston and Sallee, 2003; Wilson et al., 1999; Wilson et al., 2000; Zhong et al., 2002). However, we found that the diversity of the portal protein gene sequences does not correlate to the original host strain used for isolation or to the host range of the cyanophage (Chapter 3)(Sullivan et al.). Thus, a new genetic marker should be developed that can be used in studies correlating phage gene sequence diversity with host ITS diversity.

The sequence and structure of the tail fiber protein determines the ability of a phage to bind to receptor molecules on the surface of a potential host cell (Henning and Hashemolhosseini, 1994). Do cyanophage tail fiber gene sequences correlate with the diversity of *Prochlorococcus* and *Synechococcus* ITS gene sequences? Because these genes are commonly modular with highly divergent and highly conserved regions (Hendrix, 2003), they are suitable for the design of PCR primers that could be group-specific. We are uniquely positioned to embark on such a study as we have (1) a diversity of cyanophage in culture, (2) experience with field based Q-PCR protocols and (3) three sequenced cyanophage genomes (a fourth is on the way). The growing number of available phage genomes will be useful for the design of degenerate PCR primers that we could use to screen our diverse cyanophage collection for tail fiber amplicons. In addition to our T7-like podovirus genome and two T4-like myovirus genomes, there are currently 6 T4-like genomes (RB49, RB69, Aeh1, 44RR2.8t, T4, KVP40) and 5 T7-like genomes (T3, T7, P60, SIO1, VpV262) that can be used in primer design as well.

The Chisholm Lab has already developed a Q-PCR assay based upon the ITS gene sequence to quantify the abundance of the six known phylogenetic lineages of *Prochlorococcus*. A complementary Q-PCR assay based upon the tail fiber gene sequences could allow concurrent quantification of cyanophage that infect particular hosts. The development of such a combined Q-PCR assay that is culture-independent and can specifically quantify both the phage specific to a given host and the host itself would open the door to experiments that require parallel tracking of such difficult-to-assay populations.

The use of such a tool would have many applications, including the following. Field work suggests that overall population mortality caused by cyanophage is relatively low for *Synechococcus* (Mann, 2003; Suttle and Chan, 1994; Waterbury and Valois, 1993), while cyanophage titers from this thesis suggest that total cyanobacterial population mortality due to lytic phages might be even lower in the open oceans (Sullivan, Waterbury, and Chisholm, 2003). However, the cyanobacteria are incredibly diverse and the cyanophage infecting them have a variety of host ranges which complicates simple interpretations and modeling efforts from

quantification of general, perhaps biologically uninformative, groupings of hosts and phages. Because we can now quantify ecotypes within *Prochlorococcus* and *Synechococcus* populations, other labs are trying to assess the impact of spiking cyanophage isolates into bottled natural cyanobacterial communities during 24-48 hour incubations (J. Fuhrman, pers. comm.). We know that *Myoviridae* and *Podoviridae* have strikingly different host ranges (Chapter 2), will they have correspondingly distinct effects on natural cyanobacterial populations? If experiments were done over the course of an annual cycle, is there a seasonal effect on the interactions of cyanophage with specific parts of the cyanobacterial population? Is this driven by seasonal cycling of the host populations?

Experimental determination of cyanopodophage promoters

Due to the extensive genomic similarity between the *Prochlorococcus* podovirus P-SSP7 and coliphage T7 (Chapter 5), we expect that podovirus P-SSP7 would have similarly regulated genomic transcription. In extensive experiments done in coliphage T7 (Dunn and Studier, 1983; Kruger and Schroeder, 1981; Molineux, in press), the class I genes of the phage (including the phage-encoded RNAP) are transcribed upon entering the cells by the host RNAP and have *E. coli* promoters, while the class II and III genes are transcribed using a phage-encoded RNA polymerase (RNAP) gene that recognize promoters specific to this polymerase. If these general attributes of T7 promoters hold true in our T7-like cyanophage, then we would expect class I genes to have cyanobacterial promoters, while class II and class III genes would have promoter consensus sequences specific for the cyanophage RNAP.

To test a possible interaction between a phage-encoded T7-type RNAP and phage-DNA promotors (or host RNAP and its associated promoters in the phage), one could produce the RNAP(s) as a purified recombinant enzyme and then perform biochemical assays *in vitro* as follows. Briefly, clone the phage and host RNAP genes into separate *E. coli* constructs, over-express these genes in *E. coli* to produce a large amount of the respective protein products, purify the proteins and use the respective RNAP proteins in a DNA binding assay. By fragmenting the phage DNA, it could be determined which regions of the phage DNA are bound by the respective RNAP genes. If the interaction between the phage and/or host RNAP and phage-encoded DNA were too weak, then it might be better to directly measure transcript initiation using a 5'-RACE method as used for *Prochlorococcus* MED4 (Vogel et al., 2003). T7 RNAP acts as a single protein not requiring any accessory proteins or regulatory factors (Dunn and Studier, 1983;

Kruger and Schroeder, 1981), so for a T7-type RNAP this should prove relatively easily. However, the host RNAP may prove quite difficult until regulation in cyanobacteria is better understood as the host RNAP likely requires several subunits plus regulatory factors.

Expression analysis during infection

By examining host and phage gene expression during infection, many questions raised in this thesis (particularly chapters 4 and 5) could be addressed. Gene expression can be analyzed either at the mRNA or protein level, using a variety of tools. The transcription of specific genes of interest can be examined with RT-PCR and transcript-specific primers, while the entire transcriptome (genome-wide transcription) can be examined through the construction and implementation of a microarray. Using unfinished genome sequence files from the early stages of these phage genome projects, all ORFs from the genomes of two *Prochlorococcus* phage, P-SSM4 and P-SSP7 are represented on Affymetrix microarray chips that were designed for the *Prochlorococcus* host (D. Lindell, A. Wright, pers. comm.). This will allow simultaneous examination of genome-wide gene expression in both the host and phage during infection and over a range of experimental conditions. Protein analyses in a cell line lacking a genetic system can be done using the techniques being developed by various groups within the GTL consortium that are using mass spectrometry in conjunction with genome data to determine which proteins are produced at any given time.

There are many questions that beg hypothesis-testing as a result of the genomic observations from this thesis (Chapters 4 and 5). Specifically, the Affymetrix microarrays can be used to examine the expression during host infection of those P-SSP7 and P-SSM4 genes relevant to photosynthesis (*psbA, hli*), carbon mobilization (*tal* gene), phosphate metabolism (*pstS, phoH* genes) and cobalt usage (*cobS* gene). In P-SSP7 (the T7-like virus) the expression of the integrase gene could be assayed specifically using RT-PCR or using the Affymetrix microarrays to better understand the possible development of lysogeny (as described in more detail below).

The role of lysogeny

The fact that the podovirus, P-SSP7, is a T7-like phage that includes an integrase gene (Chapter 5) begs the question of whether this integrase gene is functional and allows P-SSP7 to integrate into the host genome as a prophage. To date, there are no T7-like phages that contain integrase genes and there are no temperate marine cyanophage in culture. Experimental evidence

to show that P-SSP7 can integrate into the host genome *show* that this phage is temperate and capable of integrating into host DNA as a prophage. Using either RT-PCR or microarrays to follow the expression of the phage-encoded integrase gene will be critical to show whether and when this gene is expressed during infection, and how lysogeny could be affected by differing experimental conditions. If the integrase gene is expressed, it will be important to subsequently demonstrate that the phage DNA is actually integrated into the host genome. This could be done through southern hybridization with a phage genome probe against DNA extracted from washed host cells (as above). Alternatively, convincing support for prophage integration might be achieved by the use of a carefully designed PCR assay using the putative lysogen DNA as a template. In this assay, one primer could be specifically designed as an outward-facing primer at one end of the phage genome while a second primer could be random to obtain a variety of products. Subsequent sequencing could identify whether the product contained both phage and host DNA. Though more difficult than a southern, this assay would also provide information on where the prophage integration site is within the genome.

While the experimental steps required to confirm that the prophage is indeed integrated into the host genome seems relatively manageable, the difficult step will likely be in finding the conditions and/or host strain where the phage actually integrates into the host genome. As a starting point for finding such conditions, temperate phage classically integrate under conditions where the host cell is growing with limited nutrients, because it is under these conditions that the host is unable to supply the building blocks and energy required to produce progeny phage particles (Calendar, 1988).

Are 'host' genes prevalent among cyanophages infecting marine cyanobacteria?

The photosynthetic genes *psbA* and *hliP* are found in all three *Prochlorococcus* cyanophage genomes sequenced to date (Lindell et al., submitted), but not in all *Synechococcus* cyanophage (Mann et al., 2003; Millard et al., submitted). It would be interesting to assay for the presence of these genes and others (e.g., *psbD, pstS, phoH, talC*) in the other isolates of the MIT cyanophage collection. Since PCR can amplify DNA from very few template copies and because these genes are also found in the hosts, such a study would require appropriate controls. Phage and host DNA could be separated in one of two ways: (1) phage particles can be separated from host DNA by density centrifugation using cesium chloride gradients, or (2) host DNA be degraded with DNase while the phage DNA is still protected within the phage heads. The phage

particles (purified by either cesium chloride or DNase) could then be used as template for PCR reactions with primers designed to amplify the gene of interest. Prior efforts to amplify, clone and sequence the genes of interest from each of the putative host strains would provide a library of host gene sequences that could help determine whether a sequenced PCR product from a phage lysate was amplified from contaminating host DNA or packaged phage DNA. Finally, southern blots using probes for both a phage structural gene and a potential "host" gene of interest could confirm that the given gene is indeed in the phage genome. Together, these two experiments would build a strong case that the host gene of interest was actually present in the phage genome.

The big picture and big obstacles

While the abundance of phages in the marine environment is well recognized (Bergh, 1989; Bratbak, 1990; Proctor and Fuhrman, 1990), their genetic and functional diversity is only beginning to be appreciated. Phages are likely influential on host microbial processes as reservoirs of genetic information that affects the evolution and physiology of their hosts (Lawrence, Hendrix, and Casjens, 2001; Lindell et al., submitted). The small size of phage genomes means that 100 phage genomes can be sequenced for the same cost as a large bacterial genome (Paul et al., 2002). Such value for investment means that phage genomic sequencing should yield genomic data that would provide insight toward understanding evolutionary and biogeochemical processes of the marine microbiological food web.

However, there are two significant barriers to be overcome. First, it is difficult, and in some cases prohibitive, to obtain concentrated phage stocks to provide sufficient material for genomic sequencing. Collected efforts with marine phage-host systems to optimize growth conditions that yield high phage titers and improve purification methods should prove invaluable to preparation of phages for sequencing. Second, experimental confirmation of genome-enabled hypotheses requires the development of well characterized phage-host systems that can be usefully probed in the laboratory. As evidenced in chapters 4 and 5, bioinformatics only gets us so far. The proof lies in experimental documentation.

In addition to learning about new functionality between phages and their hosts, it is important to quantify particular phage types in an ecologically meaningful way. Currently, we lack the ability to quantify phages that infect particular hosts without using biased culture-based methods. Thus, the development of unbiased culture-independent assays are essential to accurately assess phage roles in mortality and co-evolution in the environment. While a

significant database of portal protein gene sequences from myoviruses exists, it is unclear whether this taxonomic marker will indicate which hosts the phage can infect (Chapter 3). Alternatively, we suggest a new taxonomic marker using the distal tail fiber gene. The tail fiber protein determines host specificity of the phage (Henning and Hashemolhosseini, 1994), thus it seems a logical target for the development of such an assay. Due to the modular and mosaic nature of these genes, such work is likely to be quite complicated and require expertise in protein biology, bioinformatics and molecular biology. However, the availability of many phage genomes for the types of phages most commonly found in the marine environment (T4-like and T7-like phages), should provide sufficient starting material for rigorous identification of important conserved regions within these genes that could be used for targetted PCR primers to span divergent regions that could provide host range information.

The age of phage is upon us. I look forward to what we will learn.

REFERENCES

Amla, D. V. (1981). Isolation and characterization of a minute plaque forming mutant of cyanophage AS-1. *Biochem Physiol Pflanzen* **176,** 83-89.

Azam, F., and Worden, A. Z. (2004). Microbes, molecules and marine ecosystems. *Science* **303,** 1622-1624.

Bergh, O. (1989). High abundance of viruses found in aquatic environments. *Nature* **340,** 467-468.

Bohannan, B. J. M., and Lenski, R. E. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecol. Letters* **3,** 362-377.

Bratbak, G. (1990). Viruses as partners in spring bloom microbial trophodynamics. *Applied and Environmental Microbiology* **56,** 1400-1405.

Calendar, R. (1988). "The Bacteriophages." (R. Calendar, Ed.), I Plenum, New York.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev* **67**(2), 238-76.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* **49**(2), 277-300.

Cavender-Bares, K. K., Karl, D. M., and Chisholm, S. W. (2001). Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure, and comparison to patterns in the Pacific Ocean. *Deep-Sea Research, Part I* **48,** 2373-2395.

Cseke, C. S., and Farkas, G. L. (1979). Effects of light on the attachment of cyanophage AS-1 to *Anacystis nidulans*. *J. Bacteriol.* **137,** 667-669.

Dunn, J. J., and Studier, F. W. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166,** 477-535.

Frederickson, C. M., and Suttle, C. A. (2001). *Fourteenth Evergreen International Phage Meeting, Olympia, WA.*

Hendrix, R. W. (2003). Bacteriophage genomics. *Curr Opin Microbiol* **6**(5), 506-11.

Hennes, K. P., and Suttle, C. A. (1995). Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. *Limnol. Oceanogr.* **40,** 1050-1055.

Henning, U., and Hashemolhosseini, S. (1994). Recepor recognition by T-even-type coliphages. *In* "Molecular biology of bacteriophage T4" (J. Karam, Ed.), pp. 291-298. ASM Press, Washington D.C.

Kruger, D. H., and Schroeder, C. (1981). Bacteriophage T3 and bacteriophage T7 virus-host cell interactions. *Microbiol Rev* **45**(1), 9-51.

Lawrence, J. G., Hendrix, R. W., and Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**(11), 535-40.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A., Rohwer, F., and Chishiolm, S. W. (submitted). Transfer of photosynthesis genes to and from *Prochlorococcus* phage.

Lu, J., Chen, F., and Hodson, R. E. (2001). Distribution, isolation, host specificity, and diversity of cyanophages infecting marine Synechococcus spp. in river estuaries. *Appl Environ Microbiol* **67**(7), 3285-90.

Mann, E. L., Ahlgren, N., Moffett, J. W., and Chisholm, S. W. (2002). Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47,** 976-988.

Mann, N. H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev* **27**(1), 17-34.

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature* **424,** 741.

Marston, M. F., and Sallee, J. L. (2003). Genetic diversity and temporal variation in the cyanophage community infecting marine Synechococcus species in Rhode Island's coastal waters. *Appl Environ Microbiol* **69**(8), 4639-47.

Martin, E. L., Leach, J. E., and Kuo, K. J. (1978). Biological regulation of bloom-causing blue-green algae. *In* "Microbial Ecology" (M. W. Loutit, and J. A. R. Miles, Eds.), pp. 62-67. Springer Verlag, Berlin.

McDaniel, L., Houchin, L. A., Williamson, S. J., and Paul, J. H. (2002). Lysogeny in marine *Synechococcus*. *Nature* **415,** 496.

Millard, A., Clokie, M., Shub, D. A., and Mann, N. H. (submitted). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains, submitted.

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., Lee, J., Szczypinski, B., and Fraser, C. M. (2003). Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185**(17), 5220-33.

Molineux, I. (in press). The T7 group. *In* "The Bacteriophages" (R. Calendar, Ed.).

Noble, R. T., and Fuhrman, J. A. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aq. Microb. Ecol.* **14,** 113-118.

Ortmann, A. C., Lawrence, J. E., and Suttle, C. A. (2002). Lysogeny and Lytic Viral Production during a Bloom of the Cyanobacterium *Synechococcus* spp. *Microb. Ecol.* **in press**.

Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E. A., Partensky, F., and Larimer, F. W. (2003). The genome of a motile marine *Synechococcus*. *Nature* **424,** 1037-1041.

Paul, J. H., Sullivan, M. B., Segall, A. M., and Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**(4), 463-76.

Proctor, L. M., and Fuhrman, J. A. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature* **343,** 60-62.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., 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., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042-1047.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45,** 408-418.

Safferman, R. S., Diener, T. O., Desjardins, P. R., and Morris, M. E. (1972). Isolation and characterization of AS-1, a phycovirus infecting the blue-green algae, *Anacystis nidulans* and *Synechococcus cedorum*. *Virology* **47,** 105-113.

Samimi, B., and Drews, G. (1978). Adsorption of cyanophage AS-1 to unicellular cyanobacteria and isolation of receptor material from *Anacystis nidulans*. *J. Virol.* **25,** 164-174.

Sherman, L. A., and Connelley, M. (1976). Isolation and characterization of a cyanophage infecting the unicellular blue-green algae *A. nidulans* and *S. cedorum*. *Virology* **72,** 540-544.

Stent, G. S. (1963). "Molecular biology of bacterial viruses." W.H. Freeman and Company, San Francisco.

Steward, G., Smith, D. C., and Azam, F. (1996). Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Mar. Ecol. Prog. Ser.* **131,** 287-300.

Sullivan, M. B., and al., e. *Prochlorococcus* cyanophage genomes: Phages adapted for infection of open-ocean photosynthetic cells, in prep.

Sullivan, M. B., Rosenkrantz, J. R., Tan, G. P., Waterbury, J. B., and Chisholm, S. W. Diversity of portal genes from *Prochlorococcus* and *Synechococcus* myophage isolates. *Applied Environmental Microbiology* **in prep.**

Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W. (2003). Cyanophage infecting the oceanic cyanobacterium, *Prochlorococcus*. *Nature* **424,** 1047-1051.

Suttle, C. A. (2000). Cyanophages and their role in the ecology of cyanobacteria. *In* "The Ecology of Cyanobacteria" (B. A. Whitton, and M. Potts, Eds.), pp. 563-589. Kluwer Academic Publishers, Netherlands.

Suttle, C. A., and Chan, A. M. (1993). Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* **92,** 99-109.

Suttle, C. A., and Chan, A. M. (1994). Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Applied. Environ. Microbiol.* **60**(9), 3167-3174.

Tetart, F., Desplats, C., and Krisch, H. M. (1998). Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity. *J Mol Biol* **282**(3), 543-56.

Tetart, F., Repoila, F., Monod, C., and Krisch, H. M. (1996). Bacteriophage T4 host range is expanded by duplications of a small domain of the tail fiber adhesin. *J Mol Biol* **258**(5), 726-31.

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnol. Oceanogr.* **45**(6), 1320-1328.

Vogel, J., Axmann, I. M., Herzel, H., and Hess, W. R. (2003). Experimental and computational analysis of transcriptional start sites in the cyanobacterium Prochlorococcus MED4. *Nucleic Acids Res* **31**(11), 2890-9.

Waterbury, J. B., and Valois, F. W. (1993). Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**(10), 3393-3399.

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.* **214**, 71-120.

Wilson, W. H., Carr, N. G., and Mann, N. H. (1996). The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp. WH7803. *J. Phycol.* **32**, 506-516.

Wilson, W. H., Fuller, N. J., Joint, I. R., and Mann, N. H. (1999). Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. *In* "Marine Cyanobacteria" (L. Sharpy, and A. W. D. Larkum, Eds.), Vol. 19, pp. 209-216. Bulletin de l'Institut oceanographique, Monaco.

Wilson, W. H., Fuller, N. J., Joint, I. R., and Mann, N. H. (2000). Analysis of cyanophage diversity in the marine environment using denaturing gradient gel electrophoresis. *In* "Microbial Biosystems: New Fronteirs. Proceedings of the 8th International Symposium on Microbial Ecology" (C. R. Bell, M. Brylinksky, and P. Johnson-Green, Eds.). Atlantic Canada Society for Microbial Ecology, Halifax, Nova Scotia, Canada.

Wilson, W. H., Joint, I. R., Carr, N. G., and Mann, N. H. (1993). Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl Environ Microbiol* **59**, 3736-3743.

Zhong, Y., Chen, F., Wilhelm, S. W., Poorvin, L., and Hodson, R. E. (2002). Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl Environ Microbiol* **68**(4), 1576-84.

APPENDIX A

THE MIT CYANOPHAGE COLLECTION

Table of Established Clonal Cyanophage Isolates and Environmental Lysates

The following pages describe the various clonal phage isolates and lysates in the MIT cyanophage collection. Many of these phage lysates are unpublished, but may prove valuable resources for future studies of phage. TEM Morph = Transmission Electron Microscopy morphology (M=*Myoviridae*, S=*Siphoviridae*, P=*Podoviridae*), Head diam.= diameter across the head in nanometers, Tail length and width are measured in nanometers, # / window = number of particles observed per window on a 200 mesh copper carbon type B grid (Ted Pella, Redding CA #01811).

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
9211-1	BATS	100	6-Jun-00	0.1	M	60	100	20		7-Aug-00	
9211-2	BATS	100	6-Jun-00	0.1	M	60	100	20		7-Aug-00	
9211-3	BATS	100	6-Jun-00	0.01	M ?	60	15	15		7-Aug-00	probably a myo
9211-4	BATS	75	6-Jun-00	0.1	M ?	60	110	15	10	13-Sep-00	
9211-5	BATS	75	6-Jun-00	1	P?	40			5	13-Sep-00	
9211-6	BATS	45	Sep-99	1	M ?						
9211-7	BATS	100	Sep-99	1							
9211-8	BATS	120	Sep-99	1							
9211-9	BATS	120	Sep-99	0.1							
9211-10	BATS	70	Sep-99	1						30-Jan-01	
9211-11	BATS	100	Sep-99	0.1						30-Jan-01	
9211-12	BATS	100	Sep-99	1						30-Jan-01	
9211-13	BATS	120	Sep-99	0.1						30-Jan-01	
9211-14	BATS	120	Sep-99	1						30-Jan-01	
9211-15	BATS	45	Sep-99	1						25-Feb-01	
9211-16	Sarg	30	Sep-01	plaque	M ?					28-Jan-02	smooth edge
9211-17	Sarg	50	Sep-01	plaque						28-Jan-02	clear, crisp edge
9211-18	Sarg	110	Sep-01	plaque						28-Jan-02	big, rough edge
9211-19	Sarg	130	Sep-01	plaque						28-Jan-02	big, rough edge
9215-1	BATS	100	6-Jun-00	0.1	P	40			200	15-Aug-00	
9215-2	BATS	10	6-Jun-00	1	P	40			10	15-Aug-00	
9215-3	BATS	100	6-Jun-00	1	P	50			1000	13-Aug-00	8.33x10^6 plaque assay
9215-4	Red Sea	sfc	15-Jul-00	0.01	P	40			200	29-Aug-00	
9215-5	Red Sea	sfc	15-Jul-00	0.01	P	40	12		200	29-Aug-00	

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
9215-5	–	–	–	–	P?	65			10		large Podos never clear / sharp, but common
9215-6	Red Sea	sfc	15-Jul-00	0.01	S?	40	150	6	1	29-Aug-00	R2 liquid, only podos (40nm)
9215-6	–	–	–	–	M ?	60	100	12	40		
9215-6	–	–	–	–	P	55			50		CLONAL
9215-7	Red Sea	sfc	15-Jul-00	0.01	NOTHING					29-Aug-00	
9215-8	BATS	75	6-Jun-00	1	P	40			100	13-Sep-00	
9215-8	–	–	–	–	P	60			1		
9215-9	Red Sea	5	13-Sep-00	0.1	P	40			50	25-Oct-00	
9215-10	Red Sea	5	13-Sep-00	1	P	40			50	25-Oct-00	
9215-10	–	–	–	–	P	60			2		
9215-11	Red Sea	50	13-Sep-00	0.01	P	40			300	25-Oct-00	
9215-11	–	–	–	–	P	60			3		
9215-12	Red Sea	50	13-Sep-00	0.1	P	40			100	25-Oct-00	
9215-13	Red Sea	50	13-Sep-00	1	P	40			30	25-Oct-00	
9215-14	Red Sea	100	13-Sep-00	0.1	P	40			50	25-Oct-00	
9215-14	–	–	–	–	P	60			3		
9215-15	Red Sea	100	13-Sep-00	1	P	40			50	25-Oct-00	
9215-15	–	–	–	–	P	60			5		
9215-16	Red Sea	130	13-Sep-00	0.1	P	40			30	25-Oct-00	
9215-17	Red Sea	130	13-Sep-00	1	P	40			50	25-Oct-00	
9215-17	–	–	–	–	P	60			2		
9215-18	BATS	45	Sep-99	1	P	40			20	25-Feb-01	
9215-19	Red Sea	5	13-Sep-00	1	P	40			30	7-Mar-01	
9215-20	Red Sea	50	13-Sep-00	1						7-Mar-01	
9215-21	Red Sea	100	13-Sep-00	0.1						7-Mar-01	
9215-22	Red Sea	130	13-Sep-00	0.1						7-Mar-01	
9302-1	Red Sea	sfc	15-Jul-00	0.1	P	40			50	15-Aug-00	
9302-2	BATS	100	6-Jun-00	1	P	40			20	15-Aug-00	
9302-3	BATS	100	6-Jun-00	1	M	60	100	12		15-Aug-00	2 shots contractile tail
9302-4	BATS	100	6-Jun-00	1	P	40			20	15-Aug-00	
9302-5	BATS	100	6-Jun-00	1	P	40			20	15-Aug-00	
9302-6	Red Sea	sfc	15-Jul-00	1	P	40			50	1-Aug-00	
9302-6	–	–	–	–	M ?	60	100	12	1		
9302-7	Red Sea	sfc	15-Jul-00	1	P	40			100	1-Aug-00	
9302-7	–	–	–	–	M	60	120	12	40		
9302-8	BATS	75	6-Jun-00	0.1	P	40			200	8-Sep-00	
9302-8	–	–	–	–	M ?	60	none?		1		
9302-9	BATS	75	6-Jun-00	1	NOTHING					8-Sep-00	
9302-10	Red Sea	5	13-Sep-00	1	P	40			30	25-Oct-00	
9302-11	Red Sea	50	13-Sep-00	1	P	40			30	25-Oct-00	

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
9302-12	Red Sea	100	13-Sep-00	1	P	40			10	25-Oct-00	
9302-13	BATS	100	Sep-99	0.1	P	40			500	4-Jan-01	
9302-13	–	–	–	–	P	60			1		
9302-14	BATS	120	Sep-99	0.1	P	40			300	4-Jan-01	
9302-15	Red Sea	5	13-Sep-00	0.1						7-Mar-01	
9302-16	Red Sea	50	13-Sep-00	0.01						7-Mar-01	
9302-17	Red Sea	100	13-Sep-00	0.1						7-Mar-01	
9302-18	Red Sea	130	13-Sep-00	1						7-Mar-01	
9303-1	BATS	100	6-Jun-00	1	M ?	60			10	15-Aug-00	probably myos
9303-2	BATS	100	6-Jun-00	1	M	60			10	15-Aug-00	CLONAL, HR, PCR
9303-3	Red Sea	sfc	15-Jul-00	1	M	60	100	12	30	8-Sep-00	CLONAL, HR, PCR
9303-4	BATS	75	6-Jun-00	0.1	NOTHING					8-Sep-00	
9303-5	BATS	75	6-Jun-00	0.1	M ?				2	8-Sep-00	
9303-6	BATS	75	6-Jun-00	1	M	60	110	12	5	8-Sep-00	Cannot find stock ???
9303-7	BATS	100	6-Jun-00	1	M	60	110	12	2	8-Sep-00	Cannot find stock ???
9303-8	Red Sea	5	13-Sep-00	plaque	M	100	75	25		11-May-01	large plaques ~3+mm
9303-9	Red Sea	100	13-Sep-00	plaque						11-May-01	large plaques ~3+mm
9303-10	Red Sea	130	13-Sep-00	plaque	M	75	75	25		11-May-01	large plaques ~3+mm
9312-1	BATS	10	6-Jun-00	1	P	40			40	15-Aug-00	NOT a virus ...
9312-1					P	60	25		5		
9312-2	BATS	100	6-Jun-00	1	P	40				13-Aug-00	
9312-3	BATS	100	6-Jun-00	1	P	40			20	13-Aug-00	
9312-4	BATS	100	6-Jun-00	1	P	40			1	13-Aug-00	
9312-5	BATS	100	6-Jun-00	1	P	40			1	13-Aug-00	
9312-6	BATS	100	6-Jun-00	1	P	40			1	13-Aug-00	
9312-7	BATS	75	6-Jun-00	1	P	40			5	8-Sep-00	
9312-8	Red Sea	50	13-Sep-00	1	P	40			25	25-Oct-00	
9312-9	BATS	120	Sep-99	0.1	P	40			50	4-Jan-01	CLONAL
9312-10	BATS	100	Sep-99	0.1	P	55			50	4-Jan-01	CLONAL
9312-11	BATS	70	Sep-99	1	P	55			100	4-Jan-01	7.7x10^7 plaque assay count, CLONAL
9312-12	BATS	45	Sep-99	0.01	P	40			10	4-Jan-01	CLONAL
9312-12					P	60			2		
9312-13	BATS	3	Sep-99	1	P	40			10	4-Jan-01	CLONAL
9312-14	BATS	15	Sep-99	0.1	P	50			300	17-Jan-01	CLONAL
9312-15	Red Sea	50	13-Sep-00	1	P	40			30	2-Mar-01	
9313-1	Shelf	40	Sep-01	plaque	M					25-Nov-01	big, clear plaques with uneven edges, 2x plaque purified
9313-2	Shelf	0	Sep-01	plaque	M					25-Nov-01	2xplaque purified
9313-3	Slope	60	17-Sep-01	plaque	S	60				12-Dec-01	2xplaque purified

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
9313-4	Slope	83	17-Sep-01	plaque	S	50	250	15	30	12-Dec-01	100 nm elongated head, 2x plaque purified
9515-1	Red Sea	sfc	15-Jul-00	0.1	P	40			50		
9515-2	Red Sea	sfc	15-Jul-00	1	P	40			50		
9515-3	BATS	10	6-Jun-00	0.1	P	40			5		
9515-4	BATS	10	6-Jun-00	1	P	40			30		
9515-5	BATS	75	6-Jun-00	0.01	P	40			50		
9515-6	BATS	75	6-Jun-00	0.1	P	40			30		
9515-7	BATS	75	6-Jun-00	1	P	40			30		
9515-8	BATS	100	6-Jun-00	0.1	P	40			10		
9515-9	BATS	100	6-Jun-00	1	P	40			10		
9515-10	BATS	120	Sep-99	0.1	P	55			300	4-Jan-01	2x10^8 plaque count
9515-11	BATS	100	Sep-99	0.01	P	55			1000	4-Jan-01	2.1x10^8 plaque count
9515-12	BATS	70	Sep-99	0.1	P	40			1000	4-Jan-01	2.5x10^8 plaque count
9515-13	BATS	45	Sep-99	0.1	P	40			1000	4-Jan-01	2.9x10^8 plaque count
9515-14	BATS	3	Sep-99	1	P	40			20	8-Jan-01	1.8x10^8 plaque count
9515-14	–	–	–	–	P	60			5		
9515-15	Red Sea	5	13-Sep-00	0.1	P	40			50	2-Mar-01	
9515-16	Red Sea	50	13-Sep-00	0.1	P	50 ?			100	2-Mar-01	
9515-17	Red Sea	100	13-Sep-00	0.1	P	50 ?			100	2-Mar-01	
9515-18	Red Sea	130	13-Sep-00	0.1	P	50 ?			100	2-Mar-01	
SS120-1	BATS	100	Sep-99	LV	P	60	15	10	40	13-Feb-00	0.2um filtered
SS120-2	BATS	100	Sep-99	LV	P	50			40	16-Feb-00	0.2um filtered
SS120-3	BATS	120	Sep-99	1	P	50			1000		
SS120-4	BATS	45	Sep-99	1	P	50			1000		
SS120-5	Red Sea	130	13-Sep-00	1	M	80 +	30+	25	10	7-Mar-01	
SS120-6	Slope	83	17-Sep-01	plaque	P	50	10		1000 0s	29-Dec-01	CLONAL
MED4-1	BATS	3	Sep-99	LV	P	40			few	15-May-00	0.2um filtered
MED4-2	BATS	15	Sep-99	LV	P	40			few	15-May-00	0.2um filtered
MED4-3	BATS	40	Sep-99	LV	P	40			100	15-May-00	0.2um filtered
MED4-3	–	–	–	–	M/ S?	65	110		few		
MED4-4	BATS	70	Sep-99	LV	P	40			300	15-May-00	0.2um filtered
MED4-5	BATS	100	Sep-99	LV	P	40			500	2-Feb-00	0.2um filtered, CLONAL
MED4-6	Gulf Stream	3	Sep-99	LV	P	40			?	15-May-00	0.2um filtered, CLONAL
MED4-7	Gulf Stream	40	Sep-99	LV	P	40			?	15-May-00	0.2um filtered
MED4-7	–	–	–	–	M ?	65	110		?		
MED4-8	Gulf Stream	80	Sep-99	LV	P	45			50	2-Feb-00	0.2um filtered, CLONAL

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
MED4-9	Core	3	Sep-99	LV	P	40			10	17-Jul-00	0.2um filtered, CLONAL
MED4-10	Core	40	Sep-99	LV	P	40			50	15-May-00	0.2um filtered
MED4-10	–	–	–	–	P	60			few		
MED4-11	Core	80	Sep-99	LV	P	40			100	15-May-00	0.2um filtered
MED4-12	Core	100	Sep-99	LV	P	40			450	15-May-00	0.2um filtered
MED4-13	Core	120	Sep-99	LV	P	40			80	15-May-00	0.2um filtered
MED4-13	–	–	–	–	M ?	65	110		few		
MED4-14	BATS	10	6-Jun-00	1	P	40			1000	15-Aug-00	
MED4-15	BATS	100	6-Jun-00	1	P	60			20	15-Aug-00	
MED4-16	Slope	70	Sep-99	LV	P	40			?	14-Mar-00	0.2um filtered
MED4-17	Red Sea	sfc	15-Jul-00	0.001	P	40			10	15-Aug-00	
MED4-18	Red Sea	sfc	15-Jul-00	0.01	P	40			100	15-Aug-00	
MED4-19	Red Sea	sfc	15-Jul-00	0.01	P	50-60?			few	15-Aug-00	
MED4-20	Red Sea	sfc	15-Jul-00	0.1	NOTHING					15-Aug-00	
MED4-21	BATS	70	Sep-99	0.01	P	60			30	8-Sep-00	
MED4-22	BATS	45	Sep-99	0.1	P	40			300	8-Sep-00	
MED4-23	BATS	45	Sep-99	0.1	NOTHING					8-Sep-00	
MED4-24	BATS	70	Sep-99	0.1	P	40			2	13-Sep-00	
MED4-25	BATS	70	Sep-99	0.1	P	60			4	13-Sep-00	
MED4-26	BATS	100	Sep-99	0.01	P	40			50	13-Sep-00	
MED4-27	BATS	100	Sep-99	0.1	P	40			5	13-Sep-00	
MED4-28	BATS	100	Sep-99	0.1	P	40			5	13-Sep-00	
MED4-29	BATS	75	6-Jun-00	0.001	P	40			1	13-Sep-00	low contrast
MED4-30	BATS	75	6-Jun-00	0.01	P	40			5	13-Sep-00	
MED4-31	BATS	75	6-Jun-00	0.1	P	40			10	13-Sep-00	
MED4-32	BATS	15	Sep-99	1	P	40			20	24-Sep-00	
MED4-33	BATS	15	Sep-99	1	P	40			5	24-Sep-00	
MED4-34	BATS	3	Sep-99	1	P	40			10	24-Sep-00	
MED4-35	BATS	3	Sep-99	0.1	P	40				24-Sep-00	
MED4-36	BATS	3	Sep-99	0.1	P	40			80	24-Sep-00	
MED4-37	BATS	120	Sep-99	0.01	P	40			50	24-Sep-00	
MED4-38	BATS	120	Sep-99	0.1	P	40			40	24-Sep-00	
MED4-39	BATS	120	Sep-99	1	P	40			100	24-Sep-00	pos, neg stained podos
MED4-40	Red Sea	5	13-Sep-00	0.001	P	40			15	25-Oct-00	
MED4-41	Red Sea	5	13-Sep-00	0.01	P	40			50	25-Oct-00	
MED4-42	Red Sea	5	13-Sep-00	0.1	P	40			100	25-Oct-00	CLONAL
MED4-43	Red Sea	5	13-Sep-00	1	P	40			10	25-Oct-00	
MED4-44	Red Sea	50	13-Sep-00	0.01	P	40			30	25-Oct-00	CLONAL
MED4-45	Red Sea	50	13-Sep-00	1	P	40			50	25-Oct-00	
MED4-46	Red Sea	100	13-Sep-00	0.01	P	40			10	25-Oct-00	
MED4-47	Red Sea	100	13-Sep-00	1	P	40			50	25-Oct-00	
MED4-48	Red Sea	130	13-Sep-00	0.1	P	40			50	25-Oct-00	CLONAL

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
MED4-49	Red Sea	130	13-Sep-00	1	P	40			30	25-Oct-00	
MED4-50	BATS	100	Sep-99	0.01	P	40			50	3-Nov-00	
MED4-51	BATS	100	Sep-99	0.1	P	40			300	3-Nov-00	
MED4-52	BATS	100	Sep-99	1	P	40			100	3-Nov-00	
MED4-53	Red Sea	5	13-Sep-00	0.01	P	55			30	2-Mar-01	no tails observed at all
MED4-54	Red Sea	50	13-Sep-00	0.001	P	55			100	2-Mar-01	no tails observed at all
MED4-55	Red Sea	100	13-Sep-00	0.1	P	60			30	2-Mar-01	no tails observed at all
MED4-56	Red Sea	130	13-Sep-00	1	P	60			30	2-Mar-01	no tails observed at all
NATL1A-1	BATS	100	6-Jun-00	0.01	P?	40				7-Aug-00	TEM~1x10^6
NATL1A-1	–	–	–	–	M	65	100	12	20		plaque CLONAL (2x)
NATL1A-2	BATS	100	6-Jun-00	0.01	M	65	100	12	50	7-Aug-00	
NATL1A-3	BATS	100	6-Jun-00	0.01	P	55-60				7-Aug-00	
NATL1A-4	BATS	100	6-Jun-00	0.01	NOTHING					7-Aug-00	
NATL1A-5	Red Sea	5	13-Sep-00	1						7-Mar-01	
NATL1A-6	Red Sea	50	13-Sep-00	0.1						7-Mar-01	
NATL1A-7	Red Sea	130	13-Sep-00	1	M					7-Mar-01	
NATL1A-8											missing
NATL1A-9	BATS	75	6-Jun-00	0.1	M				100	16-Oct-00	
NATL1A-10	BATS	75	6-Jun-00	1	P	40			1000	16-Oct-00	
NATL1A-11	BATS	45	Sep-99	1	P	40			10	6-Jan-01	
NATL1A-12	BATS	70	Sep-99	0.1	P	40			1000	6-Jan-01	
NATL1A-13	BATS	100	Sep-99	0.01	P	40			100	6-Jan-01	
NATL1A-14	BATS	100	Sep-99	0.01	P	40			1000	6-Jan-01	
NATL1A-15	BATS	120	Sep-99	1	M	75	30+	25	30	6-Jan-01	
NATL2A-1	BATS	100	6-Jun-00	0.01	M	65	170	20			
NATL2A-2	BATS	100	6-Jun-00	0.01	P	40			100	7-Aug-00	
NATL2A-2	–	–	–	–	P	60			10		
NATL2A-3	BATS	100	6-Jun-00	0.01	M	75	85	15	1	7-Aug-00	plaque CLONAL
NATL2A-4	BATS	10	6-Jun-00	1	M	80	160	20	100	11-Aug-00	plaque CLONAL
NATL2A-5	Red Sea	sfc	15-Jul-00	1	M	60	110	12	5	4-Sep-00	PLATES 15 JAN 01
NATL2A-6	Red Sea	sfc	15-Jul-00	1	M	60	110	12	3	4-Sep-00	
NATL2A-7	Red Sea	sfc	15-Jul-00	1	M	60	110	12	3	4-Sep-00	spilled, only ~1ml left
NATL2A-8	Red Sea	sfc	15-Jul-00	0.01						8-Sep-00	
NATL2A-9	Red Sea	sfc	15-Jul-00	0.01	M	60	110	12	10	8-Sep-00	NEG STAINED
NATL2A-10	BATS	75	6-Jun-00	0.1	P?	60			150	8-Sep-00	probably myos
NATL2A-11	BATS	75	6-Jun-00	0.1	P?	60			few	13-Sep-00	probably myos
NATL2A-12	BATS	75	6-Jun-00	0.1	M	60	110	12	3	13-Sep-00	
NATL2A-13	BATS	15	Sep-99	0.01	M	60	110	12	3	14-Sep-00	no photo but similar to 12
NATL2A-14	BATS	15	Sep-99	0.01	M	70	80+	20	3	14-Sep-00	CLONAL
NATL2A-15	BATS	15	Sep-99	0.1	P?	60			2	14-Sep-00	probably myos
NATL2A-16	BATS	3	Sep-99	0.1	M	60	110	12	1	14-Sep-00	
NATL2A-17	BATS	3	Sep-99	0.1	M	75	160	20	10	14-Sep-00	

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
NATL2A-18	BATS	3	Sep-99	0.1	M ?	75			5	14-Sep-00	
NATL2A-19	BATS	40	Sep-99	0.1	M				1	27-Sep-00	new stock (21Jan01), CLONAL
NATL2A-20	BATS	40	Sep-99	0.1	??					27-Sep-00	spilled,none left,never looked at
NATL2A-21	BATS	40	Sep-99	1						27-Sep-00	new stock (21Jan01)
NATL2A-22	BATS	70	Sep-99	0.01						27-Sep-00	new stock (21Jan01)
NATL2A-23	BATS	70	Sep-99	0.1						27-Sep-00	new stock (21Jan01)
NATL2A-24	BATS	70	Sep-99	1	P	40			10	27-Sep-00	new stock (21Jan01)
NATL2A-25	BATS	120	Sep-99	0.01	M	75	110 +	20	500	27-Sep-00	
NATL2A-26	BATS	120	Sep-99	0.1	M ?	75			50	27-Sep-00	
NATL2A-27	BATS	120	Sep-99	1	M	75	150	20	3	27-Sep-00	
NATL2A-28	Red Sea	50	13-Sep-00	0.001	M	75	30+	25	5	21-Oct-00	
NATL2A-29	Red Sea	50	13-Sep-00	0.01	NOTHING					21-Oct-00	
NATL2A-30	BATS	100	Sep-99	0.01	M	75	100	25	5	21-Oct-00	CLONAL
NATL2A-31	BATS	100	Sep-99	1	P	50 ?			10	21-Oct-00	CLONAL
NATL2A-32	Red Sea	5	13-Sep-00	0.1	M	75	75+	20	5	3-Nov-00	
NATL2A-33	Red Sea	5	13-Sep-00	1	M	75	110	20	5	3-Nov-00	
NATL2A-34	Red Sea	50	13-Sep-00	0.001	M	75	30+	?	5	3-Nov-00	
NATL2A-35	Red Sea	50	13-Sep-00	0.01	M	75	30+	?	5	3-Nov-00	
NATL2A-36	Red Sea	100	13-Sep-00	0.001	M	75 +	30+	?	5	3-Nov-00	
NATL2A-37	Red Sea	100	13-Sep-00	0.01	M	75	150	20	5	3-Nov-00	
NATL2A-38	Red Sea	130	13-Sep-00	0.01	P	40			1000	3-Nov-00	
NATL2A-39	Red Sea	130	13-Sep-00	1						3-Nov-00	
NATL2A-40	Red Sea	50	13-Sep-00	plaque	M	75	120	25	300	3-Nov-00	CLONAL, great photos
NATL2A-41	Red Sea	50	13-Sep-00	plaque	M	80	80+	25	30	3-Nov-00	CLONAL, great photos
NATL2A-42	Red Sea	50	13-Sep-00	plaque	P	40			1000	3-Nov-00	CLONAL
NATL2A-43	Red Sea	50	13-Sep-00	plaque	P	50 ?			300	3-Nov-00	CLONAL
NATL2A-44	Red Sea	5	13-Sep-00	plaque						23-Apr-01	"clear" plaque
NATL2A-45	Red Sea	5	13-Sep-00	plaque						23-Apr-01	"old" plaque
NATL2A-46	Red Sea	130	13-Sep-00	plaque						23-Apr-01	
NATL2A-47	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-48	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-49	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-50	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-51	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-52	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-53	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-54	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-55	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-56	Sarg	0	Sep-2001	plaque						Nov/Dec2002	

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
NATL2A-57	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-58	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-59	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-60	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-61	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-62	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-63	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-64	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-65	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-66	Sarg	0	Sep-2001	plaque						Nov/Dec2002	
NATL2A-67	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-68	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-69	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-70	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-71	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-72	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-73	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-74	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-75	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-76	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-77	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-78	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-79	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-80	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
NATL2A-81	Sarg	95	Sep-2001	plaque						Nov/Dec2002	
9107-1	Red Sea	50	13-Sep-00	plaque							
9201-1	Red Sea	50	13-Sep-00	plaque						5-May-01	CLONAL
9202-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
9202-2	Red Sea	50	13-Sep-00	plaque						5-May-01	CLONAL
9301-1	Red Sea	50	13-Sep-00	plaque							
9311-1	Red Sea	50	13-Sep-00	plaque							
9314-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
9314-2	Red Sea	50	13-Sep-00	plaque						5-May-01	CLONAL
9401-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
MED1-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
AS9601-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
DV-1	Red Sea	50	13-Sep-00	plaque						13-Apr-01	CLONAL
GP2-1	Red Sea	50	13-Sep-00	plaque							
SP-1	Red Sea	50	13-Sep-00	plaque							
6501-1	Slope	0	17-Sep-01	plaque	M	80	80+	25		28-Nov-01	CLONAL
6501-2	Slope	15	17-Sep-01	plaque						28-Nov-01	
6501-3	Slope	40	17-Sep-01	plaque						28-Nov-01	
6501-4	Slope	60	17-Sep-01	plaque						29-Nov-01	

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
6501-5	Shelf	0	16-Sep-01	plaque	M	80	110	25		29-Nov-01	CLONAL
6501-6	Shelf	40	16-Sep-01	plaque						29-Nov-01	
6501-7	Shelf	50	16-Sep-01	plaque						29-Nov-01	
6501-8	Shelf	27	16-Sep-01	plaque						29-Nov-01	
6501-9	Sarg	70	22-Sep-01	plaque	M	80	110	25		29-Nov-01	CLONAL, great photos
6501-10	Sarg	30	22-Sep-01	plaque						8-Dec-01	
6501-11	WHOI	0	5-Sep-01	plaque						4-Feb-02	
7803-1	Shelf	0	16-Sep-01	plaque						4-Nov-01	Big plaque
7803-2	Slope	0	17-Sep-01	plaque						4-Nov-01	
7803-3	Sargasso	0	22-Sep-01	plaque						4-Nov-01	
7803-4	Slope	40	17-Sep-01	plaque						4-Nov-01	
7803-5	Sargasso	70	22-Sep-01	plaque						4-Nov-01	
7803-6	Shelf	50	16-Sep-01	plaque						4-Nov-01	
7803-7	Slope	83	17-Sep-01	plaque						4-Nov-01	
7803-8	Sargasso	130	22-Sep-01	plaque						4-Nov-01	
7803-9	Shelf	0	16-Sep-01	plaque						9-Nov-01	little plaque
7803-10	Slope	60	17-Sep-01	plaque						9-Nov-01	
8017-1	Slope	15	17-Sep-01	plaque						28-Feb-03	2xplaque purified
8017-2	Slope	15	17-Sep-01	plaque						28-Feb-03	2xplaque purified
8018-1	Shelf	0	16-Sep-01	plaque						4-Nov-01	
8018-2	Slope	0	17-Sep-01	plaque						4-Nov-01	
8018-3	Sargasso	0	22-Sep-01	plaque						4-Nov-01	2xplaque purified
8018-4	Shelf	40	16-Sep-01	plaque						4-Nov-01	
8018-5	Slope	60	17-Sep-01	plaque						4-Nov-01	
8018-6	Slope	15	17-Sep-01	plaque						29-Nov-01	
8018-7	Sargasso	50	22-Sep-01	plaque						29-Nov-01	
8018-8	Sargasso	110	22-Sep-01	plaque						29-Nov-01	2xplaque purified
8018-9	Shelf	50	16-Sep-01	plaque						9-Nov-01	
8101-1	Slope	0	17-Sep-01	plaque						9-Nov-01	
8102-1	Slope	0	17-Sep-01	plaque						15-Nov-01	
8102-2	Slope	40	17-Sep-01	plaque						15-Nov-01	
8102-3	Slope	60	17-Sep-01	plaque						15-Nov-01	
8102-4	Shelf	0	16-Sep-01	plaque	M	70 ?	100 +	20 +		15-Nov-01	mid-size, clear plaque
8102-5	Shelf	0	16-Sep-01	plaque						15-Nov-01	large turbid plaque
8102-6	Shelf	50	16-Sep-01	plaque						15-Nov-01	
8102-7	Shelf	40	16-Sep-01	plaque						15-Nov-01	
8102-8	Sargasso	0	22-Sep-01	plaque	M	80	200	25		15-Nov-01	CLONAL
8102-9	Shelf	27	16-Sep-01	plaque						8-Dec-01	large plaque
8102-10	Shelf	27	16-Sep-01	plaque						8-Dec-01	tiny plaque
8102-11	Sargasso	0	22-Sep-01	plaque						29-Dec-01	clear plaque
8102-12	Sargasso	95	22-Sep-01	plaque						29-Dec-01	clear plaque
8102-13	Sargasso	110	22-Sep-01	plaque						29-Dec-01	clear plaque

Virus	Locale	Depth (m)	Date Collected	Dilution	TEM Morph	Head diam.	Tail Length	Tail Width	# / window	Date isolated	Notes
8109-1	WHOI	0	5-Sep-01	plaque						4-Feb-02	
8109-2	Sargasso	70	22-Sep-01	plaque						4-Feb-02	
8109-3	Sargasso	95	22-Sep-01	plaque						4-Feb-02	
Syn 1											from JW Jan 2003
Syn 2					M	66	149	17			from JW Jan 2003
Syn 5					P	60					from JW Jan 2003
Syn 9					M	87	153	19			from JW Jan 2003
Syn 10					M	100	145	19			from JW Jan 2003
Syn 12					P	45	8	10			
Syn 14					M	93	136	21			
Syn 19					M	60					from JW Jan 2003
Syn 26											from JW Jan 2003
Syn 30											from JW Jan 2003
Syn 33											
S-WHM1					M	88	108	23			from WW June 2002
S-PM2					M	90	165	20			from WW June 2002

APPENDIX B

PROPHAGE IN HOST GENOMES

APPENDIX B. PROPHAGE IN HOST GENOMES

Introduction

There are two classes of phages: lytic and temperate. Lytic phages infect their host, use the host cellular processes to build new phage particles, and then burst the host cell releasing progeny particles. Other phages (temperate phages) infect their hosts and temporarily insert their DNA into the host genome as a prophage. Expression of prophage genes often fundamentally changes the host's physiology – a process known as lysogenic conversion (Calendar, 1988). This is best studied where prophage encode medically important, pathogen-associated toxin genes (Boyd, Davis, and Hochhut, 2001; Miao and Miller, 1999; Wagner and Waldor, 2002). Many prophage-encoded virulence factors confer advantages to the cells they infect. These virulence factors range from toxins (e.g., diphtheria toxin, cholera toxin) to conversion genes (e.g., superoxide dismutase, LPS conversion) and are often advantageous to the host cell, for example by allowing pathogens to infect a broader range of hosts, or by directly improving the cell's fitness. In some cases, prophage are even responsible for horizontally transferring such toxins between microbial species (Banks, Beres, and Musser, 2002).

How common are prophage in the microbial world? Approximately 70% of the 56 completed bacterial genome sequences surveyed by Canchaya et al (2003) contain at least one prophage >10kb in size (Canchaya et al., 2003). Many microbial genomes contain multiple prophages: 51 of 82 bacterial species surveyed harbor 230 putative prophage (Casjens, 2003). Prophage are so common that they often account for a significant fraction of the "strain-specific" DNA between closely related microbial strains (Baba et al., 2002; Simpson et al., 2000; Smoot et al., 2002). An extreme example of such strain-specificity being due to prophage can be found in a comparison of the group A *Streptococcus* (GAS) serotype M1, M3 and M18 genomes. While these three strains share a core set of genes (~90% = 1.7 Mbp), almost all of the strain-specific variation was attributable between 4 and 6 phage and phage-like elements that accounted for 130 kb (7.1%, 4 phages), 235 kb (~12.4%, 6 phages), and 204 kb (10.8%, 5 phages) in the M1, M3 and M13 genomes, respectively (Beres et al., 2002).

Prophage genes in microbial genomes are often highly expressed as seen using genome-wide expression arrays to evaluate gene expression of a GAS strain between two temperatures (Smoot et al., 2001). Of the 144 genes that were differentially expressed (~9% of 1605 genes on the microarray), the majority in any one category (22) were designated as from mobile genes and phages. Another study compared free-living and biofilm *Pseudomonas aeruginosa* cells

(Whiteley et al., 2001). Of only 73 genes that were differentially expressed between these two populations, the most highly activated genes were from a temperate bacteriophage. Such expression of prophage genes often fundamentally changes the host's physiology – a process known as lysogenic conversion (Calendar, 1988). These changes have been best studied where prophage encode medically important, pathogen-associated toxin genes, such as cholera toxin (Waldor and Mekalanos, 1996), enterotoxin A (Coleman et al., 1989), streptococcal pyrogenic exotoxin A (Johnson, Schlievert, and Watson, 1980) C (Goshorn, Bohach, and Schlievert, 1988), L and M (Smoot et al., 2002).

Together, these findings emphasize that prophage are not only widespread in prokaryotes, but also frequently account for strain diversification at both the genome and transcriptome (expression) levels often altering the host cell's physiology. However, the genomes of currently available freshwater cyanobacterial genomes lack intact prophage (Canchaya et al., 2003; Casjens, 2003) and no direct observation has confirmed the existence of prophage in marine cyanobacteria. Indirect measures from induction experiments in the field suggest that temperate cyanophage can be induced from *Synechococcus* strains (McDaniel et al., 2002; Ortmann, Lawrence, and Suttle, 2002), but without evidence showing intact prophage genomes integrated in the host genome, plausible alternative explanations can be made. During this thesis, three cyanobacterial genomes were sequenced by the Department of Energy Joint Genome Institute (Palenik et al., 2003; Rocap et al., 2003). We wondered whether these cyanobacterial hosts contained intact prophage and whether such prophage might influence the metabolic capacity of their hosts. To this end, I analyzed these host genomes for the presence of intact prophage.

Methods

The search for prophages was greatly facilitated by conversations with Dr. Sherwood Casjens. At the time, there were no standard methodologies for searching for prophages, but there was a growing belief that prophage were common in microbial genomes though their ubiquity was little mentioned in the literature until two recent reviews of prophage identified from the plethora or microbial genomes that had become available (Canchaya et al., 2003; Casjens, 2003).

I used the following method to find prophage (Casjens, pers. comm.):

(1) Use BLASTp searching (e-value cut-off <-3) to look for phage genes within the host genome that clustered within ~60 kb of DNA sequence

(2) Treat phage genes encoding the portal and terminase proteins as strong indicators of a prophage due to a lack of a function in the hosts (termed "cornerstone" genes), while phage genes such as ribonucleotide reductases commonly found in phage are also found in hosts so should not be used for identification of a prophage.

(3) In each clustering of phage-related genes within a genome, use synteny (order of genes) arguments and iterative PSI-BLASTing to determine if other genes in the surrounding region might also be phage-related to build confidence in putative prophages

(4) If a convincing prophage is identified, the ends can be found by looking for >10 bp direct repeats that act as the site of integration

Results

Examination of the BLAST analyses for the *Prochlorococcus* MED4, *Prochlorococcus* MIT9313 and *Synechococcus* WH8102 suggested that 5 putative prophage regions were worth further investigation because these regions contained clusters of 3-6 phage-like genes.

Putative prophage regions initially identified by clustering of putative phage genes were:

- In *Prochlorococcus* MED4 (617,000 to 665,000, includes ORFs or1228 to or1260 and or0916 to or0963; 1,328,000 to 1,367,000, includes ORFs or1709 to or1756 and or0437 to or0476)
- In *Prochlorococcus* MIT 9313 (714,000 to 758,000, includes ORFs or0963 to or1001 and or2925 to or2945)
- In *Synechcoccus* WH 8102 (1,127,000 to 1,175,000, includes ORFs or0142 to oro178 and or3452 to or3418; 1,291,000 to 1,345,000, includes ORFs or0244 to or0300 and or3289 to or3335)

Iterative PSI-BLASTing and synteny (gene order) was used to see if further phage genes could be detected within these regions (Frank Larimer's group at ORNL wrote a script to automate the PSI-BLASTing for 3 iterations). These analyses did not reveal any more phage genes, but did identify many "host" genes within all five of these regions. Further, three of the five putative prophage were significantly homologous to regions in one or both of the other genomes which suggested these regions were vertically transferred from ancestral genomes

None of these putative prophage passed the remaining tests for identifying prophage (steps #2, 3, 4 from methods section), so it was concluded that there are no prophage in these genomes (see Discussion).

Discussion

While 5 regions from the three cyanobacterial genomes contained clusters of 3-6 phage genes, none of these regions proved bonafide prophage through subsequent testing. There were no "cornerstone" phage genes in these regions and there were many "host" genes that suggested these regions were not prophage. One caveat to these analyses is that the sequence-based searching tools may be limited in finding phage genes (and phage structural genes in particular) due to the absence of temperate cyanophage genes and genomes in the databases. However, sequencing of three *Prochlorococcus* cyanophage genomes by the Department of Energy Joint Genome Institute shows that many cyanophage genes are of significant homology to other known phage genes (Chapter 5), so it is likely that these results are accurate and did not miss intact prophage due to limitations in the database.

While none of the three marine cyanobacterial genomes examined contained intact prophage, it is unlikely that prophage do not exist in these hosts in the oceans. The presence of intact and degraded phage-related integrases in all three genomes (this Appendix) suggests that prophage have integrated into these genomes (Palenik et al., 2003). It is plausible that even if these cyanobacterial isolates once contained prophage, such prophage could have been induced during stressful culture conditions many, many stationary phases ago (in our batch culturing over the past 10+ years).

Integrase genes as evidence of past prophage events?

Integrase family recombinases are used by prophage and mobile elements to integrate their genome into a host genome (Nunes-Duby et al., 1998; Williams, 2002). However, the diversity of function of these integrase genes is larger than simply the function of prophage integration – these genes might also be involved in conjugative transposition, resolution of concatenated DNA circles, regulation of plasmid copy number, DNA excision to control gene expression for nitrogen fixation in Anabaena and DNA inversions controlling expression of cell surface proteins or DNA replication (Nunes-Duby et al., 1998).

To determine if there were phage-related integrase genes in our genomes, I searched the three host genomes to find all ORFs with sequence homology to known integrases. I then examined this list of candidate integrase genes using Clustal alignments against known integrases and resolvases to identify a suite of 'conserved motifs' identified in a review of 105 integrase family recombinases (Nunes-Duby et al., 1998): known conserved amino acids that are important

for the secondary structure (R-H-R-Y) as well as some less conserved sites ("GT" and "LLGH")(Table I). Interestingly the one MED4 integrase has an orthologue in MIT9313 (or3443) and two MIT9313 ORFs that are an integrase or a fragment integrase are orthologous to ORFs in Syn (or3056, or1261).

In spite of finding many putative integrase family recombinases (Table I), the regions surrounding these ORFs do not appear to be part of an intact prophage (i.e., a region containing many phage-related genes without "obviously non-phage related genes"). Interestingly, many of these integrase genes appear fragmented (MIT ORFs or3853, or2797, or0731, or1796 and Syn ORFs or0150, or3298, or3591, or3050) and a few of these degenerate and putative integrase genes are associated with a deviation in %GC and/or are next to a tRNA gene. Taken together, the high number of intact and fragmented integrase genes is suggestive evidence of past prophage integration events (Palenik et al., 2003).

The possible integrase gene in MED4 (or1025) may have been part of an ancient evolutionary event as it is part of an orthologous region in MED4, MIT9313 and WH8102 and its closest sequence homology is to the cyanobacterial *xisA* gene. In Anabaena PCC7120, this gene has evolved functions useful to the host as it acts as a site-directed recombinase involved in excision of large DNA fragments that disrupt the nitrogenase gene cluster during differentiation of cells into heterocysts (Apte and Prabhavathi, 1994). Further analysis could be done with this alternate function in mind for these recombinases in these marine cyanobacteria.

Table I: Possible integrase genes in *Prochlorococcus* and *Synechococcus* genomes as well as in the *Prochlorococcus* cyanophage P-SSP7. Data included here represent amino acid motifs identified from the alignments of 105 integrase genes surveyed in (Nunes-Duby et al., 1998); R-H-R-Y was specifically identified in this paper, while the GT and LLGH motifs were added to my search. Also included are notes about surrounding gene arrangement within each of these genomes in the "notes from genome" column. All of the ORFs that have "Y" in the "INT?" column are related to the INT family of site-specific recombinases by BLAST homology and have the necessary known critical amino acids to be classified as an INT. Note that not all INT genes are prophage related.

Table I: Putative integrase genes in marine cyanobacterial genomes

		INT?	R	H	R	Y	GT	LLGH	size (aa)	%GC	tRNA?	unique?	Notes from genome
MED4	or1025	Y	+	Y	+	+	GY	+	388			MITor3443	ilvB -- cobO, pyrH/smbA, frr/rrf
MIT9313	or2572	Y	+	K	+	+	GS	+	383			U	unchar. Reg. -- yqjL, hupE, murE, rpsD
	or2631	Y	+	+	+	+	FT	?	421	??	Y	SYNor3056	RNA_4, unchar. Reg. -- cobD, clpP3, clpP4, panB, RNA_49, ftsZ, 3ORFs, RNA_41
	or3853								49	Y			C terminus only?
	or2797								65	Y	Y		C terminus only? RNA_7, pyrC, gir, 3 ORFs, godA, ~10 unchar.
	or0731			Y		+		+	207				C terminus only? Unchar. Reg.
	or3443	Y	+	Y	+	+	GY	+	387			MEDor1025, SYN0611	ilvB -- cobO, pyrH/smbA, frr/rrf, tal
	or1796		+		K				185			SYNor1261	C terminus only? hisB, fabI -- degT, folK, chlD
	or2231	Y?	+	Y	+	+	GS		509	Y		U	spoT, unchar.reg.--unchar.reg., argJ, coaE
Syn	or2191	Y	+	Y	+	+	GF	+	426	Y		U	RNA_47, unchar.reg. -- unchar.reg., or2184, or1352
	or2184	Y	+	Y	+	+	GF	?	399	??		U	near or2191, or1352
	or1352		+						302			U	near or2191, or2184
	or1730	Y	+	Y	+	+	GY	+	390				fur, icc, unchar.reg. -- gltS, unchar.reg., RNA_37, RNA_36, 2x psbY genes
	or0150		+	+	+	+		?	191			U	C terminus only? Unchar. Reg. __ near or3442, RNA_34
	or3442	Y	+	S	+	+	GY	+	379		Y	U	or0150 -- nrdJ, prfC
	or3408	Y	+	Y	+	+	GC	+	450			U	unchar.reg. -- uncarh.reg., fmt, pmbA, tldD
	or3298	Y?	S	+	+	+		?	185			U	C terminus only? petF3, 2ORFs -- unchar. Reg.
	or3123	Y?	+	Y		+	GY	+	379			U	15 unchar.ORFs--15unchar.ORFs, RNA, argF, ftsH3
	or0499	Y	+	Y	+	+	GL	+	419			U	opposite strand from or3123
	or3056	Y	+	+	+	+	FT	?	421			MITor2631	also close to or3591; RNA_48, panB, hemN, clpP3, clpP4, ilvC, cobD, 10 unchar. ORFs -- or3055, or3052, INT, INT, INT, RNA
	or3591								46				grouped with or3051, or3050; C terminus only?
	or3051	Y	+	+	+	+	FS	?	308			U	grouped with or3591, or3050
	or3050		K	Y		H			114		Y	U	see notes for or3591
	or0611	Y	+	Y	+	+	GY	+	396			MEDor1025, MITor3443	ilvB, hemH -- or0612, cobD, pyrH/smbA, rrf, 6 ORFs, tal, ftsI, 2 genes cobC
	or2809	Y	+	+	+	+	FT	?	425		Y	U	RNA -- 10 unchar. ORFs, psaI, psaL
	or0914	Y?		Y	Y	+	+	+	472	Y		U	RNA_21, ORF, rpaA, holB, 3ORFs -- 3ORFs, radA, rpaB, plsX (SPLIT rpa operon?)
	or2481	Y	+	Y	+	+	GF	+	421			U	30 unchar. ORFs -- RNA_22
P-SSP7	INT	Y	+	+	+	+	+	+	291				

REFERENCES

Apte, S. K., and Prabhavathi, N. (1994). Rearrangements of nitrogen-fixation (nif) genes in the heterocystous cyanobacteria. *J. Biosci.* 19, 579-602.

Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A., Nagai, Y., Iwama, N., Asano, K., Naimi, T., Kuroda, H., Cui, L., Yamamoto, K., and Hiramatsu, K. (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* 359(9320), 1819-27.

Banks, D. J., Beres, S. B., and Musser, J. M. (2002). The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microbiol* 10(11), 515-21.

Beres, S. B., Sylva, G. L., Barbian, K. D., Lei, B., Hoff, J. S., Mammarella, N. D., Liu, M. Y., Smoot, J. C., Porcella, S. F., Parkins, L. D., Campbell, D. S., Smith, T. M., McCormick, J. K., Leung, D. Y., Schlievert, P. M., and Musser, J. M. (2002). Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci U S A* 99(15), 10078-83.

Boyd, E. F., Davis, B. M., and Hochhut, B. (2001). Bacteriophage-bacteriophage interactions in the evolution of pathogenic bacteria. *Trends in Microbiology* 9, 137-144.

Calendar, R. (1988). "The Bacteriophages." (R. Calendar, Ed.), I Plenum, New York.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol Mol Biol Rev* 67(2), 238-76.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* 49(2), 277-300.

Coleman, D. C., Sullivan, D. J., Russell, R. J., Arbuthnott, J. P., Carey, B. F., and Pomeroy, H. M. (1989). Staphylococcus aureus bacteriophages mediating the simultaneous lysogenic conversion of beta-lysin, staphylokinase and enterotoxin A: molecular mechanism of triple conversion. *J Gen Microbiol* 135 (Pt 6), 1679-97.

Goshorn, S. C., Bohach, G. A., and Schlievert, P. M. (1988). Cloning and characterization of the gene, speC, for pyrogenic exotoxin type C from Streptococcus pyogenes. *Mol Gen Genet* 212(1), 66-70.

Johnson, L. P., Schlievert, P. M., and Watson, D. W. (1980). Transfer of group A streptococcal pyrogenic exotoxin production to nontoxigenic strains of lysogenic conversion. *Infect Immun* 28(1), 254-7.

McDaniel, L., Houchin, L. A., Williamson, S. J., and Paul, J. H. (2002). Lysogeny in marine *Synechococcus*. *Nature* 415, 496.

Miao, E. A., and Miller, S. I. (1999). Bacteriophages in the evolution of pathogen-host interactions. *Proc Natl Acad Sci U S A* 96, 9452-9454.

Nunes-Duby, S. E., Kwon, H. J., Tirumalai, R. S., Ellenberger, T., and Landy, A. (1998). Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* 26(2), 391-406.

Ortmann, A. C., Lawrence, J. E., and Suttle, C. A. (2002). Lysogeny and Lytic Viral Production during a Bloom of the Cyanobacterium *Synechococcus* spp. *Microb. Ecol.* in press.

Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E. A., Partensky, F., and Larimer, F. W. (2003). The genome of a motile marine *Synechococcus*. *Nature* 424, 1037-1041.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., 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., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424, 1042-1047.

Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., Alves, L. M., Araya, J. E., Baia, G. S., Baptista, C. S., Barros, M. H., Bonaccorsi, E. D., Bordin, S., Bove, J. M., Briones, M. R., Bueno, M. R., Camargo, A. A., Camargo, L. E., Carraro, D. M., Carrer, H., Colauto, N. B., Colombo, C., Costa, F. F., Costa, M. C., Costa-Neto, C. M., Coutinho, L. L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A. P., Ferreira, A. J., Ferreira, V. C., Ferro, J. A., Fraga, J. S., Franca, S. C., Franco, M. C., Frohme, M., Furlan, L. R., Garnier, M., Goldman, G. H., Goldman, M. H., Gomes, S. L.,

Gruber, A., Ho, P. L., Hoheisel, J. D., Junqueira, M. L., Kemper, E. L., Kitajima, J. P., Krieger, J. E., Kuramae, E. E., Laigret, F., Lambais, M. R., Leite, L. C., Lemos, E. G., Lemos, M. V., Lopes, S. A., Lopes, C. R., Machado, J. A., Machado, M. A., Madeira, A. M., Madeira, H. M., Marino, C. L., Marques, M. V., Martins, E. A., Martins, E. M., Matsukuma, A. Y., Menck, C. F., Miracca, E. C., Miyaki, C. Y., Monteriro-Vitorello, C. B., Moon, D. H., Nagai, M. A., Nascimento, A. L., Netto, L. E., Nhani, A., Jr., Nobrega, F. G., Nunes, L. R., Oliveira, M. A., de Oliveira, M. C., de Oliveira, R. C., Palmieri, D. A., Paris, A., Peixoto, B. R., Pereira, G. A., Pereira, H. A., Jr., Pesquero, J. B., Quaggio, R. B., Roberto, P. G., Rodrigues, V., de, M. R. A. J., de Rosa, V. E., Jr., de Sa, R. G., Santelli, R. V., Sawasaki, H. E., da Silva, A. C., da Silva, A. M., da Silva, F. R., da Silva, W. A., Jr., da Silveira, J. F., Silvestri, M. L., Siqueira, W. J., de Souza, A. A., de Souza, A. P., Terenzi, M. F., Truffi, D., Tsai, S. M., Tsuhako, M. H., Vallada, H., Van Sluys, M. A., Verjovski-Almeida, S., Vettore, A. L., Zago, M. A., Zatz, M., Meidanis, J., and Setubal, J. C. (2000). The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* 406(6792), 151-7.

Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F., Parkins, L. D., Beres, S. B., Campbell, D. S., Smith, T. M., Zhang, Q., Kapur, V., Daly, J. A., Veasy, L. G., and Musser, J. M. (2002). Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* 99(7), 4668-73.

Smoot, L. M., Smoot, J. C., Graham, M. R., Somerville, G. A., Sturdevant, D. E., Migliaccio, C. A., Sylva, G. L., and Musser, J. M. (2001). Global differential gene expression in response to growth temperature alteration in group A Streptococcus. *Proc Natl Acad Sci U S A* 98(18), 10416-21.

Wagner, P. L., and Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infect Immun* 70, 3985-3993.

Waldor, M. K., and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272(5270), 1910-4.

Whiteley, M., Bangera, M. G., Bumgarner, R. E., Parsek, M. R., Teitzel, G. M., Lory, S., and Greenberg, E. P. (2001). Gene expression in Pseudomonas aeruginosa biofilms. *Nature* 413(6858), 860-4.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nuc. Acids Res.* 30, 866-875.

APPENDIX C

Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation

Reprinted with permission from *Nature*

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., 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., and Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042-1047.

4. Scanlan, D. J. & West, N. J. Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol. Ecol.* **40**, 1–12 (2002).
5. Ferris, M. J. & Palenik, B. Niche adaptation in ocean cyanobacteria. *Nature* **396**, 226–228 (1998).
6. Toledo, G., Palenik, B. & Brahamsha, B. Swimming strains of marine *Synechococcus* with widely different photosynthetic pigment ratios form a monophyletic group. *Appl. Environ. Microbiol.* **65**, 5247–5251 (1999).
7. Bushman, F. *Lateral DNA Transfer Mechanisms and Consequences* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002).
8. Brahamsha, B. An abundant cell-surface polypeptide is required for swimming by the nonflagellated marine cyanobacterium *Synechococcus*. *Proc. Natl Acad. Sci. USA* **93**, 6504–6509 (1996).
9. Monger, B. C., Landry, M. R. & Brown, S. L. Feeding selection of heterotrophic marine nanoflagellates based on the surface hydrophobicity of their picoplankton prey. *Limnol. Oceanogr.* **44**, 1917–1927 (1999).
10. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301**, 75–101 (2000).
11. Willey, J. M. & Waterbury, J. B. Chemotaxis toward nitrogenous compounds by swimming strains of marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **55**, 1888–1894 (1989).
12. Harano, Y. *et al.* Identification and nitrogen regulation of the cyanase gene from the cyanobacteria *Synechocystis* sp. strain PCC6803 and *Synechococcus* sp. strain PCC7942. *J. Bacteriol.* **179**, 5744–5750 (1997).
13. Clark, L. L., Ingall, E. D. & Benner, R. Marine phosphorus is selectively remineralized. *Nature* **393**, 426 (1998).
14. Zehr, J. P. & Ward, B. B. Nitrogen cycling in the ocean: new perspectives on processes and paradigms. *Appl. Environ. Microbiol.* **68**, 1015–1024 (2002).
15. Long, R. A. & Azam, F. Antagonistic interactions among marine pelagic bacteria. *Appl. Environ. Microbiol.* **67**, 4975–4983 (2001).
16. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47**, 976–988 (2002).
17. Palenik, B. & Dyhrman, S. T. in *Phosphorus in Plant Biology: Regulatory Roles in Molecular, Cellular, Organismic, and Ecosystem Processes* (eds Lynch, J. P. & Deikman, J.) 26–38 (American Society of Plant Physiologists, Rockville, MD, 1998).
18. Nyyssola, A., Kerovuo, J., Kaukinen, P., von Weymarn, N. & Reinikainen, T. Extreme halophiles synthesize betaine from glycine by methylation. *J. Biol. Chem.* **275**, 22196–22201 (2000).
19. Nomura, M., Ishitani, M., Takabe, T., Rai, A. K. & Takabe, T. *Synechococcus* sp. PCC7942 transformed with *Escherichia coli bet* genes produces glycine betaine from choline and acquires resistance to salt stress. *Plant Physiol.* **107**, 703–708 (1995).
20. Badger, M. R., Hanson, D. & Price, G. D. Evolution and diversity of CO2 concentrating mechanisms in cyanobacteria. *Funct. Plant Biol.* **29**, 161–173 (2002).
21. de Lorimier, R., Guglielmi, G., Bryant, D. A. & Stevens, S. E. J. Structure and mutation of a gene encoding a M, 33 000 phycocyanin-associated linker polypeptide. *Arch. Microbiol.* **153**, 541–549 (1990).
22. Collier, J. L. & Grossman, A. R. A small polypeptide triggers complete degradation of light-harvesting phycobiliproteins in nutrient-deprived cyanobacteria. *EMBO J.* **13**, 1039–1047 (1994).
23. Dolganov, N. & Grossman, A. R. A polypeptide with similarity to phycocyanin alpha-subunit phycocyanobilin lyase involved in degradation of phycobilisomes. *J. Bacteriol.* **181**, 610–617 (1999).
24. Volz, K. in *Two-Component Signal Transduction* (eds Hoch, J. A. & Silhavy, T. J.) 53–64 (American Society for Microbiology, Washington DC, 1995).
25. Goto-Seki, A., Shirokane, M., Masuda, S., Tanaka, K. & Takahashi, H. Specificity crosstalk among group 1 and group 2 sigma factors in the cyanobacterium *Synechococcus* sp PCC7942: *in vitro* specificity and a phylogenetic analysis. *Mol. Microbiol.* **34**, 473–484 (1999).
26. Brahamsha, B. A genetic manipulation system for oceanic cyanobacteria of the genus *Synechococcus*. *Appl. Environ. Microbiol.* **62**, 1747–1751 (1996).
27. Badger, J. H. & Olsen, G. J. CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16**, 512–524 (1999).
28. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
29. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequences. *Nucleic Acids Res.* **25**, 955–964 (1997).
30. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W:. improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank the members of the *Prochlorococcus* annotation team: G. Rocap, S. W. Chisholm, D. Lindell, N. Algren, M. Coleman, W. Hess, A. Post, S. Shaw, C. Steglich, C. Ting, M. Sullivan, A. Tolonen, Z. Johnson and E. Zinser. We also thank T. Lane for discussions about carbonic anhydrases. This research was funded by the Biological and Environmental Research Program and the US Department of Energy's Office of Science. Sequencing was carried out and managed at the Joint Genome Institute. Computational analysis was performed at Oak Ridge National Laboratory, managed by UT-BATTELLE for the US Department of Energy. Additional support was provided by a DOE grant to B.P, B.B. and I.P., and an NSF grant to B.B. F.P. and A.D. were supported by the EC program Margenes, and by the Region Bretagne.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.P. (bpalenik@ucsd.edu). The sequence for the chromosome of *Synechococcus* sp. strain WH8102 is deposited in GenBank under accession number BX548020.

Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation

Gabrielle Rocap[1], Frank W. Larimer[2,3], Jane Lamerdin[3], Stephanie Malfatti[3], Patrick Chain[3,4], Nathan A. Ahlgren[1], Andrae Arellano[3], Maureen Coleman[5], Loren Hauser[2,3], Wolfgang R. Hess[9]*, Zackary I. Johnson[5], Miriam Land[2,3], Debbie Lindell[5], Anton F. Post[10], Warren Regala[3], Manesh Shah[2,3], Stephanie L. Shaw[6]*, Claudia Steglich[9], Matthew B. Sullivan[7], Claire S. Ting[8], Andrew Tolonen[7], Eric A. Webb[11], Erik R. Zinser[5] & Sallie W. Chisholm[5,8]

[1]*School of Oceanography, University Of Washington, Seattle, Washington 98195, USA*
[2]*Computational Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[3]*Joint Genome Institute, Walnut Creek, California 94598, USA*
[4]*Lawrence Livermore National Laboratory, Livermore, California 94550, USA*
[5]*Department of Civil and Environmental Engineering,* [6]*Department of Earth, Atmospheric and Planetary Sciences,* [7]*Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution, and* [8]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[9]*Institute of Biology, Humboldt-University, D-10115 Berlin, Germany*
[10]*Interuniversity Institute of Marine Science, 88103 Eilat, Israel*
[11]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA*

* Present addresses: Ocean Genome Legacy, Beverly, Massachusetts 01915, USA (W.R.H.); Department of Environmental Science Policy and Management, University of California, Berkeley, California 94720, USA (S.L.S.)

The marine unicellular cyanobacterium *Prochlorococcus* is the smallest-known oxygen-evolving autotroph[1]. It numerically dominates the phytoplankton in the tropical and subtropical oceans[2,3], and is responsible for a significant fraction of global photosynthesis. Here we compare the genomes of two *Prochlorococcus* strains that span the largest evolutionary distance within the *Prochlorococcus* lineage[4] and that have different minimum, maximum and optimal light intensities for growth[5]. The high-light-adapted ecotype has the smallest genome (1,657,990 base pairs, 1,716 genes) of any known oxygenic phototroph, whereas the genome of its low-light-adapted counterpart is significantly larger, at 2,410,873 base pairs (2,275 genes). The comparative architectures of these two strains reveal dynamic genomes that are constantly changing in response to myriad selection pressures. Although the two strains have 1,350 genes in common, a significant number are not shared, and these have been differentially retained from the common ancestor, or acquired through duplication or lateral transfer. Some of these genes have obvious roles in determining the relative fitness of the ecotypes in response to key environmental variables, and hence in regulating their distribution and abundance in the oceans.

As an oxyphototroph, *Prochlorococcus* requires only light, CO_2 and inorganic nutrients, thus the opportunities for extensive niche differentiation are not immediately obvious—particularly in view of the high mixing potential in the marine environment (Fig. 1a). Yet co-occurring *Prochlorococcus* cells that differ in their ribosomal DNA sequence by less than 3% have different optimal light intensities for growth[6], pigment contents[7], light-harvesting efficiencies[5], sensitivities to trace metals[8], nitrogen usage abilities[9] and cyanophage specificities[10] (Fig. 1b, c). These 'ecotypes'—distinct genetic lineages with ecologically relevant physiological differences—would be lumped together as a single species on the basis of their rDNA similarity[11], yet they have markedly different distributions within a stratified oceanic water column, with high-

Table 1 **General features of two *Prochlorococcus* genomes**

Genome feature	MED4	MIT9313
Length (bp)	1,657,990	2,410,873
G+C content (%)	30.8	50.7
Protein coding (%)	88	82
Protein coding genes	1,716	2,275
With assigned function	1,134	1,366
Conserved hypothetical	502	709
Hypothetical	80	197
Genes with orthologue in:		
Prochlorococcus MED4	–	1,352
Prochlorococcus MIT9313	1,352	–
Synechococcus WH8102	1,394	1,710
Genes without orthologue in:		
MED4 and WH8102	–	527
MIT9313 and WH8102	284	–
Transfer RNA	37	43
Ribosomal RNA operons	1	2
Other structural RNAs	3	3

light-adapted ecotypes most abundant in surface waters, and their low-light-adapted counterparts dominating deeper waters[12] (Fig. 1a). The detailed comparison between the genomes of two *Prochlorococcus* ecotypes we report here reveals many of the genetic foundations for the observed differences in their physiologies and vertical niche partitioning, and together with the genome of their close relative *Synechococcus*[13], helps to elucidate the key factors that regulate species diversity, and the resulting biogeochemical cycles, in today's oceans.

The genome of *Prochlorococcus* MED4, a high-light-adapted strain, is 1,657,990 base pairs (bp). This is the smallest of any oxygenic phototroph—significantly smaller than that of the low-light-adapted strain MIT9313 (2,410,873 bp; Table 1). The genomes of MED4 and MIT9313 consist of a single circular chromosome (Supplementary Fig. 1), and encode 1,716 and 2,275 genes respectively, roughly 65% of which can be assigned a functional category (Supplementary Fig. 2). Both genomes have undergone numerous large and small-scale rearrangements but they retain conservation of local gene order (Fig. 2). Break points between the orthologous gene clusters are commonly flanked by transfer RNAs, suggesting that these genes serve as loci for rearrangements caused by internal homologous recombination or phage integration events.

The strains have 1,352 genes in common, all but 38 of which are also shared with *Synechococcus* WH8102 (ref. 13). Many of the 38 '*Prochlorococcus* -specific' genes encode proteins involved in the atypical light-harvesting complex of *Prochlorococcus*, which contains divinyl chlorophylls *a* and *b* rather than the phycobilisomes that characterize most cyanobacteria. They include genes encoding the chlorophyll *a*/*b*-binding proteins (*pcb*)[14], a putative chlorophyll *a* oxygenase, which could synthesize (divinyl) chlorophyll *b* from (divinyl) chlorophyll *a*[15], and a lycopene epsilon cyclase involved in the synthesis of alpha carotene[16]. This remarkably low number of 'genera defining' genes illustrates how differences in a few gene families can translate into significant niche differentiation among closely related microbes.

MED4 has 364 genes without an orthologue in MIT9313, whereas MIT9313 has 923 that are not present in MED4. These strain-specific genes, which are dispersed throughout the chromosome (Fig. 2), clearly hold clues about the relative fitness of the two strains under different environmental conditions. Almost half of the 923 MIT9313-specific genes are in fact present in *Synechococcus* WH8102, suggesting that they have been lost from MED4 in the course of genome reduction. Lateral transfer events, perhaps

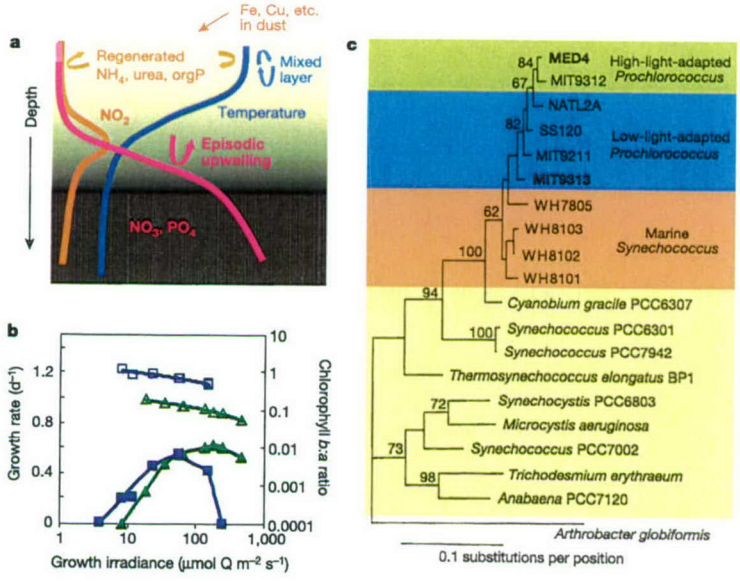


Figure 1 Ecology, physiology and phylogeny of *Prochlorococcus* ecotypes. **a**, Schematic stratified open-ocean water column illustrating vertical gradients allowing niche differentiation. Shading represents degree of light penetration. Temperature and salinity gradients provide a mixing barrier, isolating the low-nutrient/high-light surface layer from the high-nutrient/low-light deep waters. Photosynthesis in surface waters is driven primarily by rapidly regenerated nutrients, punctuated by episodic upwelling. **b**, Growth rate (filled symbols) and chlorophyll *b*:*a* ratio (open symbols) as a function of growth irradiance for MED4 (ref. 7) (green) and MIT9313 (ref. 6) (blue). **c**, Relationships between *Prochlorococcus* and other cyanobacteria inferred using 16S rDNA.

mediated by phage[10], may also be a source of some of the strain-specific genes (Supplementary Figs 3–6).

Gene loss has played a major role in defining the *Prochlorococcus* photosynthetic apparatus. MED4 and MIT9313 are missing many of the genes encoding phycobilisome structural proteins and enzymes involved in phycobilin biosynthesis[15]. Although some of these genes remain, and are functional[17], others seem to be evolving rapidly within the *Prochlorococcus* lineage[18]. Selective genome reduction can also be seen in the photosynthetic reaction centre of *Prochlorococcus*. Light acclimation in cyanobacteria often involves differential expression of multiple, but distinct, copies of genes encoding photosystem II D1 and D2 reaction centre proteins (*psbA* and *psbD* respectively)[19]. However, MED4 has a single *psbA* gene, MIT9313 has two that encode identical photosystem II D1 polypeptides, and both possess only one *psbD* gene, suggesting a diminished ability to photoacclimate. MED4 has also lost the gene encoding cytochrome *c*550 (*psbV*), which has a crucial role in the oxygen-evolving complex in *Synechocystis* PCC6803 (ref. 20).

There are several differences between the genomes that help account for the different light optima of the two strains. For example, the smaller MED4 genome has more than twice as many genes (22 compared with 9) encoding putative high-light-inducible proteins, which seem to have arisen at least in part through duplication events[15]. MED4 also possesses a photolyase gene that has been lost in MIT9313, probably because there is little selective pressure to retain ultraviolet damage repair in low light habitats. Regarding differences in light-harvesting efficiencies, it is noteworthy that MED4 contains only a single gene encoding the chlorophyll *a*/*b*-binding antenna protein Pcb, whereas MIT9313 possesses two copies. The second type has been found exclusively in low-light-adapted strains[21], and may form an antenna capable of binding more chlorophyll pigments.

Both strains have a low proportion of genes involved in regulatory functions. Compared with the freshwater cyanobacterium *Thermosynechococcus elongatus* (genome size <2.6 megabases)[22], MIT9313 has fewer sigma factors, transcriptional regulators and two-component sensor-kinase systems, and MED4 is even more reduced (Supplementary Table 1). The circadian clock genes provide an example of this reduction as both genomes lack several components (*pex*, *kaiA*) found in the model *Synechococcus* PCC7942 (ref. 23). However, genes for the core clock proteins (*kaiB*, *kaiC*) remain in both genomes, and *Prochlorococcus* cell division is tightly synchronized to the diel light/dark cycle[24]. Thus, loss of some circadian components may imply an alternative signalling pathway for circadian control.

Gene loss may also have a role in the lower percentage of G+C content of MED4 (30.8%) compared with that of MIT9313 (50.74%), which is more typical of marine *Synechococcus*. MED4 lacks genes for several DNA repair pathways including recombinational repair (*recJ*, *recQ*) and damage reversal (*mutT*). Particularly, the loss of the base excision repair gene *mutY*, which removes adenosines incorrectly paired with oxidatively damaged guanine residues, may imply an increased rate of G•C to T•A transversions[25]. The tRNA complement of MED4 is largely identical to MIT9313 and is not optimized for a low percentage G+C genome, suggesting that it is not evolving as fast as codon usage.

Analysis of the nitrogen acquisition capabilities of the two strains points to a sequential decay in the capacity to use nitrate and nitrite during the evolution of the *Prochlorococcus* lineage (Fig. 3a). In *Synechococcus* WH8102—representing the presumed ancestral state—many nitrogen acquisition and assimilation genes are grouped together (Fig. 3a). MIT9313 has lost a 25-gene cluster, which includes genes encoding the nitrate/nitrite transporter and nitrate reductase. The nitrite reductase gene has been retained in MIT9313, but it is flanked by a proteobacterial-like nitrite transporter rather than a typical cyanobacterial nitrate/nitrite permease (Supplementary Fig. 4), suggesting acquisition by lateral gene transfer. An additional deletion event occurred in MED4, in which the nitrite reductase gene was also lost (Fig. 3a). As a result of these serial deletion events MIT9313 cannot use nitrate, and MED4 cannot use nitrate or nitrite[9]. Thus each *Prochlorococcus* ecotype uses the N species that is most prevalent at the light levels to which they are best adapted: ammonium in the surface waters and nitrite at depth (Fig. 1a). *Synechococcus*, which is the only one of the three that has nitrate reductase, is able to bloom when nitrate is upwelled (Fig. 1a), as occurs in the spring in the North Atlantic[3] and the north Red Sea[26].

The two *Prochlorococcus* strains are also less versatile in their organic N usage capabilities than *Synechococcus* WH8102 (ref. 13). MED4 contains the genes necessary for usage of urea, cyanate and oligopeptides, but no monomeric amino acid transporters have been identified. In contrast, MIT9313 contains transporters for urea, amino acids and oligopeptides but lacks the genes necessary for cyanate usage (cyanate transporter and cyanate lyase) (Fig. 3a). As expected, both genomes contain the high-affinity ammonium transporter *amt1* and both lack the nitrogenase genes essential for nitrogen fixation. Finally, both contain the nitrogen transcriptional regulator encoded by *ntcA* and there are numerous genes in both genomes, including *ntcA*, *amt1*, the urea transport and GS/GOGAT genes (glutamine synthetase and glutamate synthase, both involved in ammonia assimilation), with an upstream NtcA-binding-site consensus sequence.

The genomes also have differences in genes involved in phosphorus usage that have obvious ecological implications. MED4, but not MIT9313, is capable of growth on organic P sources (L. R. Moore and S.W.C., unpublished data), and organic P can be the prevalent form of P in high-light surface waters[27]. This difference may be due to the acquisition of an alkaline phosphatase-like gene in MED4 (Supplementary Fig. 5). Both genomes contain the high-affinity phosphate transport system encoded by *pstS* and *pstABC*[28], but MIT9313 contains an additional copy of the phosphate-binding component *pstS*, perhaps reflecting an increased reliance on orthophosphate in deeper waters. MED4 contains

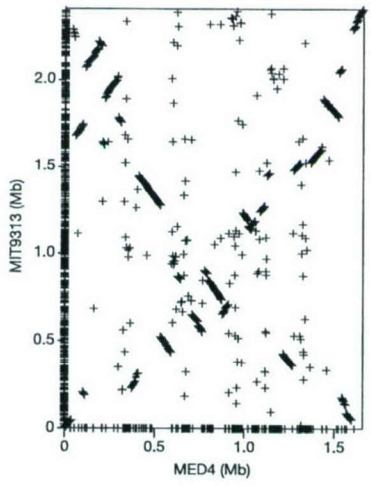


Figure 2 Global genome alignment as seen from start positions of orthologous genes. Genes present in one genome but not the other are shown on the axes. The 'broken X' pattern has been noted before for closely related bacterial genomes, and is probably due to multiple inversions centred around the origin of replication. Alternating slopes of many adjacent gene clusters indicate that multiple smaller-scale inversions have also occurred.

several P-related regulatory genes including the *phoB*, *phoR* two-component system and the transcriptional activator *ptrA*. In MIT9313, however, *phoR* is interrupted by two frameshifts and *ptrA* is further degenerated, suggesting that this strain has lost the ability to regulate gene expression in response to changing P levels.

Both *Prochlorococcus* strains have iron-related genes that are missing in *Synechococcus* WH8102, which may explain its dominance in the iron-limited equatorial Pacific[2]. These genes include flavodoxin (*isiB*), an Fe-free electron transfer protein capable of replacing ferredoxin, and ferritin (located with the ATPase component of an iron ABC transporter), an iron-binding molecule implicated in iron storage. Additional characteristics of the iron acquisition system in these genomes include: an Fe-induced transcriptional regulator (Fur) that represses iron uptake genes; numerous genes with an upstream putative *fur* box motif that are candidates for a high-affinity iron scavenging system; and absence of genes involved in Fe–siderophore complexes.

Prochlorococcus does not use typical cyanobacterial genes for inorganic carbon concentration or fixation. Both genomes contain a sodium/bicarbonate symporter but lack homologues to known families of carbonic anhydrases, suggesting that an as yet unidentified gene is fulfilling this function. One of the two carbonic anhydrases in *Synechococcus* WH8102 was lost in the deletion event that led to the loss of the nitrate reductase (Fig. 3a); the other is located next to a tRNA and seems to have been lost during a genome rearrangement event. Similar to other *Prochlorococcus* and marine *Synechococcus*, MED4 and MIT9313 possess a form IA ribulose-1,5-bisphosphate carboxylase/oxygenase, rather than the typical cyanobacterial form IB. The ribulose-1,5-bisphosphate carboxylase/oxygenase genes are adjacent to genes encoding structural carboxysome shell proteins and all have phylogenetic affinity to genes in the γ-proteobacterium *Acidithiobacillus ferroxidans*[15], suggesting lateral transfer of the extended operon.

Prochlorococcus has been identified in deep suboxic zones where it is unlikely that they can sustain themselves by photosynthesis alone[29], thus we looked for genomic evidence of heterotrophic capability. Indeed, the presence of oligopeptide transporters in both genomes, and the larger proportion of transporters (including some sugar transporters) in the MIT9313 strain-specific genes (Supplementary Fig. 2), suggests the potential for partial hetero-

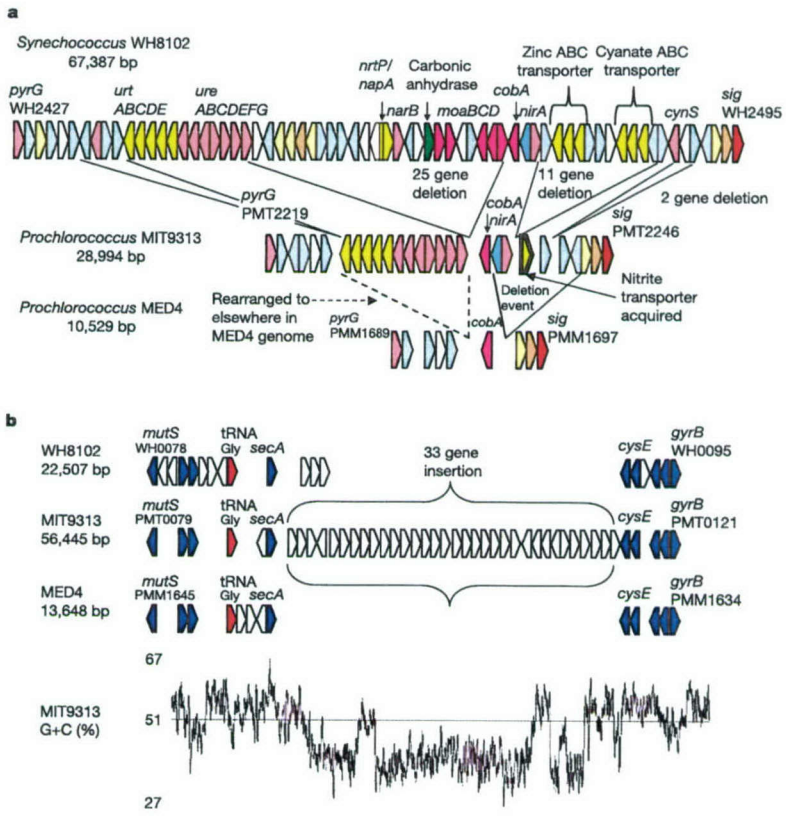


Figure 3 Dynamic architecture of marine cyanobacterial genomes. **a**, Deletion, acquisition and rearrangement of nitrogen usage genes. In MIT9313, 25 genes including the nitrate/nitrite transporter (*nrtP*/*napA*), nitrate reductase (*narB*) and carbonic anhydrase have been deleted. The cyanate transporter and cyanate lyase (*cynS*) were probably lost after the divergence of MIT9313 from the rest of the *Prochlorococcus* lineage, as MED4 possesses these genes. MIT9313 has retained nitrite reductase (*nirA*) and acquired a nitrite transporter. In MED4 *nirA* has been lost and the urea transporter (*urt* cluster) and urease (*ure* cluster) genes have been rearranged (dotted line). Genes in different functional categories are colour-coded to guide the eye. **b**, Lateral transfer of genes involved in lipopolysaccharide biosynthesis including sugar transferases, sugar epimerases, modifying enzymes and two pairs of ABC-type transporters. Blue, genes in all three genomes; pink, genes hypothesized to have been laterally transferred; red, tRNAs; white, other genes. The percentage of G + C content in MIT9313 along this segment is lower (42%) than the whole-genome average (horizontal line).

trophy. However, neither genome contains known pathways that would allow for complete heterotrophy. They are both missing genes for steps in the tricarboxylic acid cycle, including 2-oxoglutarate dehydrogenase, succinyl-CoA synthetase and succinyl-CoA-acetoacetate-CoA transferase.

Cell surface chemistry has a major role in phage recognition and grazing by protists and thus is probably under intense selective pressure in nature. The two *Prochlorococcus* genomes and the *Synechococcus* WH8102 genome show evidence of extensive lateral gene transfer and deletion events of genes involved in lipopolysaccharide and/or surface polysaccharide biosynthesis, reinforcing the role of predation pressures in the creation and maintenance of microdiversity. For example, MIT9313 has a 41.8-kilobase (kb) cluster of surface polysaccharide genes (Fig. 3b), which has a lower percentage G+C composition (42%) than the genome as a whole, implicating acquisition by lateral gene transfer. MED4 has acquired a 74.5-kb cluster consisting of 67 potential surface polysaccharide genes (Supplementary Fig. 6a) and has lost another cluster of surface polysaccharide biosynthesis genes shared between MIT9313 and *Synechococcus* WH8102 (Supplementary Fig. 6b).

The approach we have taken in describing these genomes highlights the known drivers of niche partitioning of these closely related organisms (Fig. 1). Detailed comparisons with the genomes of additional strains, such as *Prochlorococcus* SS120 (ref. 30), will enrich this story, and the analysis of whole genomes from *in situ* populations will be necessary to understand the full expanse of genomic diversity in this group. The genes of unknown function in all of these genomes hold important clues for undiscovered niche dimensions in the marine pelagic zone. As we unveil their function we will undoubtedly learn that the suite of selective pressures that shape these communities is much larger than we have imagined. Finally, it may be useful to view *Prochlorococcus* and *Synechococcus* as important 'minimal life units', as the information in their roughly 2,000 genes is sufficient to create globally abundant biomass from solar energy and inorganic compounds. □

Methods

Genome sequencing and assembly

DNA was isolated from the clonal, axenic strain MED4 and the clonal strain MIT9313 essentially as described previously[4]. The two whole-genome shotgun libraries were obtained by fragmenting genomic DNA using mechanical shearing and cloning 2–3-kb fragments into pUC18. Double-ended plasmid sequencing reactions were carried out using PE BigDye Terminator chemistry (Perkin Elmer) and sequencing ladders were resolved on PE 377 Automated DNA Sequencers (Perkin Elmer). The whole-genome sequence of *Prochlorococcus* MED4 was obtained from 27,065 end sequences (7.3-fold redundancy), whereas *Prochlorococcus* MIT9313 was sequenced to ×6.2 coverage (33,383 end sequences). For *Prochlorococcus* MIT9313, supplemental sequencing (×0.05 sequence coverage) of a pFos1 fosmid library was used as a scaffold. Sequence assembly was accomplished using PHRAP (P. Green). All gaps were closed by primer walking on gap-spanning library clones or PCR products. The final assembly of *Prochlorococcus* MED4 was verified by long-range genomic PCR reactions, whereas the assembly of *Prochlorococcus* MIT9313 was confirmed by comparison to the fosmid clones, which were fingerprinted with *Eco*RI. No plasmids were detected in the course of genome sequencing, and insertion sequences, repeated elements, transposons and prophages are notably absent from both genomes. The likely origin of replication in each genome was identified based on G+C skew, and base pair 1 was designated adjacent to the *dnaN* gene.

Genome annotation

The combination of three gene-modelling programs, Critica, Glimmer and Generation, were used in the determination of potential open reading frames and were checked manually. A revised gene/protein set was searched against the KEGG GENES, Pfam, PROSITE, PRINTS, ProDom, COGs and CyanoBase databases, in addition to BLASTP against the non-redundant peptide sequence database from GenBank. From these results, categorizations were developed using the KEGG and COGs hierarchies, as modified in CyanoBase. Manual annotation of open reading frames was done in conjunction with the *Synechococcus* team. The three-way genome comparison was used to refine predicted start sites, add additional open reading frames and standardize the annotation across the three genomes.

Genome comparisons

The comparative genome architecture of MED4 and MIT9313 was visualized using the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/). Orthologues were determined by aligning the predicted coding sequences of each gene with the coding sequences of the other genome using BLASTP. Genes were considered orthologues if each was the best hit of the other one and both *e*-values were less than e^{-10}. In addition, bidirectional best hits with *e*-values less than e^{-8} and small proteins of conserved function were manually examined and added to the orthologue lists.

Phylogenetic analyses used PAUP*, logdet distances and minimum evolution as the objective function. The degree of support at each node was evaluated using 1,000 bootstrap resamplings. Ribosomal DNA analyses used 1,160 positions. The Gram-positive bacterium *Arthrobacter globiformis* was used to root the tree.

Received 11 May; accepted 25 July 2003; doi:10.1038/nature01947.
Published online 13 August 2003.

1. Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334,** 340–343 (1988).
2. Campbell, L., Liu, H., Nolla, H. & Vaulot, D. Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep Sea Res. I* **44,** 167–192 (1997).
3. DuRand, M. D., Olson, R. J. & Chisholm, S. W. Phytoplankton population dynamics at the Bermuda Atlantic Time-series Station in the Sargasso Sea. *Deep Sea Res. II* **48,** 1983–2003 (2001).
4. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S rDNA internal transcribed spacer (ITS) sequences. *Appl. Environ. Microbiol.* **68,** 1180–1191 (2002).
5. Moore, L. R. & Chisholm, S. W. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.* **44,** 628–638 (1999).
6. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393,** 464–467 (1998).
7. Moore, L. R., Goericke, R. E. & Chisholm, S. W. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar. Ecol. Prog. Ser.* **116,** 259–275 (1995).
8. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47,** 976–988 (2002).
9. Moore, L. R., Post, A. F., Rocap, G. & Chisholm, S. W. Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* **47,** 989–996 (2002).
10. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424,** 1047–1051 (2003).
11. Hagström, Å. *et al.* Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Appl. Environ. Microbiol.* **68,** 3628–3633 (2002).
12. West, N. J. *et al.* Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147,** 1731–1744 (2001).
13. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424,** 1037–1042 (2003).
14. La Roche, J. *et al.* Independent evolution of the prochlorophyte and green plant chlorophyll a/b light harvesting proteins. *Proc. Natl Acad. Sci. USA* **93,** 15244–15248 (1996).
15. Hess, W. *et al.* The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. *Photosynth. Res.* **70,** 53–71 (2001).
16. Stickforth, P., Steiger, S., Hess, W. R. & Sandmann, G. A novel type of lycopene ε-cyclase in the marine cyanobacterium *Prochlorococcus marinus* MED4. *Arch. Microbiol.* **179,** 407–415 (2003).
17. Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *Plant Cell* **13,** 965–978 (2001).
18. Ting, C., Rocap, G., King, J. & Chisholm, S. W. Phycobiliprotein genes of the marine prokaryote *Prochlorococcus*: Evidence for rapid evolution of genetic heterogeneity. *Microbiology* **147,** 3171–3182 (2001).
19. Golden, S. S., Brusslan, J. & Haselkorn, R. Expression of a family of *psbA* genes encoding a photosystem II polypeptide in the cyanobacterium *Anacystis nidulans* R2. *EMBO J.* **5,** 2789–2798 (1986).
20. Shen, J. R., Qian, M., Inoue, Y. L. & Burnap, R. L. Functional characterization of *Synechocystis* sp. PCC 6803 Δ*psbU* and Δ*psbV* mutants reveals important roles of cytochrome c-550 in cyanobacterial oxygen evolution. *Biochemistry* **37,** 1551–1558 (1998).
21. Garczarek, L., van der Staay, G. W. M., Hess, W. R., Le Gall, F. & Partensky, F. Expression and phylogeny of the multiple antenna genes of the low-light-adapted strain *Prochlorococcus marinus* SS120 (Oxyphotobacteria). *Plant Mol. Biol.* **46,** 683–693 (2001).
22. Nakamura, Y. *et al.* Complete genome structure of the thermophilic Cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res.* **9,** 123–130 (2002).
23. Ishiura, M. *et al.* Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobateria. *Science* **281,** 1519–1523 (1998).
24. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the Equatorial Pacific Ocean. *Science* **268,** 1480–1482 (1995).
25. Michaels, M. L. & Miller, J. H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-Hydroxyguanine (7,8-Dihydro-8-Oxoguanine). *J. Bacteriol.* **174,** 6321–6325 (1992).
26. Lindell, D. & Post, A. F. Ultraphytoplankton succession is triggered by deep winter mixing in the Gulf of Aqaba (Eilat). *Red Sea. Limnol. Oceanogr.* **40,** 1130–1141 (1995).
27. Karl, D. M., Bidigare, R. R. & Letelier, R. M. Long-term changes in plankton community structure and productivity in the North Pacific Subtropical Gyre: The domain shift hypothesis. *Deep Sea Res. II* **48,** 1449–1470 (2001).
28. Scanlan, D. J., Mann, N. H. & Carr, N. G. The response of the picoplanktonic marine cyanobacterium *Synechococcus* species WH7803 to phosphate starvation involves a protein homologous to the periplasmic phosphate binding protein of *Escherichia coli*. *Mol. Microbiol.* **10,** 181–191 (1993).
29. Johnson, Z. *et al.* Energetics and growth kinetics of a deep *Prochlorococcus* spp. population in the Arabian Sea. *Deep Sea Res. II* **46,** 1719–1743 (1999).
30. Dufresne, A. *et al.* Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a near minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA* (in the press).

Supplementary Information accompanies the paper on **www.nature.com/nature.**

Acknowledgements This research was funded by the Biological and Environmental Research Program of the US Department of Energy's Office of Science. The Joint Genome Institute managed the overall sequencing effort. Genome finishing was carried out under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory. Computational annotation was carried out at the Oak Ridge National Laboratory, managed by UT-BATTELLE for the US Department of Energy. Additional support was provided by the DOE, NSF and the Seaver Foundation to S.W.C., the Israel–US Binational Science Foundation to A.F.P. and S.W.C., and FP5-Margenes to W.R.H. and A.F.P. We thank the *Synechococcus* WH8102 annotators (B. Palenik, B. Brahamsha, J. McCarren, E. Allen, F. Partensky, A. Dufresne and I. Paulsen) for their help with curating the *Prochlorococcus* genomes and E. V. Armbrust and L. Moore for critical reading of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu). The complete nucleotide sequences and sequences of predicted open reading frames have been deposited in the EMBL/GenBank/DDBJ databases under accession numbers BX548174 (MED4) and BX548175 (MIT9313).

Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*

Matthew B. Sullivan[1], John B. Waterbury[3] & Sallie W. Chisholm[2]

[1]*MIT/Woods Hole Oceanographic Institution Joint Program in Biological Oceanography, and* [2]*Department of Civil and Environmental Engineering and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Woods Hole Oceanographic Institution, Department of Biology, Woods Hole, Massachusetts 02543, USA*

***Prochlorococcus* is the numerically dominant phototroph in the tropical and subtropical oceans, accounting for half of the photosynthetic biomass in some areas[1,2]. Here we report the isolation of cyanophages that infect *Prochlorococcus*, and show that although some are host-strain-specific, others cross-infect with closely related marine *Synechococcus* as well as between high-light- and low-light-adapted *Prochlorococcus* isolates, suggesting a mechanism for horizontal gene transfer. High-light-adapted *Prochlorococcus* hosts yielded *Podoviridae* exclusively, which were extremely host-specific, whereas low-light-adapted *Prochlorococcus* and all strains of *Synechococcus* yielded primarily *Myoviridae*, which has a broad host range. Finally, both *Prochlorococcus* and *Synechococcus* strain-specific cyanophage titres were low ($<10^3$ ml^{-1}) in stratified oligotrophic waters even where total cyanobacterial abundances were high ($>10^5$ cells ml^{-1}). These low titres in areas of high total host cell abundance seem to be a feature of open ocean ecosystems. We hypothesize that gradients in cyanobacterial population diversity, growth rates, and/or the incidence of lysogeny underlie these trends.**

Phages are thought to evolve by the exchange of genes drawn from a common gene pool through differential access imposed by host range limitations[3]. Similarly, horizontal gene transfer, important in microbial evolution[4,5], can be mediated by phages[6] and is probably responsible for many of the differences in the genomes of closely related microbes[5]. Recent detailed analyses of molecular phylogenies constructed for marine *Prochlorococcus* and *Synechococcus*[7,8] (Fig. 1) show that these genera form a single group within the marine picophytoplankton clade[9] (>96% identity in 16S ribosomal DNA sequences), yet display microdiversity in the form of ten well-defined subgroups[8]. We have used members of these two groups to study whether phage isolated on a particular host strain cross-infect other hosts, and if so, whether the probability of cross-infection is related to rDNA-based evolutionary distance between the hosts.

Analyses of host range were conducted (Fig. 1) with 44 cyanophages, isolated as previously described[10] from a variety of water depths and locations (see Supplementary Information) using 20 different host strains chosen to represent the genetic diversity of *Prochlorococcus* and *Synechococcus*[8]. Although we did not examine how these patterns would change if phage were propagated on different hosts, this would undoubtedly add another layer of complexity due to host range modifications as a result of methylation of phage DNA[6]. Similar to those that infect other marine bacteria[11] and *Synechococcus*[10–14], our *Prochlorococcus* cyanophage isolates fell into three morphological families: *Myoviridae*, *Siphoviridae* and *Podoviridae*[15].

As would be predicted[10–14], *Podoviridae* were extremely host specific with only two cross-infections out of a possible 300 (Fig. 1). Similarly, the two *Siphoviridae* isolated were specific to their hosts. In instances of extreme host specificity, *in situ* host abundance would need to be high enough to facilitate phage–host contact. It is noteworthy in this regard that members of the high-light-adapted *Prochlorococcus* cluster, which yielded the most host-specific cyanophage, have high relative abundances *in situ*[16]. The *Myoviridae* exhibited much broader host ranges, with 102 cross-infections out of a possible 539. They not only cross-infected among and between *Prochlorococcus* ecotypes but also between *Prochlorococcus* and *Synechococcus*. Those isolated with *Synechococcus* host strains have broader host ranges and are more likely to cross-infect low-light-adapted than high-light-adapted *Prochlorococcus* strains. The low-light-adapted *Prochlorococcus* are less diverged from *Synechococcus* than high-light-adapted *Prochlorococcus*[7,8], suggesting a relationship, in this instance, between the probability of cross-infection and rDNA relatedness of hosts. Finally, we tested the *Myoviridae* for cross-infection against marine bacterial isolates closely related to *Pseudoalteromonas*, which are known to be broadly susceptible to diverse bacteriophages (bacterial strains HER1320, HER1321, HER1327, HER1328)[11]. None of the *Myoviridae* cyanophages infected these bacteria.

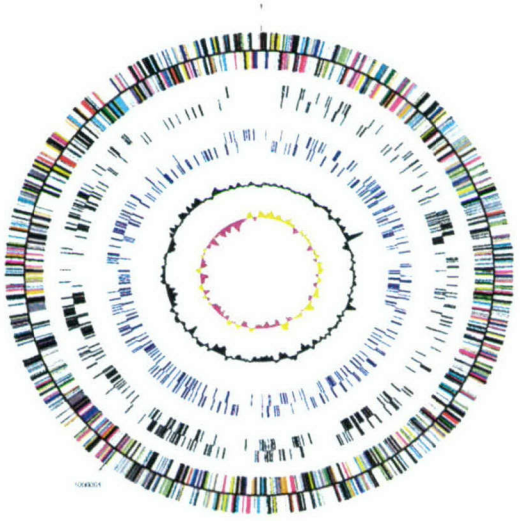
Phage morphotypes isolated were determined, to some degree, by the host used for isolation (Fig. 1). For example, ten of ten cyanophages isolated using high-light-adapted *Prochlorococcus* strains were *Podoviridae*. In contrast, all but two cyanophages isolated on *Synechococcus* were *Myoviridae*, a bias that has been reported by others[14], and over half of those isolated on low-light-adapted *Prochlorococcus* belonged to this morphotype. We further substantiated these trends by examining lysates (as opposed to plaque-purified isolates) from a range of host strains, geographic locations and depths—of 58 *Synechococcus* lysates 93% contained *Myoviridae*, of 43 low-light-adapted *Prochlorococcus* lysates 65% contained *Myoviridae*, and of 107 high-light-adapted *Prochlorococcus* lysates 98% contained *Podoviridae* (see Supplementary Information).

Maximum cyanophage titres, using a variety of *Synechococcus* hosts, are usually found to be within an order of magnitude of the total *Synechococcus* abundance[10,14,17,18], and can be as high as 10^6 phage ml^{-1}. One study[17] has shown, for example, that along a transect in which total *Synechococcus* abundance decreased from 10^5 cells ml^{-1} to 250 cells ml^{-1}, maximum cyanophage titres remained at least as high as the total number of *Synechococcus*. We wondered whether titres of *Prochlorococcus* cyanophage in the Sargasso Sea, where *Prochlorococcus* cells are abundant (10^5 cells ml^{-1}), would be comparable to those measured in coastal oceans for *Synechococcus* where total *Synechococcus* host abundances are of similar magnitude. We assayed cyanophage titres in a depth profile in the Sargasso Sea at the end of seasonal stratification using 11 strains of *Prochlorococcus* (Fig. 2), choosing at least one host strain from each of the six phylogenetic clusters that span the rDNA-based genetic diversity of our culture collection[8].

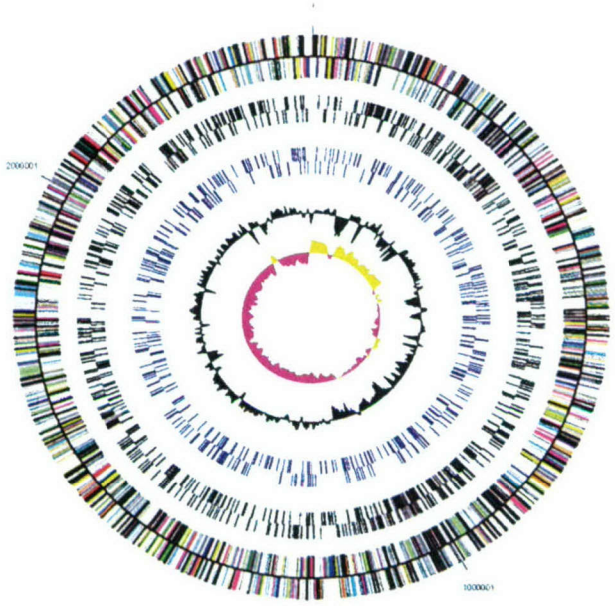
Three *Prochlorococcus* host strains (MIT 9303, MIT 9313 and SS120) yielded low or no cyanophage. Other hosts yielded titres

Chisholm Supplementary Figure 1

a

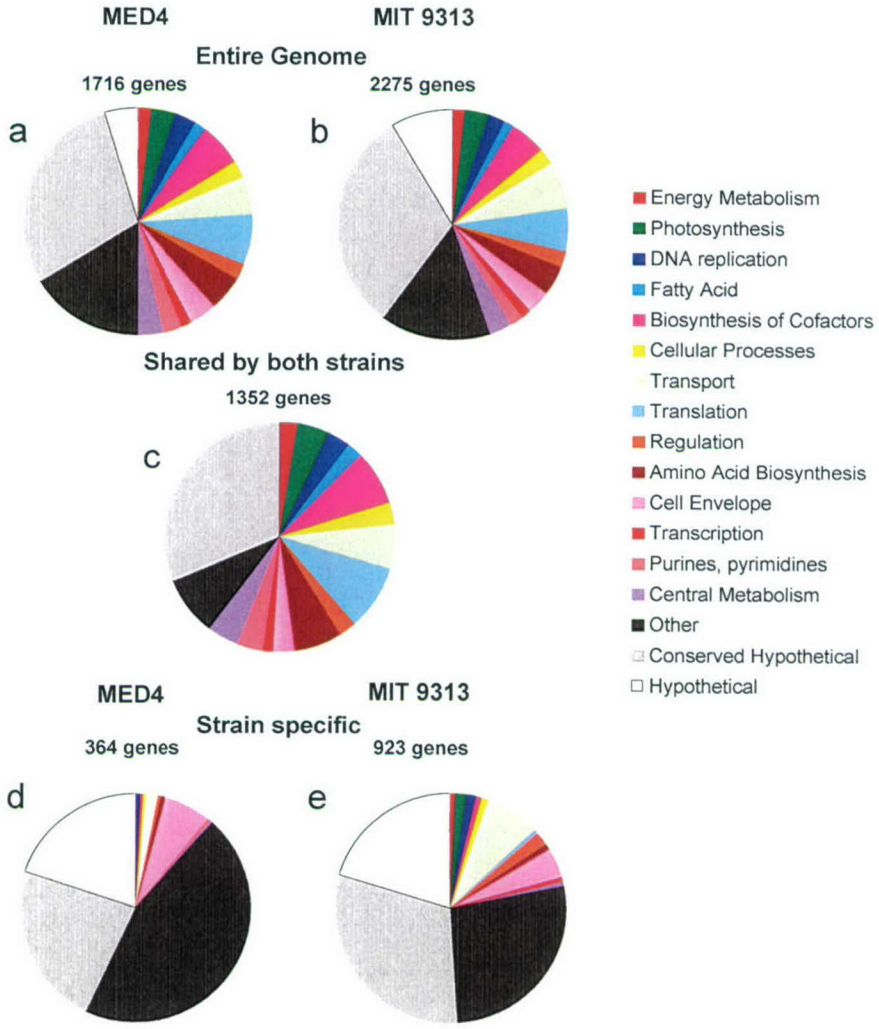


b



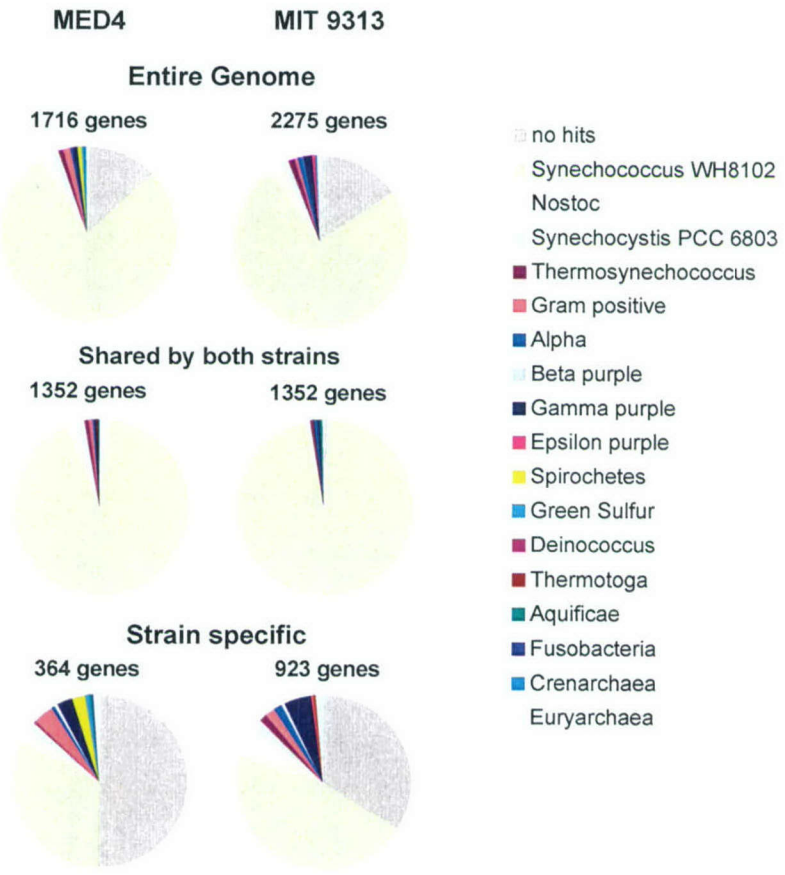
Supp. Figure 1. Circular representation of the *Prochlorococcus* genomes. **a,** MED4. **b,** MIT 9313. For both genomes outermost circles (1 and 2) are predicted protein coding regions on the plus and minus strands, respectively. Color coding is as in Supplementary Figure 2. The next two circles show genes not present in the other *Prochlorococcus* genome on the plus (circle 3) and minus (circle 4) strands. Circles 5 and 6 show genes on the plus and minus strands, respectively that contain transmembrane domains. Circle 7 is % G+C content (deviation from average). Innermost circle (8) represents the GC skew curve.

Chisholm Supplementary Figure 2



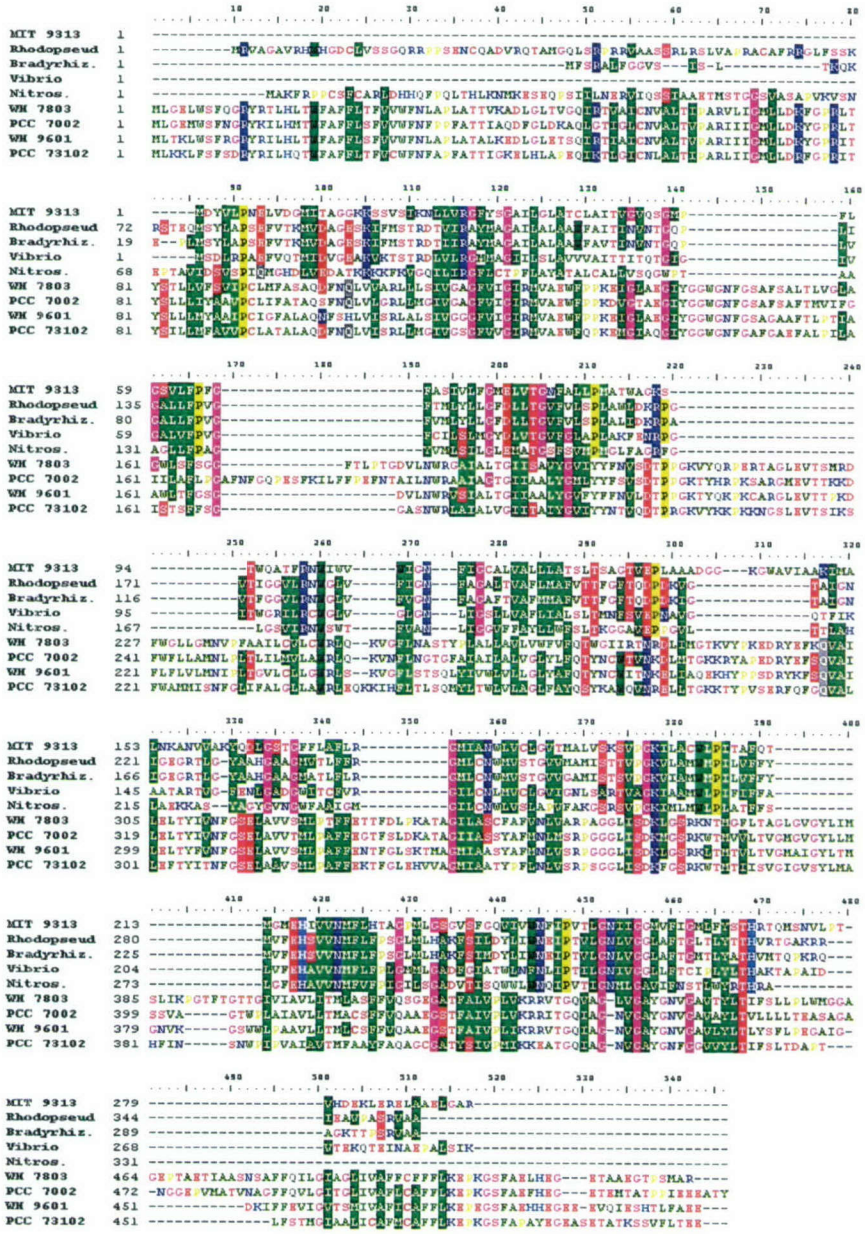
Supp. Figure 2. Functional categorization of predicted open reading frames in the *Prochlorococcus* genomes, following the classification scheme used by CyanoBase. **a,** MED4, entire genome. **b,** MIT 9313, entire genome. **C,** Genes present in both MED4 and MIT 9313. **d,** Genes in MED4 not present in MIT 9313. **e,** Genes in MIT 9313 not present in MED4.

Chisholm Supplementary Figure 3



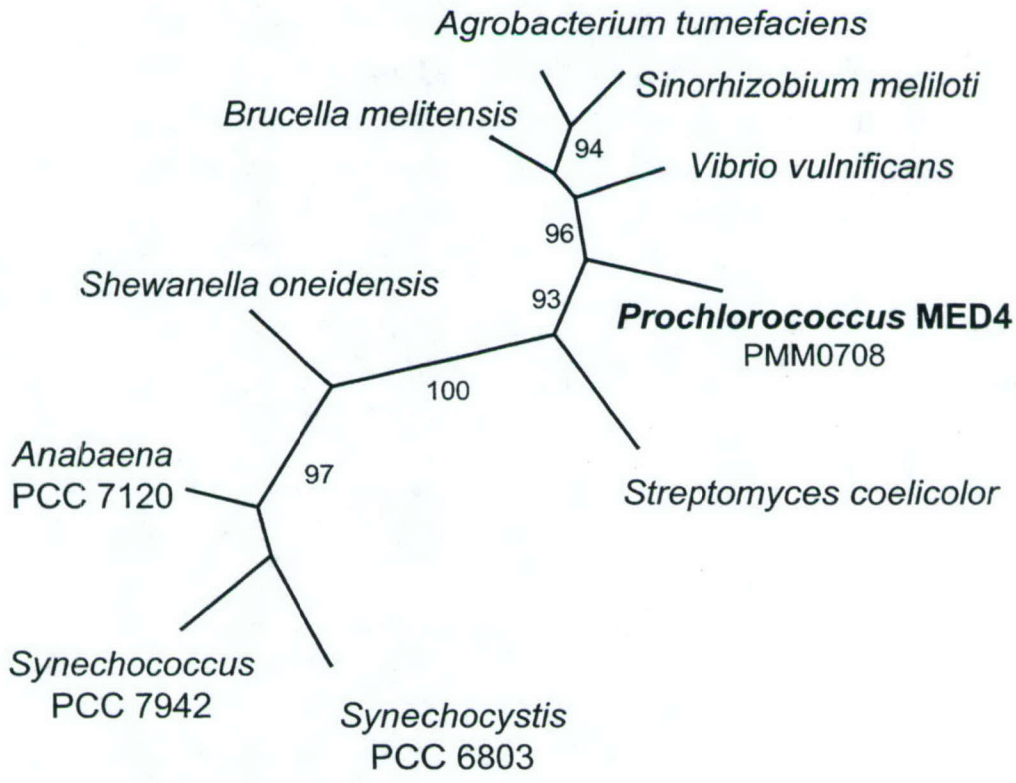
Supp. Figure 3. Comparison of *Prochlorococcus* MED4 and MIT 9313 open reading frames with those of other complete prokaryotic genomes. The predicted coding sequences of each gene in both genomes were aligned with the coding sequences of 90 bacterial genomes using BLASTP. Significant alignments were defined as having an e-value less than 10^{-6}. The bacterial genomes comprised the 89 completed bacterial genomes available from ftp.ncbi.nih.gov/genbank/genomes/Bacteria on 30 October 2002 and *Synechococcus* WH 8102[8]. **a,** MED4, entire genome. **B,** MIT 9313, entire genome. **c,** MED4 genes present in MIT 9313 **c,** MIT 9313 genes present in MED4 **e,** Genes in MED4 not present in MIT 9313 **f,** Genes in MIT 9313 not present in MED4.

Chisholm Supplementary Figure 4



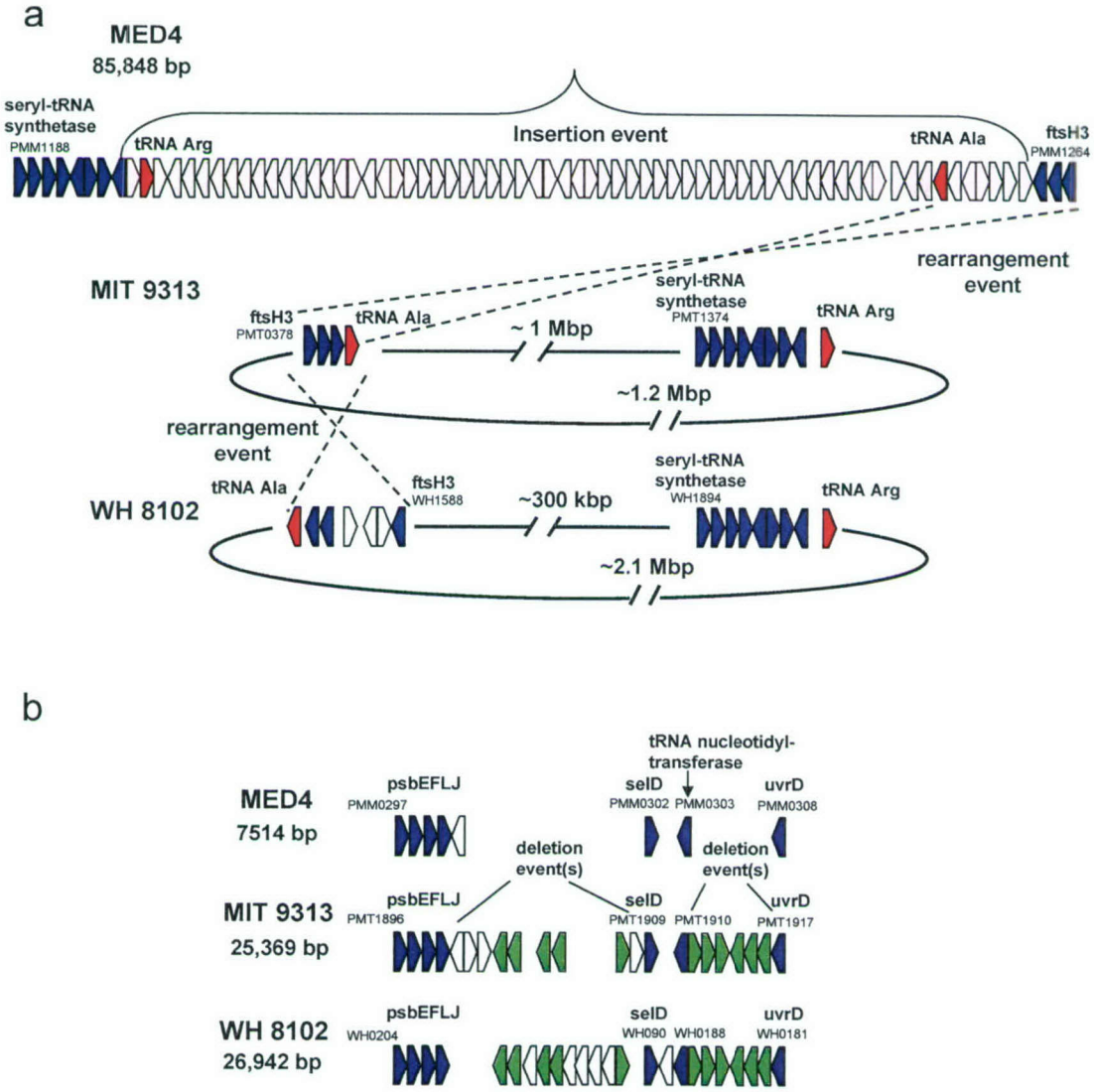
Supp. Figure 4 Alignment of the putative nitrite transporter in *Prochlorococcus* MIT9313 (PMT2240) with its most significant matches in the NR database (all proteobacteria) and with cyanobacterial nitrate/nitrate transporters. The MIT 9313 gene has a formate/nitrite transporter domain (Pfam PF01226) in contrast to the cyanobacterial nitrate transporters which are permeases of the major facilitator superfamily (Pfam PF00083). Furthermore, the MIT 9313 gene has no significant matches (BLASTP evalue < e-2) in the genomes of *Prochlorococcus* MED4, *Synechococcus* WH8102, *Synechocystis* sp. PCC 6803, *Thermosynechococcus elongatus* BP-1, or *Anabaena* sp. PCC 7120 suggesting it may have been acquired via lateral gene transfer. Alignment generated using ClustalW. Shaded residues indicate >50% similarity. Abbreviations and accession numbers as follows: Rhodopseud., *Rhodopseudomonas palustris* (ZP_00012718.1); Bradyrhiz., *Bradyrhizobium japonicum* (NP_769441); Vibrio, *Vibrio vulnificus* (NP_762336.1); Nitros., *Nitrosomonas europaea* (NP_840759); WH 7803, *Synechococcus* WH 7803 napA (AAG45172); PCC 7002, *Synechococcus* PCC 7002 nrtP (AAD45941); WH9601, *Trichodesmium* WH 9601 napA (AAF00917); PCC 73102, *Nostoc punctiforme* PCC 73102 (ZP_00107423).

Chisholm Supplementary Figure 5



Supp. Figure 5 Phylogenetic tree showing the relationship of a possible alkaline phosphatase like gene in *Prochlorococcus* MED4 (PMM0708) with the most significant matches in the NR database, which include several proteobacterial sequences, and with the atypical alkaline phosphatase of *Synechococcus* PCC 7942 and related cyanobacterial genes. Accession numbers as follows: *Brucella melitensis* (NP_541633.1), *Agrobacterium tumefaciens* str. C58 (NP_531956.1); *Sinorhizobium meliloti*, (NP_385365.1); *Vibrio vulnificus* (NP_762849.1), *Streptomyces coelicolor* A3(2) (NP_624650.1), Shewanella oneidensis MR-1 (NP_717877.1) *Anabaena* PCC 7102 (NP_489331.1), *Synechocystis* sp. PCC 6803 (NP_440276); *Synechococcus* sp. PCC 7942 (A47026).

Chisholm Supplementary Figure 6



Supp. Figure 6 Insertions, deletions and rearrangements of genes involved in lipopolysaccharide biosynthesis (LPS clusters) in MED4. Color coding is as follows: blue, orthologous genes present in all three genomes; pink, genes hypothesized to be part of lateral transfer events, many have roles in LPS biosynthesis; red, tRNAs; green, orthologous genes present in two genomes, many have roles in LPS biosynthesis; white, other genes. Length in bp represents the size of the region shown for each genome. **a,** Insertion of a 74.5 kbp cluster of LPS genes in MED4, roughly between two tRNAs. The 67 potential surface polysaccharide genes in this cluster include sugar transferases, sugar epimerases, and modifying enzymes such as aminotransferases, methyltransferases, carbamoyltransferases, and acetyltransferases. In MIT 9313 and WH 8102 the genes that flank this insertion are rearranged to other parts of the genome. **b,** Deletion of LPS biosynthesis genes in MED4. LPS related genes present in MIT 9313 and WH 8102, several of which have homologs in the acquired genes shown in part a, have been deleted. In this region a selenophosphate synthase (*selD*) and a tRNA nucleotidyl-transferase in the center of the cluster have been retained suggesting that they are essential genes and separate deletion events have occurred on either side of them.

Supp. Table 1 Number of predicted signal transduction and transcription factors suggests reduced regulatory capacity in *Prochlorococcus*

	MED4	MIT 9313	*T. elongatus*
Sigma Factors	5	8	8
Two Component systems			
Histidine Kinases	4	5	17
Response regulators	6	8	27
Ser/Thr protein Kinases	0	1	11
Transcription Factors			
LuxR family	2	5	4
LysR family	1	1	3
CRP family	3	4	3
ArsR family	1	2	2
FUR family	2	3	3
Other	2	3	3
Light sensors/transducers			
Cryptochrome	2	0	2
Bacteriophytochrome	0	0	5
Phototropin	0	0	1

APPENDIX D

Marine phage genomics

Paul, J. H., Sullivan, M. B., Segall, A. M., and Rohwer, F. (2002). Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133**(4), 463-76.

ELSEVIER

Comparative Biochemistry and Physiology Part B 133 (2002) 463–476

CBP

www.elsevier.com/locate/cbpb

Review

Marine phage genomics☆

John H. Paul[a,*], Matthew B. Sullivan[b], Anca M. Segall[c], Forest Rohwer[c]

[a]*College of Marine Sciences, University of South Florida, 140 Seventh Ave S., St. Petersburg, FL, 33701 USA*
[b]*Massachusetts Institute of Technology/Woods Hole Oceanographic Institution Joint Program in Biological Oceanography, Cambridge, MA, USA*
[c]*Biology Department, San Diego State University, San Diego, CA, USA*

Received 7 May 2002; received in revised form 3 September 2002; accepted 4 September 2002

Abstract

Marine phages are the most abundant biological entities in the oceans. They play important roles in carbon cycling through marine food webs, gene transfer by transduction and conversion of hosts by lysogeny. The handful of marine phage genomes that have been sequenced to date, along with prophages in marine bacterial genomes, and partial sequencing of uncultivated phages are yielding glimpses of the tremendous diversity and physiological potential of the marine phage community. Common gene modules in diverse phages are providing the information necessary to make evolutionary comparisons. Finally, deciphering phage genomes is providing clues about the adaptive response of phages and their hosts to environmental cues.

Keywords: Phages; Genomics; Lysogeny; Roseophage; *Synechococcus*; *Prochlorococcus*

1. Introduction

Direct counts show that there are ~3–10 virus-like particles for every cell in the marine environment (Bergh et al., 1989; reviewed in Wommack and Colwell, 2000 and Fuhrman, 1999). Bacteria and Archaea are the most common cells in seawater, and it is believed that most of the viral-like particles are phages that prey upon these prokaryotes. Since the oceans are the world's largest biosphere, marine phages are probably the most abundant biological entities on the planet. Through their lytic activities, phages modulate carbon flow through microbial food webs by attacking both autotrophic and heterotrophic microbes (reviewed in Fuhrman, 1999). As prophages, marine phages may also confer a wide range of traits to their hosts including: immunity to superinfection (Hershey, 1971); toxin production (Waldor and Mekalanos, 1996); and the capability to transfer modular blocks of genes (Jiang and Paul, 1998a; Paul, 1999).

Even though 'The Age of Genomics' was heralded by the sequencing of phage *Escherichia coli* phi 174 in 1977 (Sanger et al., 1977), there are only three completed marine phage genomes currently in GenBank. This number will undoubtedly increase over the next decade. Here, we review the current state of the field of marine phage genomics and argue that these genomes, because of their small size, offer unprecedented opportunities for exploring eco-genomics, testing evolu-

☆ Contribution to a special issue of CBP on Comparative Functional Genomics.

*Corresponding author. Tel.: +1-727-553-1168; fax: +1-727-553-1189.
E-mail address: jpaul@seas.marine.usf.edu (J.H. Paul).

1096-4959/02/$ - see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S1096-4959(02)00168-9

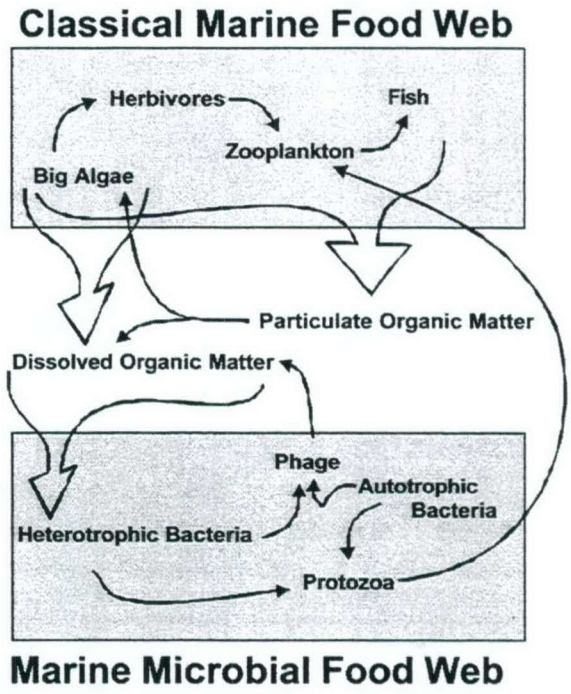


Fig. 1. Phage and the cycling of organic carbon matter through marine food webs. Arrows indicate direction of organic carbon flow.

tionary models and understanding genetic transduction within the environment.

2. Overview of marine phage ecology

2.1. Phage effects on carbon flow

The influences of phages on ecosystem dynamics are best understood in the marine environment. The marine microbial food web (MMFW) is the consortium of heterotrophic and autotrophic prokaryotes, as well as their predators, that inhabit the Earth's oceans and seas (Azam, 1998). The MMFW regulates the transfer of energy and nutrients to higher trophic levels and greatly influences global carbon (C) and nutrient cycles (Pomeroy, 1974; Azam et al., 1983). Dissolved organic matter (DOM) is the largest biogenic sink of carbon in the ocean (Kennish, 2001). Because the DOM pool is so large, heterotrophic bacterial populations are not resource limited; instead, they are controlled by predation (Fuhrman and Noble, 1995). The two predator guilds responsible for top-down control of the MMFW are the protozoa and phages (Fuhrman and Noble, 1995). In near-shore waters each of these predator guilds accounts for 50% of the microbial mortality each day (Fuhrman and Noble, 1995). To put the effects of these two bacterial predator guilds into perspective, ~49.3 Gt of C is fixed by phytoplankton per year in the world's oceans (Field et al., 1998), while global marine bacterial production is estimated to be 26–70 Gt of C per year (Wilhelm and Suttle, 1999). Therefore, the majority of the marine-biotic C is cycled into microbes and most of these microbes are killed by protozoa and phage predators.

When bacteria are eaten by protozoa, there is a possibility that the carbon can be transferred to the larger members of the marine food web (Fig. 1; Fukami et al., 1999). In contrast, when a bacterium is killed by a lytic phage, both the lysed host cell and the phage become part of the DOM pool (Middelboe et al., 1992, 1996). Since DOM is only utilized by other heterotrophic bacteria that

are also susceptible to lytic phages, this carbon never leaves the MMFW. The more rapidly this cycle repeats itself, the greater the amount of respiratory CO_2 that is produced, leaving less organic carbon stored in the world's oceans (Fig. 1). Thus phage activity may prove troublesome to proposed efforts to fertilize the oceans for increased carbon sequestration and fisheries yields.

In the open ocean, *Prochlorococcus* and *Synechococcus* are the numerically dominant autotrophs (reviewed in Waterbury et al., 1986; Partensky et al., 1999). Phages that infect these important primary producers have been isolated (Suttle, 1993; Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson et al., 1993; Sullivan, 2001). Although ecological studies of the impact of cyanophage on *Synechococcus* communities suggest that direct mortality is low relative to that observed for heterotrophic bacteria (reviewed in Suttle, 2000), the impact of cyanophage pressures on population structure and diversity of these systems may be significant (Waterbury and Valois, 1993). In near-shore communities, viruses have been isolated that infect the major 'large' phytoplankton species (e.g. eukaryotic diatoms and dinoflagellates). While viruses are not thought to be the major agent of cyanobacterial mortality, virus-induced mortality may be responsible for the 'sudden crashes' that terminate many blooms of eukaryotic algae (Sieburth et al., 1988; Bratbak, 1993; Nagasaki et al., 1993; Bratbak et al., 1995, 1998).

2.2. Transduction in marine environments

Besides their enormous influence on marine biogeochemistry, phages have important effects on genetic exchange in the marine environment (Jiang and Paul, 1998a). Phages can mediate DNA exchange between different bacteria by transduction, which occurs when host DNA is accidentally packaged into the phage during assembly (Masters, 1996; Weisberg, 1996). When the mispackaged phage infects another bacterium, instead of injecting phage DNA, it transfers DNA from its former host. Jiang and Paul (1998a) estimated that 1.3×10^{14} transduction events occur per year in the Tampa Bay Estuary, Florida. Extrapolation suggests that marine phages transduce 10^{28} base pairs of DNA per year in the world's oceans.

2.3. Lysogeny in marine environments

Not all phage-host encounters lead to host cell lysis, many rather result in lysogeny or pseudolysogeny (Ackermann and DuBow, 1987). Through meticulous work at the single cell level with *Bacillus*, lysogeny was first described as 'the hereditary power to produce bacteriophage' (Lwoff, 1953). This 'hereditary power' is due to the integration of invading phage DNA into the host cell genome (now termed a prophage) rather than proceeding through the lytic pathway. The prophage will remain integrated in the host cell genome until it is induced to 'abandon ship' and proceed through the lytic pathway. The molecular mechanism underlying prophage integration and excision are well understood in model systems (Hershey, 1971). In contrast, pseudolysogeny is a poorly understood phenomenon. It is often invoked to describe conditions where constant phage production occurs in the presence of a high abundance of host cells, thus allowing large numbers of host cells and their phage to coexist. Two mechanisms that might explain such observations are the following: (1) a mixture of sensitive and resistant host cells; or (2) a mixture of temperate and virulent phages (Williamson et al., 2001).

Lysogeny has been shown to improve the general fitness of the host (Edlin et al., 1975), largely from lysogenic conversion, or the expression of prophage-encoded genes. A common lysogenic conversion phenotype is immunity to superinfection (Hershey, 1971), but lysogenic conversion can also result in altered structural characteristics (Pruzzo and Satta, 1988; Vaca-Pacheco et al., 1999; Mirold et al., 2001), as well as resistance to antibiotics (Mlynarczyk et al., 1997) and reactive oxygen species (Figueroa-Bossi and Bossi, 1999). Of particular importance and global significance is the spread of toxin/virulence genes (often termed 'pathogenicity islands') by lysogenic conversion. Diphtheria, botulinum, cholera, pertussis, shiga and many other exotoxins are prophage encoded (reviewed in Davis and Waldor, 2002). Recent evidence suggests that these toxin genes can be transferred by transduction and other lateral gene transfer mechanisms (Boyd and Waldor, 1999; Faruque et al., 1999; Yaron et al., 2000).

Lysogeny is a common phenomenon in the marine environment. A recent study suggests that lysogeny in oligotrophic waters is common amongst cultivated bacteria, as 40% of 110 marine

bacterial isolates produced phage or bacteriocin-like particles upon treatment with an inducing agent (Jiang and Paul, 1998b). Efforts to quantify lysogeny in natural populations have resulted in a wide range of values for the proportion of the population lysogenized. For example, Weinbauer and Suttle (1996) found that in the Gulf of Mexico, 1.5–11.4% of the microbial population was lysogenized, whereas a detailed seasonal study of lysogeny in Tampa Bay indicated that the lysogenic fraction could range from 0 to 100% (average 27.6±37.1%; Williamson et al., 2002). During the seasonal study, lysogeny was primarily detected in winter months, consistent with the theory that lysogeny is favored in times of low host cell density. The environmental factors that lead to the control of lysogeny in the marine environment are largely unknown although links to nutrients such as phosphate have been suggested (Tuomi et al., 1995; Wilson et al., 1996, 1997). The molecular control of lysogeny in many phage host systems is complex (Ptashne, 1992; Friedman and Court, 2001) and usually involves genomic elements termed lysogeny modules (Lucchini et al., 1999). Essentially, nothing is known of the molecular or environmental control of these genomic elements in marine phages. In addition to investigations of bacterial lysogens in the environment, two recent studies have suggested that natural populations of the cyanobacterium *Synechococcus* can be lysogenized (McDaniel et al., 2002; Ortmann et al., 2002).

2.4. Microbial and phage diversity in the marine environment

Through their role as species-specific predators, phages may also help maintain microbial diversity. In the absence of host cell resistance and providing that contact rates remain high, lytic phages could potentially lyse all individuals of a species—thus phage attack can result in a rapid succession of microbial species (Thingstad and Lignell, 1997; Wommack and Colwell, 2000). Experimental evidence that phage exert a strong selective pressure on microbial populations comes from host-range analysis of phage isolates and the observation that very closely related bacterial species and even strains of the same species are infected by different phages (Moebus, 1991; Suttle and Chan, 1993, 1994; Waterbury and Valois, 1993).

The biodiversity of marine phage is essentially unknown. The few studies that have addressed this question suggest that diversity is high (reviewed in Borsheim, 1993). Moebus and colleagues (Moebus, 1991, 1992a,b Moebus and Nattkemper, 1991) screened over 900 isolates of culturable marine bacteria and found that approximately one-third were susceptible to at least one, and often multiple, lytic phages. Over the course of these studies, the authors concluded that: (1) the majority of bacterial strains were probably susceptible to phage infection; and (2) the phages isolated in these studies were specific to single hosts. Further work characterized a subset of these phages using DNA–DNA hybridizations, %GC, genome size estimations and host-range analysis and showed that they were genetically diverse (Wichels et al., 1998). Kellogg et al. (1995) isolated 60 phages from Florida and Hawaii that infected *Vibrio parahaemolyticus* and analyzed them by restriction fragment length polymorphism (RFLP), host specificity and Southern blotting. RFLP analysis separated the 60 phage isolates into six distinct groups that were then further genetically characterized through Southern blotting using a 1.5-kb DNA probe cloned from one of the isolates. Host range analysis showed that these phages were host-specific as none of the 60 isolates were able to infect closely related *Vibrio* species. Together these findings strongly suggest that phage diversity is at least as high, and probably higher, than the diversity of bacteria. That is, for each bacterium there is probably at least one, and often multiple, phages capable of infecting it and each phage is usually specific to a single microbial species or strain.

3. The current state of marine phage genomics

3.1. Cultured marine phage genomes

The first marine phage genome to be completely sequenced, *Pseudoaltermonas espejiana* BAL-31 phi PM2, was isolated off the coast of Chile in the 1960s (Espejo and Canelo, 1968). Phage PM2 was also the first lipid-containing phage ever isolated and it serves as the type phage for the International Committee of Viral Taxonomy (ICTV) family *Corticoviridae* (Murphy et al., 1995). The genome of phi PM2 is circular and 10 079 bp long that is replicated in a rolling circle fashion (Mannisto et al., 1999). The genes that encode structural and replication proteins have

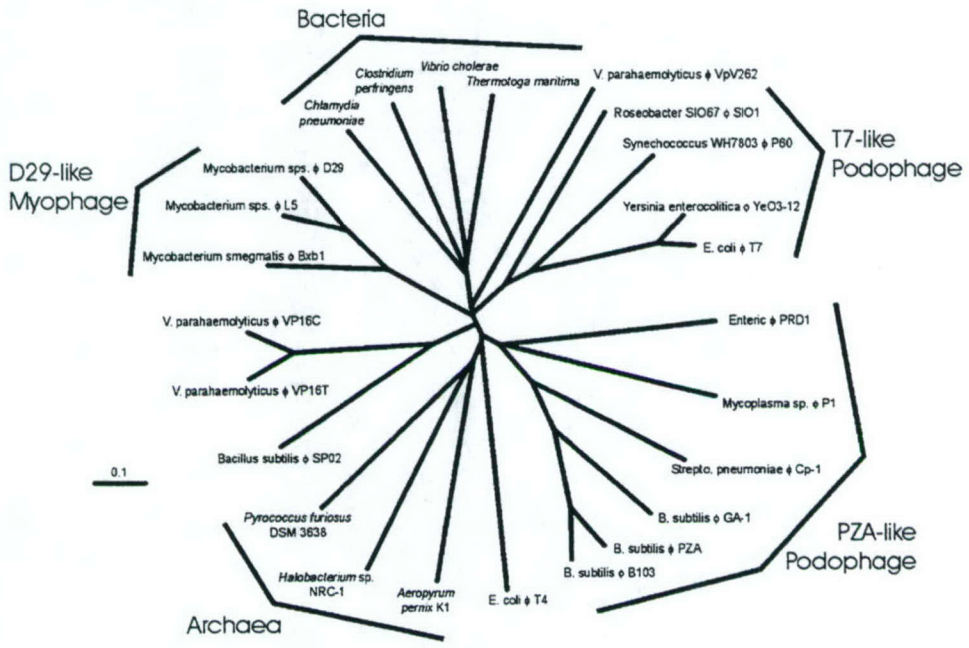


Fig. 2. Phylogenetic trees of DNA polymerase genes found in bacteria, Archaea and various phages of marine and terrestrial origin. Marine phage DNA polymerase sequences include *V. parahaemolyticus* ϕVpV262, ϕVP16C and ϕVP16T; *Roseobacter* SIO67 ϕSIO1 and Synechococcus WH7803 ϕP60.

been identified amongst the 42 potential open reading frames (ORFs) (Mannisto et al., 1999). The phi PM2 genome also contains a plasmid-like maintenance region, suggesting that the genome may be transferred among bacteria either as a plasmid or as a free phage (Mannisto et al., 1999). Essentially nothing is known about the ecology of phi PM2.

The second marine phage genome to be sequenced, *Roseobacter* SIO67 phi SIO1, was isolated off the Scripps Institution of Oceanography pier in 1990 (Rohwer et al., 2000). Phage SIO1 has a 39 906 bp dsDNA genome with 34 predicted ORFs (Rohwer et al., 2000). The phi SIO1 primase/helicase, DNA polymerase and endodeoxyribonuclease 1 share significant similarity to presumed homologs in *Escherichia coli* phi T3 and T7 (Rohwer et al., 2000). The *Synechococcus* WH7803 phi P60 genome (Chen and Lu, in press), discussed below also contains ORFs with significant BLAST hits to these genes. This suggests that there is a group of marine phage that uses a DNA replication mechanism much like phi T3 and T7 (Fig. 2). The ORFs that probably encode the phi SIO1 structural proteins, as suggested by their position in the genome, do not share significant similarity to other phage proteins currently in GenBank, but may be related to sequences found in the marine phages *Vibrio parahaemolyticus* phi VpV 262 (Hardies et al., 2002). The phi SIO1 genome also contains a number of tantalizing hints about its ecology. In particular, it appears the phage has linked its lifecycle to the phosphate metabolism of its host cell (see below).

The first cyanophage genome to be completely sequenced, *Synechococcus* WH7803 phi P60, has a 47 872 bp dsDNA genome that contains 80 potential ORFs (Chen and Lu, in press). The DNA replication genes are related to those of phi SIO1 and phi T7. Interestingly, the phi P60 DNA polymerase gene appears to be more closely related to those encoded by the non-marine phage T3, T7 and phi-YeO3-12 than to the marine phi SIO1

Table 1
The cultured marine phage that have completed genome sequences

Phage	Genome Size (kb) and morphology	Genome access and refs.
Pseudoalteromonas	10.1	NC 000867
espejiana phi PM2	Corticovirus	(Mannisto et al. 1999)
Roseobacter SIO67 phi	39.9	NC 002519
SIO1	Podovirus	(Rohwer et al. 2000)
Synechococcus WH7803	47.8	AF338467
phi P60	Podovirus	(Chen and Lu, in press)
Vibrio parahaemolyticus	49.7	
phi TB16T	Myovirus	
Vibrio parahaemolyticus	47.5	
phi TB16C	Myovirus	
Vibrio parahaemolyticus	45.9	http://biochem.uthscsa.edu/~hs_lab/phage.html
phi VpV 262	Podovirus	(Hardies et al., 2002)

Currently there are only three marine phage genomes available in GenBank. Three other genomes have been sequenced and are currently being annotated.

gene (Chen and Lu, in press). Phage P60 and phi SIO1 also encode a ribonucleotide reductase that is probably involved in recycling host rNTPs to dNTPs that can be incorporated into nascent phage genomes. Phage P60 encodes a RNA polymerase, which appears to be absent from the phi SIO1 genome. Since the RNA polymerase is essential to invasion and transcription of T7 (Calendar, 1988), phi P60 and phi SIO1 must have very different lifecycles.

Three marine phages that infect two strains of the human pathogen *Vibrio parahaemolyticus* have also been completely sequenced (Table 1). *Vibrio parahaemolyticus* phi VpV 262 is a Podovirus isolated from the Strait of Georgia, BC, Canada (Hardies and Serwer, 2002). The genome is 45 874 bp linear dsDNA genome with 73 predicted ORFs. The genome contains DNA polymerase, primase and helicase genes that appear to be more closely related to their bacterial homologues than to the phi T7 or phi SIO1 genes.

Two other Vibriophages, phi TB16T and phi TB16C, were separated from each other from a lysate of phage VP16, isolated from Tampa Bay (Kellogg et al., 1995). Although the original phage was thought to be a myovirus based on its contractile tail (Kellogg et al., 1995), the DNA sequence similarities and presence of a cos site suggest that the viruses are more likely to be siphoviruses (Segall and Rohwer, unpublished data). The two viruses are closely related (73–91%) over roughly 80% of their genomes, but differ by multiple unique insertions ranging between approximately 200 and 5000 bp in size. Each virus encodes 63-64 ORFs. The structural genes are clustered on the left side of the genome near the cos site and include a gene related to phage lambda's large terminase subunit, as well as genes related to those encoding tail and tail fiber proteins, portal, sheath and tape measure proteins of various phages and prophages. The rest of the genome includes ORFs similar to a DNA polymerase, a helicase, a polypeptide deformylase, and two genes weakly similar to transcriptional regulators. The two Vibriophages also contain numerous other genes similar to ORFs encoding hypothetical proteins from other phage, prophage and bacterial genomes. Although the original phage VP16 gave clear plaques, TB16T and TB16C were separated from each other based on their plaque morphology—TB16T gave turbid plaques with clear centers, whereas TB16C gave entirely clear plaques. Although lysogeny has not been proven, there is significant evidence from the genome sequence predicting that these phages have a temperate lifestyle—including the putative regulatory elements and the ability to circularize. Of great interest is that the sequences of the mixture of two genomes formed separate contigs during sequencing and assembly, despite substantial regions of identity between the two genomes (Rohwer and Segall, unpublished data). This suggests that even very closely related viruses can be sequenced from a mixed lysate, even when, as indicated by restriction digests, one genome makes up less than 5% of the DNA in the lysate. Subsequent sequencing of libraries made of DNA from the isolated phages showed that we did not

obtain any chimeric sequences between the two genomes.

As of this writing, the genomic sequencing of cyanophage S-PM2 is being finished (Nicholas H. Mann, personal communication). The genome is ~170 kb and includes several very large genes, including several that encode proteins >3000 amino acids in length. There are some rather surprising genes such as one that is homologous to a gene involved in complementary chromatic adaptation. When finished, this genome should be particularly exciting because it will be the first marine representative of the Myoviruses related to coliphage T4.

In addition to marine phages, a number of groups are starting to sequence marine viruses. *Emiliania huxleyi* is a marine coccolithophorid, with a world-wide distribution that forms vast coastal and mid-oceanic blooms (Holligan et al., 1993). Double stranded DNA viruses that infect *E. huxleyi* (EhV) have recently been isolated (Wilson et al., in press). Phylogenetic analysis of DNA polymerase gene fragments of these viruses suggests that EhVs belong to a new genus with the proposed name *Coccolithovirus*, within the family of algal viruses Phycodnaviridae (Schroeder et al., in press). Work is currently underway (shotgun sequencing completed, finishing begun) to sequence the 410-kb genome of one of these viruses with a completion date due in summer 2002 (William Wilson, personal communication). Tai et al. (2002) have also reported the complete sequence of a virus associated with lysis of the eukaryotic fish-killer *Heterosigma akashiwo*.

3.2. Prophage in completed marine bacterial genomes

In the near future, prophage contained in marine bacterial genomes will be an important data source when studying marine phage. The best studied prophage system in a marine bacterial genome is CTX-phi, the lysogenic phage that encodes the cholera toxin in pathogenic strains of *Vibrio cholerae* (Waldor and Mekalanos, 1996). CTX-phi is a temperate, filamentous phage that is secreted by the same extracellular protein secretion system used for cholera toxin production (Davis et al., 2000a). The phage uses a type IV pilus as a receptor that is encoded by an adjacent genetic element, the TCP gene cluster, itself a putative prophage (Karaolis et al., 1999). Three strains of CTX-phi have been characterized from classical, El Tor and Calcutta isolates of *V. cholerae*. Interestingly, the former strain encodes functional phage genes that are expressed, but lacks the ability to produce infectious virions due to their occurrence as single prophage elements (Davis et al., 2000b). In contrast, the latter two phage strains occur in tandem arrays of multiple prophages or single prophage plus prophage parts and these strains are capable of generating infectious virions (Davis and Waldor, 2000). The increasing numbers of publicly available microbial genomes have led to the discovery that prophage elements may exist in nearly every microbial genome (S. Casjens; personal communication). Putative prophage have been identified by inspecting regions of microbial chromosomes for the following characteristics: (1) genes possessing homology to known phage genes; (2) a contiguous group of genes containing few, if any, obviously non-phage genes; (3) the phage genes organized in a phage-like manner (e.g. integrase near the end, structural genes clustered, correctly ordered, etc.); and (4) 'unknown' genes in the putative prophage not obviously organized in a non-phage-like manner. Using this approach, we have identified putative prophages in the complete marine genomes of *Prochlorococcus* MED4, *Prochlorococcus* MIT 9313 and *Synechococcus* WH 8102 (http://www.jgi.doe.gov/). However, these regions were later shown not to be prophage regions as the phage genes were too few, were not organized in a phage-like manner and were interupted by obvious non-phage genes.

There is significant evidence that some prophages integrate into select regions of a host cell genome. Thirty-four of 58 cases of the integration of genetic elements (including prophage) were found to occur in attB sites within tRNA or tmRNA genes (Williams, 2002). Prophage integration at these sites might be beneficial because: (1) tRNA and tmRNA genes have ~four- to ninefold lower mutation rates than other protein encoding regions; and (2) these genes are small thus requiring a smaller region to be mimicked by the phage attP. Of particular interest to phage ecologists is the fact that tRNA promoters are known to be regulated by growth rate (Swenson et al., 1994)—in effect, allowing a prophage integrating into a tRNA gene to monitor the physiological state of its host through transcriptional coupling to the tRNA gene. It is unknown how integrase genes of genetic elements are able to recognize tRNA-like

elements, but Williams (2002) suggests that the definitive secondary structure of the DNA molecule may be involved.

To further aid in the identification of prophage in microbial genomes, we assessed the usefulness of DNA Structural Atlases (http://www.cbs.dtu.dk/services/GenomeAtlas/; Pedersen et al., 2000) as a diagnostic tool. These atlases display DNA structural characters, such as DNA curvature, DNA flexibility and DNA stability, in the form of a color-coded wheel that is useful for visually revealing interesting structural features of a genomic sequence (Pedersen et al., 2000). The five models used to predict these structural characters are based on either empirical data (e.g. DNase I sensitivity, X-ray crystallography data, trinucleotide preferences and gel mobility) or quantum mechanical calculations. To evaluate this tool for predicting prophage, we examined the DNA structural atlases of genomes containing well-characterized prophage elements to qualitatively look for diagnostic trends. Although there often appears to be significant structure within the prophage regions, this structure is not associated with all known prophage and often occurs throughout the genome in known non-prophage regions. However, factors such as the type of prophage and the length of time since its last activity might greatly affect these DNA structural properties and require a more intensive, quantitative analysis to identify diagnostic trends.

3.3. Uncultured marine phage genomes

Based upon pulsed-field gel electrophoresis of natural marine phage communities, we know that marine viral genomes fall into three size ranges: 35–40, 50–65 and 120–140 kb (Wommack et al., 1999; Steward et al., 2000). It has long been known that there is little similarity between the marine bacteria that have been cultivated and those phylotypes that are known to be prevalent as determined by 16S rDNA analyses (Fuhrman and Campbell, 1998). Therefore, cultured marine phage will most probably not be representative of the community. To circumvent the limitations imposed by culturing, a number of laboratories have started to sequence the genomes of total marine viral communities.

Breitbart et al. (in press) have constructed a shotgun library from an uncultured, near-shore marine viral community. Only 30% of the sequences from this library possessed appreciable similarity to those in the GenBank. Of the significant hits, 32% were phage in origin and 3% were most closely related to eukaryotic viruses. Among the phage genes showing similarity, Podovirus genes were most common (43%). Representatives of the Sipho- and Myoviridae were also found. These broad trends have now been observed in a second, near shore library (Breitbart and Rohwer, unpublished data). Another research group has partially sequenced a shotgun library made from a phage community from 70 m in Monterey Bay (Steward and Preston, unpublished data). As with the near-shore phage community, most of the sequences show no similarity to Genbank entries. Both libraries contained sequences with significant similarity to phage DNA polymerase genes, RNA polymerase, integrases, transposases and reverse transcriptases.

Breitbart et al. (in press) proposed a mathematical model that uses the number of observed contigs to predict phage richness and diversity in the sample. According to their calculations, the most abundant phage in the sample made up 4% of the population and phage diversity is very high (e.g. Shannon–Weaver index value of 7–8; Shannon and Weaver, 1963). This model also predicts that it is technically possible to sequence an entire marine viral community.

4. Uses of marine phage genomes

4.1. Classification of marine phages

A major goal of phage genomic sequencing projects should be to provide the information necessary to classify marine phage into guilds that reflect their biology. Current phage taxonomy relies on the morphological characteristics of the free phage particle as established by the International Committee on Taxonomy of Viruses (ICTV) (Murphy et al., 1995). The ICTV classification, however, provides very little information about the ecological niches or lifestyles of phage. Additionally, the ICTV system does not have sufficient resolution to address phage biodiversity questions, nor will it be useful for analyzing uncultured marine phage or prophage genomes. In response to these shortcomings, numerous groups are actively constructing phage taxonomical systems based on completed genomic sequences (Lawrence et al., 2002; Rohwer and Edwards, 2002). These

systems will help classify marine phages into families that provide information about their life-cycles and ecological roles, as well as identify phage types that deserve more detailed analyses.

Marine phage genomes are already helping to differentiate phages into operational taxonomic units (OTUs) that predict biological properties. For example, total genome analyses and individual DNA polymerase sequences show that Podoviruses belong to two OTUs with fundamentally different DNA replication mechanisms (Pecenkova and Paces, 1999; Chen and Lu, in press, Rohwer and Edwards, 2002). The first group, the phi T7-like Podophages, includes: phi T7; phi T3; phi P60; phi SIO1; and *Yersinia enterocolitica* phi Ye03-12 (Fig. 2; Rohwer and Edwards, 2002). This group of T7-like phages replicates their genomes by a primase/DNA polymerase mechanism (Ackermann and DuBow, 1987). A second group, the PZA-like Podophage includes: phi PRD1; *Bacillus subtilis* phi PZA; *B. subtilis* phi GA-1; *B. subtilis* phi B103; *Streptococcus pneumoniae* phi Cp-1; and *Mycoplasma* sp phi P1 (Rohwer and Edwards, 2002). The PZA-like Podophage replicate their DNA using a covalently linked 5′ terminal protein primer (Salas, 1991).

Completed phage genomes are also beginning to help identify conserved sequences to facilitate studies of phage evolutionary history, biodiversity and biogeography. Rohwer and Edwards (2002) have suggested that conserved sequences within phage groups be called 'signature genes'. It should be noted that as is often the case with studies of natural diversity, a growing database of more sequences and genomes will greatly facilitate the identification of novel phage taxa and signature genes.

4.2. Evolution of marine phage

Genomic data enable determination of evolutionary relationships between marine and non-marine phages (Fuller et al., 1998; Rohwer et al., 2000; Hambly et al., 2001). Additionally, since the marine environment probably represents the largest and oldest biosphere on the planet, vital clues to the origin of phages may reside in the sequences of marine phage. Fuller et al. (1998) first proposed specific evolutionary relationships between marine and non-marine phages by sequencing regions of structural proteins in *Escherichia coli* phi T4-like phage. Hambly et al. (2001) has also sequenced the entire region homologous to gp18-23 in phi T4-like phage and showed that it is conserved in the marine phage *Synechococcus cyanophage* S-PM2. The podophage phi SIO1 and phi P60, as well as data from the uncultured phage libraries, suggest that there is a large group of marine phage encoding DNA replication machinery closely related to that of T7 (Fig. 2; Rohwer et al., 2000; Rohwer and Edwards, 2002; Chen and Lu, in press). Breitbart et al. (in press) have proposed that this group of phages is numerically dominant in the world's oceans.

Using analyses similar to those employed for enteric and dairy phages, Hardies et al. (2002) proposed that the phi VpV262 genome contains an identifiable moron ('more DNA'; Hendrix et al., 2000; Juhala et al., 2000; Hardies et al., 2002). By comparing codon usage preferences, these researchers have suggested that phi VpV262 genes are in equilibrium with each other, but not with the host (Hardies et al., 2002). This observation suggests that this phage may have a broader host range than expected, extending beyond *V. parahaemolyticus*. This type of analysis will be useful for determining phylogenetic relationships among marine and non-marine phage (Blaisdell et al., 1996).

4.3. Biogeography of marine phages

Genomic sequence information enables the construction of primers and probes to detect specific phage in the environment. The *Roseobacter* SIO67 phi SIO1 genomic sequence, for example, was used to design specific primers that could detect ~10 phage in a sample. These primers have been used to show that phi SIO1 is present in the waters around Scripps Pier most of the year and that the phage population rapidly increases during *Lingulodinium polyhedrum* blooms (Breitbart, Deyanat-Yazdi, Rohwer, unpublished data). These specific primers have also been used to rapidly differentiate between phages that infect *Roseobacter* SIO67 (Rohwer et al., 2000).

To examine *Synechococcus* cyanophage, Zhong et al. (2002) redesigned the PCR primers of Fuller et al. (1998) to specifically amplify a larger region (592 bp) of the g20 homologue of marine cyanophage for use in phylogenetic and biogeographic studies. Analysis of g20 sequences from cyanophage isolates revealed that: (1) the isolates were highly diverse yet more closely related to each

other than to enteric coliphage T4; and (2) there was no correlation between genetic variation of the clones and geographic location. Analysis of g20 sequences from six clone libraries of natural virus concentrates revealed that: (1) six of nine phylogenetic clusters represented novel uncultured g20 sequences; and (2) the phylogenetic composition of the cloned sequences from varying environments and depths were different from each other. All of these results indicate a high genetic diversity of marine cyanophage assemblages.

In an attempt to understand the diversity and biogeography of viruses infecting eukaryotic algae Chen et al. (1996) designed PCR primers to selectively amplify part of the DNA polymerase genes from viruses that infect two eukaryotic algae, an endosymbiont *Chlorella*-like alga and *Micromonas pusilla*. These primers were used to amplify sequences from environmental samples for phylogenetic analyses and to examine biodiversity using denaturing gradient gel electrophoresis (DGGE) (Short and Suttle, 1999, 2000, 2002; Short et al., 2000).

4.4. Prediction of the ecological niches of marine phages

Phages are 50% DNA by weight, which means they require a high proportion of phosphate. Since phosphate is often limiting in the marine environment (Bjorkman et al., 2000; Cavender-Bares et al., 2001), phosphate concentrations may limit phage production. If we consider that the average burst size in the marine environment is ~50 virions (Borsheim, 1993; Wommack and Colwell, 2000) and that the average phage genome is ~50 kb (Steward et al., 2000), then a typical lytic cycle requires the production of ~2.5 Mb of phage DNA. This is roughly equivalent to the genome size of one marine bacterium (calculated to be 2.3 Mb from Simon and Azam, 1989). Tantalizing hints of the importance of phosphate in the life of marine phages from the available genomes include the occurrence of genes involved in recycling or scavenging more phosphate (e.g. ribonucleotide reductases, phoH, thymidine synthetases, endo- and exo-nucleases) (Wikner et al., 1993).

5. Technical challenges associated with sequencing marine phage genomes

The key to any genomic sequencing project is a high coverage shotgun library. This can be more problematic for marine phage than for 'typical' prokaryotic sequencing projects. The foremost challenge is obtaining a sufficient amount of DNA. Because the majority of marine bacterial hosts grow slowly and to lower densities than non-marine bacteria, typical yields of phage DNA are in the ng range, compared to the micro-gram quantities of DNA used in typical shotgun library protocols.

A second problem that may be encountered when sequencing marine phage is unclonable DNA. Phages often modify their genomic DNA to avoid host restriction systems or to target their DNA for activity by specialized phage-encoded enzymes. *V. parahaemolyticus* phi TB16T and phi TB16C could not be cloned using standard approaches (e.g. enzymatic digestion and cloning) (Rohwer et al., 2001). The extent of this phenomenon in other marine phage is unknown.

To circumvent problems associated with both limiting amounts and modified DNA, alternative shotgun cloning protocols have been developed. Random amplified shotgun libraries (RASLs) are constructed by first amplifying the DNA with random 10-mer oligonucleotides as primers (Rohwer et al., 2001). The resulting products are then blunt end digested and cloned. Using the RASL method, shotgun libraries sufficient to sequence phage-sized genomes can be constructed from ~20 ng of initial DNA. Recently, Breitbart et al. (in press) have used a second method for constructing high coverage shotgun libraries, called Linker amplified shotgun libraries (LASLs), from uncultured marine viral communities. To make a LASL, the DNA is physically broken into 2-kb fragments using a Hydroshear. The fragments are then end-repaired and asymmetrical linkers are ligated to the fragment ends. Primers to the linkers are then used to PCR-amplify the products before they are cloned. Using the LASL protocol, it is possible to construct libraries containing a million clones from <10 ng of initial DNA. Both RASL and LASL protocols have been shown to generate essentially random coverage without evidence of chimeric molecules.

Closing viral genomes after the initial shotgun sequencing phase is also complicated by limited amounts of DNA. One way to use less DNA than direct sequencing is to make primers to the ends of all the available contigs and perform PCR with the mixture. This approach was used when closing phi TB16C, phi TB16T, phi VpV 265 and phi

SIO1, as well as with bacterial genomes (Rohwer et al., 2000; Hardies et al., 2002; Tettelin et al., 1999). Limiting DNA also makes it almost impossible to directly sequence the ends of linear phage genomes. Cloning large pieces of phage genomes to help with closure will probably not be successful because phage genes like holins and lysozymes are usually lethal to *E. coli*.

6. The future

The field of marine phage genomics is in its infancy. Many more marine phage genomes are 'in the pipeline' for sequencing. Due to their small size, 100 phage genomes can be sequenced for the same cost as one large bacterial genome. Thus, the study of phage genomes is particularly economical. Moreover, phage genomes are easier to understand—their small size makes it practical to model phage lifecycles in the mind or on a computer. Combined with our increasingly detailed knowledge of the marine microbial food web, marine phage should be leading the effort to understand how a community of organisms and the environment interact with each other at the genomic level. Just as a reductionist's view of phage biology led to significant advances in the field of molecular biology, it is reasonable to expect that a reductionist's view will prove invaluable to our understanding of complex natural microbial systems. The ability to produce a large number of genomes and extract a wealth of useful and predictive information from them is a compelling reason for the phage field to be leading the way into massive comparative genomics.

Acknowledgments

The authors would like to thank phage and virus researchers who provided unpublished data for this review—Grieg Steward and Christina M. Preston (uncultured phage library), Curtis Suttle and Stephen C. Hardies (*Vibrio parahaemolyticus* phi VpV 262), William Wilson (*Emiliania huxleyi* coccolithovirus) and Nicholas H. Mann (cyanomyovirus S-PM2). Part of this work was sponsored by: NSF SGER OCE01-16900 to FR and Farooq Azam; an NSF grant to JHP; and an NSF Biocomplexity award to JHP and AMS.

References

Ackermann, H.-W., DuBow, M.S., 1987. Viruses of prokaryotes. General Properties of Bacteriophages, Vol. Volume I. CRC Press Inc, Boca Raton.

Azam, F., 1998. Microbial control of oceanic carbon flux: the plot thickens. Science 280, 694–696.

Azam, F., Fenchel, T., Field, J.G., Gray, J.S., Meyer-Reil, L.A., Thingstad, F., 1983. The ecological role of water-column microbes in the sea. Mar. Ecol. Prog. Ser. 10, 257–263.

Bergh, Ø., Borsheim, K.Y., Bratbak, G., Heldal, M., 1989. High abundance of viruses found in aquatic environments. Nature 340, 467–468.

Bjorkman, K., Thomson-Bulldis, A.L., Karl, D.M., 2000. Phosphorus dynamics in the North Pacific subtropical gyre. Aquat. Microb. Ecol. 22, 185–198.

Blaisdell, B.E., Campbell, A.M., Karlin, S., 1996. Similarities and dissimilarities of phage genomes. Proc. Nat. Acad. Sci. USA 93, 5854–5859.

Borsheim, K., 1993. Native marine bacteriophages. FEMS Microbiol. Ecol. 102, 141–159.

Boyd, E.F., Waldor, M.K., 1999. Alternative mechanism of cholera toxin acquisition by *Vibrio* cholerae: Generalized transduction of CTXPHI by bacteriophage CP-T1. Infec. Immun. 67, 5898–5905.

Bratbak, G., 1993. Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. Mar. Ecol. Prog. Ser. 93, 39–48.

Bratbak, G., Jacobsen, A., Heldal, M., 1998. Viral lysis of *Phaeocystis pouchetii* and bacterial secondary production. Aquat. Microb. Ecol. 16, 11–16.

Bratbak, G., Levasseur, M., Michaud, S., et al., 1995. Viral activity in relation to *Emiliania huxleyi* blooms: A mechanism of DMSP release? Mar. Ecol. Prog. Ser. 128, 133–142.

Breitbart, M., Segall, A., Mead, D., Azam, F., Rohwer, F., in press. Genomic analysis of an uncultivated viral community.

Calendar, R. (Ed.), 1988. The Bacteriophages. Plenum Press, New York.

Cavender-Bares, K.K., Karl, D.M., Chisholm, S.W., 2001. Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure and comparison to patterns in the Pacific Ocean. Deep-Sea Res. Part I Oceanog. Res. Papers 48, 2373–2395.

Chen, F., Lu, J., in press. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. Appl. Env. Microbiol. 68.

Chen, F., Suttle, C.A., Short, S.M., 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Appl. Env. Microbiol. 62, 2869–2874.

Davis, B.M., Lawson, E.H., Sandkvist, M., Ali, A., Sozhamannan, S., Waldor, M.K., 2000a. Convergence of the secretory pathways for cholera toxin and the filamentous phage, CTXphi. Science 288, 333–335.

Davis, B.M., Moyer, K.E., Boyd, E.F., Waldor, M.K., 2000b. CTX prophages in classical biotype Vibrio cholerae: functional phage genes but dysfunctional phage genomes. J. Bacteriol. 182, 6992–6998.

Davis, B.M., Waldor, M.K., 2000. CTXphi contains a hybrid genome derived from tandemly integrated elements. Proc. Natl. Acad. Sci. USA 97, 8572–8577.

Davis, B.M., Waldor, M.K., 2002. Mobile genetic elements and bacterial pathogenesis. In: Craig, N.L. (Ed.), Mobile DNA II. ASM Press, Washington DC.

Edlin, G., Lin, L., Bitner, R., 1975. Lambda lysogens of *Escherichia coli* reproduce more rapidly than non-lysogens. Nature 255, 735–737.

Espejo, R.T., Canelo, E.S., 1968. Properties of bacteriophage PM2: a lipid-containing bacterial virus. Virology 34, 738–747.

Faruque, S.M., Rahman, M.M., Asadulghani, Islam, K.M.N., Mekalanos, J.J., 1999. Lysogenic conversion of environmental *Vibrio mimicus* strains by CTXPHI. Infect. Immun. 67, 5723–5729.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–240.

Figueroa-Bossi, N., Bossi, L., 1999. Inducible prophages contribute to *Salmonella* virulence in mice. Mol. Microbiol. 33, 167–176.

Friedman, D.I., Court, D.L., 2001. Bacteriophage lambda: Alive and well and still doing its thing. Curr. Opin. Microbiol. 4, 201–207.

Fuhrman, J.A., 1999. Marine viruses: biogeochemical and ecological effects. Nature 399, 541–548.

Fuhrman, J.A., Campbell, L., 1998. Microbial microdiversity. Nature 393, 410–411.

Fuhrman, J.A., Noble, R.T., 1995. Viruses and protists cause similar bacterial mortality in coastal seawater. Limnol. Oceanog. 40, 1236–1242.

Fukami, K., Watanabe, A., Fujita, S., Yamaoka, K., Nishijima, T., 1999. Predation on naked protozoan microzooplankton by fish larvae. Mar. Ecol. Prog. Ser. 197, 285–291.

Fuller, N.J., Wilson, W.H., Joint, I.R., Mann, N.H., 1998. Occurence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. Appl. Env. Microbiol. 64, 2051–2060.

Hambly, E., Tetart, F., Desplats, C., Wilson, W.H., Krisch, H.M., Mann, N.H., 2001. A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. Proc. Nat. Acad. Sci. USA 98, 11411–11416.

Hardies, S.C., Comeau, A.M., Serwer, P., Suttle, C.A., 2002. Genome sequence of a novel marine phage infecting *Vibrio parahaemolyticus*. American Society of Limnology and Oceanography, Honolulu.

Hardies, S.C., Serwer, P., 2002. Bacteriophage sequence and evolution <http://biochem.uthscsa.edu/~hs_lab/phage.html>.

Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., Casjens, S., 2000. The origins and ongoing evolution of viruses. Trends Microbiol. 8, 499–500.

Hershey, A.D. (Ed.), 1971. The Bacteriophage Lambda. Cold Spring Harbour Laboratory, Cold Spring Harbour, NY.

Holligan, P.M., Fernandez, E., Aiken, J., et al., 1993. A biogeochemical study of the Coccolithophore, *Emiliania huxleyi*, in the North-Atlantic. Global Biogeochem. Cycles 7, 879–900.

Jiang, S.C., Paul, J.H., 1998. Gene transfer by transduction in the marine environment. Appl. Env. Microbiol. 64, 2780–2787.

Jiang, S.C., Paul, J.H., 1998. Significance of lysogeny in the marine environment: studies with isolates and a model for viral production. Microbial Ecol. 35, 235–243.

Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., Hendrix, R.W., 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299, 27–51.

Karaolis, D.K.R., Somara, S., Maneval, D.R., Johnson, J.A., Kaper, J.B., 1999. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature 399, 375–379.

Kellogg, C.A., Rose, J.B., Jiang, S.C., Turmond, J.M., Paul, J.H., 1995. Genetic diversity of related vibriophages isolated from marine environments around Florida and Hawaii, USA. Mar. Ecol. Prog. Ser. 120, 89–98.

Kennish, M.J., 2001. Practical Handbook of Marine Science. CRC Press, Boca Raton, FL.

Lawrence, J.G., Hatfull, G.F., Hendrix, R.W., 2002. The imbroglios of viral taxonomy: genetic exchange and the failings of phenetic approaches. J. Bacteriol. submitted for publication.

Lucchini, S., Desiere, F., Brussow, H., 1999. Similarly organized lysogeny modules in temperate Siphoviridae from low GC content gram-positive bacteria. Virology 263, 427–435.

Lwoff, A., 1953. Lysogeny Bacteriol. Rev. 17, 269–337.

Mannisto, R.H., Kivela, H.M., Paulin, L., Bamford, D.H., Bamford, J.K.H., 1999. The complete genome sequence of PM2, the first lipid-containing bacterial virus to be isolated. Virology 262, 355–363.

Masters, M., 1996. Generalized trasnduction. In: Neidhardt, F.H.C. (Ed.), *E. coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington DC.

McDaniel, L., Houchin, L., Williamson, S., Paul, J.H., 2002. Lysogeny in natural *Synechococcus* populations. Nature. 415, 496.

Middelboe, M., Jørgensen, N.O.G., Kroer, N., 1996. Effects of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton. Appl. Environ. Microbiol. 62, 1991–1997.

Middelboe, M., Nielsen, B., Søndergaard, M., 1992. Bacterial utilization of dissolved organic carbon (DOC) in coastal waters—determination of growth yield. Arch. Hydrobiol. Ergeb. Limnol. 37, 51–61.

Mirold, S., Rabsch, W., Tschaepe, H., Hardt, W.-D., 2001. Transfer of the *Salmonella* type III effector sopE between unrelated phage families. J. Mol. Biol. 312, 7–16.

Mlynarczyk, G., Mlynarcyk, A., Zabicka, D., Jeljaszewicz, J., 1997. Lysogenic conversion as a factor influencing the vancomycin tolerance phenomenon in *Staphylococcus aureus*. J. Antimicrob. Chemother. 40, 136.

Moebus, K., 1991. Preliminary observations on the concentration of marine bacteriophages in the water around Helgoland. Helgoländer Meeresunters 45, 411–422.

Moebus, K., 1992. Further investigation on the concentration of marine bacteriophages in the water around Helgoland with reference to the phage–host systems encountered. Helgolaender Meeresunters 46, 275–292.

Moebus, K., 1992. Laboratory investigations on the survival of marine bacteriophages in raw and treated seawater. Helgoländer Meeresunters 46, 251–273.

Moebus, K., Nattkemper, H., 1991. Bacteriophage sensitivity patterns among bacteria isolated from marine waters. Helgolander Meeresunters 34, 375–385.

Murphy, F.A., Fauquet, C.M., Bishop, D.H.L.et al. (Eds.), 1995. Virus Taxonomy: Sixth Report of the International

Committee on Taxonomy of Viruses. Springer–Verlag, New York,. Supplement 10.

Nagasaki, K., Ando, M., Imai, I., Itakura, S., Ishida, Y., 1993. Virus-like particles in an apochlorotic flagellate in Hiroshima Bay, Japan. Mar. Ecol. Prog. Ser. 96, 307–310.

Ortmann, A.C., Lawrence, J.E., Suttle, C.A., in press. Lysogeny and lytic viral production during a bloom of the cyanobacterium *Synechococcus* spp. Microb. Ecol.

Partensky, F., Blanchot, J., Vaulot, D., 1999. Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: A review. Bulletin de l'Institut Oceanographique (Monaco) 10, 457–475.

Paul, J.H., 1999. Microbial gene transfer: an ecological perspective. J. Mol. Microbiol. Biotechnol. 1, 45–50.

Pecenkova, T., Paces, V., 1999. Molecular phylogeny of phi29-like phages and their evolutionary relatedness to other protein-primed replicating phages and other phages hosted by gram-positive bacteria. J. Mol. Evol. 48, 197–208.

Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.-H., Ussery, D.W., 2000. A DNA structural atlas for *Escherichia coli*. J. Mol. Biol. 299, 907–930.

Pomeroy, L.R., 1974. The ocean's food web, a changing paradigm. BioScience 24, 499–504.

Pruzzo, C., Satta, G., 1988. Capsular antigenic variations by lysogenic conversion in klebsiella–pneumoniae relationship with virulence. Curr. Microbiol. 16, 259–264.

Ptashne, M., 1992. A Genetic Switch. Blackwell Scientific Publications, Cambridge, MA.

Rohwer, F., Edwards, R., in press. The phage proteomic tree: a genome based taxonomy for phage. J. Bacteriol.

Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., Azam, F., 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. Limnol. Oceanog. 42, 408–418.

Rohwer, F., Seguritan, V., Choi, D.H., Segall, A.M., Azam, F., 2001. Production of shotgun libraries using random amplification. BioTechniques 31, 108–118.

Salas, M., 1991. Protein-Priming of DNA Replication. In: Salas, M. (Ed.). Ann. Rev. Biochem. 60, Annual Reviews Inc., Palo Alto, CA, 39–72.

Sanger, F., Air, G.M., Barrell, B.G., et al., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687–695.

Schroeder, D., Oke, J., Malin, G., Wilson, W.H., in press. Coccolithovirus (Phycodnaviridae), characterisation of a new algal virus genus that infects *Emiliania huxleyi*. Arch. Virol.

Shannon, C.E., Weaver, W., 1963. The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Short, S.M., Frederickson, C.M., Suttle, C.A., 2000. DGGE analysis of DNA fragments amplified from viruses that infect eukaryotic algae and cyanobacteria. Abstracts Gen. Meeting Am. Soc. Microbiol. 100, 465.

Short, S.M., Suttle, C.A., 1999. Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. Hydrobiologia 401, 19–32.

Short, S.M., Suttle, C.A., 2000. Denaturing gradient gel electrophoresis resolves virus sequences amplified with degenerate primers. BioTechniques 28, 20–26.

Short, S.M., Suttle, C.A., 2002. Sequence analysis of marine cirus communities reveals that groups of related algal viruses are widely distributed in nature. Appl. Env. Microbiol. 68, 1290–1296.

Sieburth, J.M., Johnson, P.W., Hargraves, P.E., 1988. Ultrastructure and ecology of *Aureococcus anophagefferens* gen. et sp. nov. (Chrysophyceae) and the dominant picoplankter during a bloom in Narragansett Bay Rhode Island USA summer 1985. J. Phycol. 24, 416–425.

Simon, M., Azam, F., 1989. Protein content and protein synthesis rates of planktonic marine bacteria. Mar. Ecol. Prog. Ser. 51, 201–213.

Steward, G.F., Montiel, J.L., Azam, F., 2000. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol. Oceanog. 45, 1697–1706.

Sullivan, M.B., 2001. *Prochlorococcus* cyanophage: isolation, preliminary characterization and natural abundances. In: Sullivan, M.B. (Ed.), Biology. MAMIT/WHOI, Cambridge MA, p. 129.

Suttle, C.A., 1993. Enumeration and isolation of viruses. In: Suttle, C.A. (Ed.), Current Methods in Aquatic Microbiology. Lewis Publ, Chelsea.

Suttle, C.A., 2000. Cyanophages and their role in the ecology of cyanobacteria. In: Suttle, C.A. (Ed.), The Ecology of Cyanobacteria. Kluwer Academic Publishers, Netherlands, pp. 563–589.

Suttle, C.A., Chan, A.M., 1993. Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*-abundance, morphology, cross-infectivity and growth characteristics. Mar. Ecol. Prog. Ser. 92, 99–109.

Suttle, C.A., Chan, A.M., 1994. Dynamics and distribution of cyanophages and their effect on marine *Synechoccus* spp. Appl. Env. Microbiol. 60, 3167–3174.

Swenson, D.L., Kim, K.J., Six, E.W., Clegg, S., 1994. The gene fimU affects expression of *Salmonella typhimurium* type 1 fimbriae and is related to the *Escherichia coli* tRNA gene argU. Mol. Gen. Genet. 244, 216–218.

Tai, V., Lang, A.S., Suttle, C.A., 2002. Genomic analysis of a novel marine virus associated with lysis of the toxic bloom former, *Heterosigma akashiwo*. American Society of Limnology and Oceanography, Honolulu, HI.Abst. OS11I-04

Tettelin, H., Radune, D., Kasif, S., Khouri, H., Salzberg, S.L., 1999. Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. Genomics 62, 500–507.

Thingstad, T.F., Lignell, R., 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat. Microb. Ecol. 13, 19–27.

Tuomi, P., Fagerbakke, K.M., Bratbak, G., Heldal, M., 1995. Nutritional enrichment of a microbial community: the effects on activity, elemental composition, community structure and virus production. FEMS Microb. Ecol. 16, 123–134.

Vaca-Pacheco, S., Paniagua-Contreras, G.L., Garcia-Gonzalez, O., de la Garza, M., 1999. The clinically isolated FIZ15 bacteriophage causes lysogenic conversion in Pseudomonas aeruginosa PAO1. Curr. Microbiol. 38, 239–243.

Waldor, M.K., Mekalanos, J.J., 1996. Lysogenic conversion by a filamentous phage encoding Cholera toxin. Science 272, 1910–1914.

Waterbury, J.B., Valois, F.W., 1993. Resistance to co-occurring phages enables marine *Synechococcus* communities to coex-

ist with cyanophages abundant in seawater. Appl. Environ. Microbiol. 59, 3393–3399.

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. 214, 71–120.

Weinbauer, M.G., Suttle, C.A., 1996. Potential significance of lysogeny to bacteriophage production and bacterial mortality in coastal waters of the Gulf of Mexico. Appl. Environ. Microbiol. 62, 4375.

Weisberg, R.A., 1996. Specialized transduction. In: Neidhardt, F.H.C. (Ed.), *E. Coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington DC.

Wichels, A., Biel, S.S., Gelderblom, H.R., Brinkhoff, T., Muyzer, G., Schuett, C., 1998. Bacteriophage diversity in the North Sea. Appl. Env. Microbiol. 64, 4128–4133.

Wikner, J., Vallino, J.J., Steward, G.F., Smith, D.C., Azam, F., 1993. Nucleic acids from the host bacterium as a major source of nucleotides for three marine bacteriophages. FEMS Microbiol. Ecol. 12, 237–248.

Wilhelm, S.W., Suttle, C.A., 1999. Viruses and nutrient cycles in the sea. BioScience 49, 781–783.

Williams, K.P., 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucl. Acids Res. 30, 866–875.

Williamson, S.J., McLaughlin, M.R., Paul, J.H., 2001. Interaction of the PhiHSIC virus with its host: lysogeny or pseudolysogeny? Appl. Environ. Microbiol. 67, 1682–1688.

Williamson, S.J., Houchin, L., McDaniel, L., Paul, J.H., 2002. Seasonal variation in lysogeny as depicted by phophage induction in Tampa Bay, Florida. Appl. Environ. Microbiol. 68, 4307–4314.

Wilson, W.H., Carr, N.G., Mann, N.H., 1996. The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp. WH7803. J. Phycol. 32, 506–516.

Wilson, W.H., Joint, I.R., Carr, N.G., Mann, N.H., 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH 7803. Appl. Environ. Microbiol. 59, 3736–3743.

Wilson, W.H., Tarran, G.A., Schroeder, D., Cox, M., Oke, J., Malin, G., in press. Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. J. Marine Biol. Assoc. UK.

Wilson, W.H., Turner, S., Mann, N.H., 1997. Population dynamics of phytoplankton and viruses in a phosphate limited mesocosm and their effect on DMSP and DMS production. Estuarine Coastal Shelf Sci. 46, 49–59. (Suppl. A).

Wommack, K.E., Colwell, R.R., 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. 64, 69–114.

Wommack, K.E., Ravel, J., Hill, R.T., Chun, J.S., Colwell, R.R., 1999. Population dynamics of Chesapeake Bay virioplankton: total community analysis by pulsed-field gel electrophoresis. Appl. Environ. Microbiol. 65, 231–240.

Yaron, S., Kolling, G.L., Simon, L., Matthews, K.R., 2000. Vesicle-mediated transfer of virulence genes from *Escherichia coli* O157:H7 to other enteric bacteria. Appl. Env. Microbiol. 66, 4414–4420.

Zhong, Y., Chen, F., Wilhelm, S.W., Poorvin, L., Hodson, R.E., 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. Appl. Env. Microbiol. 68, 1576–1584.

APPENDIX E

PROTOCOLS

Plating *Prochlorococcus* and *Synechococcus* strains in top agarose for plaque assays

(GIBCO LM agarose = now called Invitrogen Life Technologies LMP agarose, 1-800-955-6288, cat# 1-5517-014 for 50 g, 1-5517-022 for 100 g)

(1) Prepare base plates using 0.4% GIBCO LM agarose (0.4g agarose / 100 ml)
   a. Make up plates using filtered, autoclaved 75% Sargasso Sea Water (SSW)
   b. Add appropriate agarose, then microwave to a boil
   c. Allow agarose to cool in a pre-heated water bath to ~35°C (solidifies at 24-28°C) and add appropriate nutrients to bring levels up to media of choice (e.g., SN, Pro99)
   d. Vortex / shake to mix nutrients thoroughly and asceptically pour plates in hood
   e. Allow plates to sit overnight to solidify before adding top agarose
      i. Anecdotal observations suggest that *Synechococcus* are less sensitive to plating conditions than *Prochlororoccus*: therefore *Synechococcus* base plates can be used 1-3 days after being made, whereas *Prochlorococcus* base plates should be used the next day
      ii. Anecdotal observations also suggest that some strains (e.g., MED4, NATL2A, WH 8102, WH 7803) can go without base plates but may not stay pigmented as long as those plates with base plates, whereas other strains do better without base plates (e.g., SS120 and MIT 9312)

(2) Prepare dilutions in 5 ml Falcon tubes ahead of plating cells in top agarose.
   a. Prepare phage dilution to a total volume 300 µl using appropriate dilution of phage stock lysate into appropriate growth media

(3) Prepare top agarose as 0.5% GIBCO Low-Melt agarose (1 part cells : 4 parts agarose yields ~0.4% agarose final concentration)
   a. Use filtered, autoclaved 75% SSW to make up top agarose
   b. Add appropriate agarose (0.5 g / 100 ml SSW) , microwave tindalize as above
   c. Allow to cool in a pre-heated water bath to ~29°C (solidifies ~24-28°C) and asceptically add filter sterilized nutrients to bring level of top agarose nutrients to desired media of choice
   d. Add appropriate volume of exponentially growing cells to top agarose (i.e., 10 ml of a dense stock~10^8 cells ml^{-1} added to 40 ml of top agarose), vortex to thoroughly mix and plate immediately
   e. Add 2.7 ml of the cell-agarose mixture to the 0.3 ml phage dilution prepared in advance, then quickly and gently vortex and pour plate immediately
   f. Incubate all plates overnight at ~2-5 uE light without parafilming the plates to allow the top agarose to solidify completely
   g. After the o/n incubation parafilm the plates (humidity control) and place at the appropriate light levels.

   Note: due to the harshness of the plating procedures (e.g., shifts in temperature, variability in media introduced from addition of agarose), anecdotal observations suggest that the light levels for plating might need be slightly reduced (~ to 33% less) to those used for liquid cultures.

(4) Inoculate a liquid culture using the same dilution chosen for the plates as an indicator of when the plates might turn green (within a few days of each other)

Prochlorococcus Phage Isolation and Purification

(1) Picking plaques for liquid lysates
   a. Prepare base plates using 0.4% GIBCO Low-Melt agarose (0.4g agarose / 100 ml SSW) as described in plating protocol
   b. Prepare 0.5% GIBCO Low-Melt agarose (0.5g agarose / 100ml) as described in plating protocol
   c. Incubate until a lawn of *Prochlorococcus* has grown up and clearings (plaques) form in the lawn
      i. To see plaques, it may be necessary to hold the plates up to the light outside the incubator as the *Prochlorococcus* lawns are light colored
   d. Pick plaques
      i. Poke the plaque in the agarose with the tip of an autoclaved pasteur pipette to core out the plaque drawing agarose chunks into the pipette.
      ii. Inoculate 2 ml cells with the phage plaque by pipetting up and down a few times, gently shake to dissociate phages from agarose matrix.
      iii. Let the plaque core sit in the incubator for ~60 minutes to allow phage to adsorb to host cells, then dilute with about 15 ml media.
      iv. Do this in parallel to a control tube (without a plaque core) and monitor the cultures for phage lysis relative to the controls.

(2) Harvesting the fresh phage lysate
   a. To harvest the phage (remove cell debris that will reduce titer during storage), decant the entire tube (~24 ml) into a 50 ml orange capped tube and spin on the Beckman JA-17 rotor, 10K (~13,800g), 17°C, 20 minutes
   b. Pre-label an acid-washed borosilicate glass tube with the phage name (e.g., MED4-57) with the phage name, the date you harvested, and "spun".
   c. Pipette the supernatant into the pre-labeled tubes and store 4°C, dark.

Large-scale phage purification using CsCl gradients

(adapted from Rowher et al. 2000, used in genome preparations for P-SSP7, P-SSM2, P-SSM4, P-SS2)

1. Grow up cells to yield ~10^{11} cells
   a. Inoculate 1 L Pro99 with exponentially growing cells (~20-50 ml, near 10^8 cells ml^{-1} density) and harvest cells when cell densities in 1L culture ~10^8 cells ml^{-1} – NOTE: LV cultures require the addition of $NaHCO_3$ (10 mM final concentration)
   b. Pellet (4500 rpm on Beckman ~ 3100*g*, 22°C, 20 minutes) using large 250 ml centrifuge tubes and resuspend pelleted cells using minimal (5 ml) Pro99
   c. Pool into 1L polycarbonate bottle
   d. Add 5 ml of stock cyanophage (>10^6 phage ml^{-1}) for 30 minutes at room temperature
   e. Add 1L of Pro99 medium and incubate until see lysis relative to control cells, then proceed to phage purification
2. Phage purification (adapted from Sambrook et al. 1989)
   a. 10 ml chloroform L^{-1} phage lysate for 15 mins at room temperature
      (strips bilayer off unburst cells to maximize phage yield; note that chloroform melts plastic, so work in glass)
   b. Incubate with ribonuclease A (RnaseA; 10µg ml^{-1}=add 1ml stock) and deoxyribonuclease 1 (DnaseI; 0.25 SU ml^{-1}=add 25µl of stock) for 1h at RT
      (chews up undigested HMW DNA to avoid losing phage in spin)
   c. Add NaCl to final concentration ~1M, incubate on ice 30 mins.
      (Peter Weigele suggests 1.75M NaCl addition and adding Triton-X as well)
      (promotes phage dissociation from particles)
   d. Centrifuge 10,000 rpm, 30 mins., 4°C – removes cell debris
      i. Filter the supernatant through a Kim Wipe tissue paper (removes chloroform-lysed lipid membranes that float into supernatant)
   e. Incubate spnt in polyethylene glycol (PEG 8,000; 100 g L^{-1}) overnight at 4°C
      (precipitates phage) ---- stored here if needed ----
   f. Collect PEG-phage precipitate by centrifugation at 9,000 rpm using a Sorvall GSA rotor for 30 mins., discard supernatant
   g. Resuspend pellet into 5 ml Pro99 (or MSM = 32.5 mM NaCl, 12 mM $MgSO_4$, 50 mM Tris, 0.1% gelatin), extract PEG using chloroform (Peter Weigele's new data suggests not to extract the PEG before CsCl)
   h. Prepare a CsCl step gradient (CsCl ρ =1.3, 1.4, 1.5, 1.65) and add sample at top – spin on ultracentrifuge (2 h, 4°C, 104,000 x *g*)
      i. NOTE: Using P22 phage as a test, we know that <$8x10^9$ phage ml^{-1} yields no apparent band on a CsCl step gradient, but 10^{12} phage ml^{-1} yields a nice 1 mm thick band between the ρ=1.4 and 1.5 CsCl bands
   i. Remove purified phage band (interface of ρ=1.4 and 1.5 bands on step gradient)
   j. Dialyze in Slide-A-Lyzer 10K MWCO dialysis cassettes (Pierce #66425) 0.5 – 3.0 ml sample volume – against buffer of choice
   k. Extract phage DNA using DNA extraction protocol (proteinase K/ SDS, phenol-chloroform extract, ethanol ppt)
   l. Resuspend the pellet into buffer of choice (Tris, Qiagen EB, water)
3. Often yields <1 µg of DNA

Phage DNA Extraction Protocol

1. Take 1 volume of dialyzed, CsCl purified phage lysate and add 50 ug/ml Proteinase K and 0.5% SDS.
   a. 400 µl lysate = add 1 µl of ProtK 20 mg/ml stock (20µg) and 20 µl 10% SDS stock (0.5% final concentration).
2. Mix and incubate 1 hour at 56°C.
3. Allow to cool to RT.
4. Add an equal volume of phenol and invert several times.
5. Spin 3000g (6000 rpm on microfuge), 5 minutes, RT.
6. Carefully transfer the supernatant with a wide-bore pipette to a fresh tube.
7. Add an equal volume of phenol:choroform (1:1), invert.
8. Spin as above and transfer the supernatant as above.
9. Add an equal volume of chloroform, invert, spin, transfer supernatant as above.

DNA Precipitation protocol

1. Add 2 volumes of ice cold ethanol, mix well and sit on ice 15-30 minutes. Let stand longer if expecting significantly low DNA yields (<µg). Can also leave overnight at -20°C.
2. Microfuge at max speed 30 minutes, 4°C.
3. Carefully remove supernatant and fill tube halfway with 70% ethanol (made from 100% with autoclaved milli-Q-water), spin at max for 5 minutes, 4°C.
4. Repeat the above step one time (2nd 70% wash).
5. Remove as much ethanol as possible without disturbing the pellet – good idea to hold onto the supernatant until confirmed DNA.
6. Leave tube open on bench ~ 15 minutes to let ethanol disperse.
7. Dissolve in TE buffer (~ pH 7.6).

Quantifying DNA with agarose gel

1. Prepare 1% slab gel, then 0.5% agarose for the high-molecular-weight DNA to run through
2. 1ul of 10xloading buffer per 10ul of sample. Prepare the appropriate dilutions of the appropriate marker (HINDIII λ DNA). NOTE, the largest size resolved on electrophoretic gels without PFGE or CHEF gels will be about 23 kb band of this marker and fragmented DNA runs as a smear smaller than 23 kb.
3. Run at 5-10V per cm of the gel bringing the current up and down gently to avoid 'frowning' bands. Watch for bromophenol blue marker to run about 3-4 cm (~1 hour) and capture the image using the Eagle Eye

Prepping phage for viewing on the TEM

(1) Concentrate phage to approximately $10^9 - 10^{10}$ PFU ml^{-1}
   a. Grow up a lysate and harvest when cells lyse relative to controls by 0.45 μm filtering the lysate to remove bulk cell parts
   b. Centrifugation to remove more cell parts
      i. Sorvall GSA rotor, 8500 rpm = 11000 g, 18°C, 20 minutes
      ii. decant the supernatant (contains the phage) into a new tube
   c. Ultracentrifugation to pellet the phage
      i. SW55 swinging bucket rotor on Beckman ultracentrifuge holds up to 6 tubes of 5 ml of lysate per rotor
      ii. Prep the samples by precisely weighing out the tubes in the sample holders
      iii. 40,000 rpm, 60 minutes, 18°C
      iv. Decant off the supernatant and re-suspend the pellet in ~50 μl buffer of choice (overnight, 4°C, shaker at ~100 RPM)

(2) Staining the phage particles
   a. Place 5 μl of phage concentrate onto the shiny side (carbon substrate side) of a Carbon type B 400 mesh copper EM grid (obtained from Ted Pella)
   b. Let sample sit for about 3 minutes
   c. Rinse sample off with three drops of 2% uranyl acetate (0.2 μm filtered) and let the fourth drop sit on the EM grid for approximately 45 seconds
   d. Gently wick away the 2% UA to ***leave a little bit*** of it on the grid to air dry

(3) Viewing on the scope
   a. Follow directions for the individual EM you are using
   b. Start with low magnification (~40-100 x) to see the whole grid and go to higher power (40K-75K x) on the grid windows with more negative staining

Adsorption Kinetics Assay

(Modified from Adams 1959 and Stent 1963)

(1) Titer a phage stock in advance (to be certain $>10^6$ phage ml^{-1})

(2) Mix dense exponentially growing cells (~5-8 x 10^7 cells ml^{-1}) and phage to yield a multiplicity of infection (MOI) ~ 1 phage per 10-100 host cells

   a. Cseke and Farkas (1979) suggest that MOIs of 0.05, 5, and 20 have similar adsorption kinetics, but the effect of MOI should be tested in each phage-host system

(3) Remove a sample every 10 minutes for 60 minutes and dilute it 100-fold in 4°C medium to stop adsorption, spin out cells with a centrifugation spin at 11,000 *g*, 5 minutes

(4) Titer the free phage in the supernatant using triplicate plaque assays

(5) If you must store any samples before plating then be sure to keep them in the dark at 4°C

(6) Fit a regression line to the average of the titers and calculate an **adsorption rate constant** (k) as per Stent (1963) using the following formula:

$$-dP/dt = kNP,$$

where P is the free phage particles remaining unadsorbed after a time t, N the concentration of bacterial cells in the adsorption mixture and k the adsorption rate constant.

(7) The von Schmoluchowski equation can also be used to predict the value of the adsorption rate constant:

$$k = 4\pi RCf,$$

where R is the radius of a sphere whose surface area is equal to that of the bacterial cell and C is the diffusion constant of the virus (can be estimated from physicochemical measurements of the virus) and f is the fraction of viral collisions which result in adsorption (appears to be close to 1 for optimal adsorption conditions for T4).

One-Step Growth Curve Assay

(modified from Adams 1959 and Stent 1963)

(1) Titer a phage stock in advance (want ~10^7 to 10^8 phage ml^{-1})
(2) Mix dense exponentially growing cells (~5-8 x 10^7 cells ml^{-1}) with phage to yield an MOI ~1 phage per 100 host cells
    a. NOTE: Stent (1963) suggests that the latent period and burst size are not greatly affected by the number of phage particles used to infect the host cell for MOIs <1 to >1. [Citing: Delbruck and Luria (1942) Interference between two bacterial viruses acting upon the same host, and mechanism of virus growth. *Arch. Biochem.* 1: 111] However, this should be tested for each phage-host system
(3) Let adsorb for 1 hour, then dilute into two tubes (100-fold and 10,000-fold)
    a. The 100-fold dilution will become the first half of the growth curve data for samples believed to have been taken before the phages begin to lyse cells
    b. The 10,000-fold dilution will become the second half of the growth curve data for samples believed to have been taken after the phages have begun to lyse the cells
(4) Take 2 samples (1 ml) at each time point as often as possible (1-3 hour intervals) for about 24 – 30 hours by removing a sample
    a. Place one tube on ice
    b. To the second tube, add 1% chloroform for 10 minutes, mix to disrupt intact cells and liberate intracellular phage, decant the supernatant into a fresh tube
(5) Centrifuge the 1 ml samples (11,000 *g*, 5 minutes) to remove cell debris
    a. Plaque assay each sample being careful to choose the dilutions plated appropriately depending upon where in the "growth cycle" the samples was taken from
    b. The PFUs from the chloroform treated cells will show you when the eclipse period has ended and phage production has started to occur, whereas the PFUs from the untreated cells will provide information about the length of the latent period of the phage population

(6) Plot the average and standard deviation of the points and fit a regression curve to the data
(7) Determine the length of the latent period from time zero until the rise period of the untreated cell PFUs begins
(8) Determine the length of the eclipse period from time zero until the rise period of the chloroform treated cell PFUs begins
(9) Calculate the burst size as the difference in PFU ml^{-1} between the latent period free phage concentration and the final phage concentration after the rise period

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. MIT/WHOI 2004-09	2.	3. Recipient's Accession No.

4. Title and Subtitle
Ecology, Diversity and Comparative Genomics of Oceanic Cyanobacterial Viruses

5. Report Date
June 2004

6.

7. Author(s)
Matthew B. Sullivan

8. Performing Organization Rept. No.

9. Performing Organization Name and Address
MIT/WHOI Joint Program in Oceanography/Applied Ocean Science & Engineering

10. Project/Task/Work Unit No.
MIT/WHOI 2004-09

11. Contract(C) or Grant(G) No.
(C) OCE-9820035
DE-FG02-02ER63445
(G) NAG5-7538
7-T32-HG0039-05

12. Sponsoring Organization Name and Address
National Science Foundation
Department of Energy
National Aeronautics and Space Administration
National Institutes of Health

13. Type of Report & Period Covered
Ph.D. Thesis

14.

15. Supplementary Notes
This thesis should be cited as: Matthew B. Sullivan, 2004. Ecology, Diversity and Comparative Genomics of Oceanic Cyanobacterial Viruses. Ph.D. Thesis. MIT/WHOI, 2004-09.

16. Abstract (Limit: 200 words)
This thesis describes the isolation and subsequent characterization of viruses (cyanophages) that infect the numerically dominant primary producers in the oceans, *Prochlorococcus* and *Synechococcus*. These cyanophage isolates belong to one of three morphological families, *Myoviridae, Podoviridae* and *Siphoviridae,* with host strains of similar ecotypes often yielding cyanophages of the same family. Host-range analyses demonstrated varying levels of specificity, ranging from infection of a single strain to infection across ecotypes and even across both cyanobacterial genera. Strain-specific cyanophage titers were low in open ocean waters where total cyanobacterial abundances were high, suggesting low phage titers might be a feature of open oceans. The diversity of *Myoviridae* isolates, examined using the portal protein gene, suggested that cultured isolates were not representative of naturally occurring portal protein gene diversity. Finally, three *Prochlorococcus* cyanophage genome sequences revealed these genomes were similar to well-studied T7- and T4-like phages, but additionally suggested modification for infection of photosynthetic hosts, that live in nutrient-limited environments. Many non-phage genes were found to be full-length and conserved in sequence suggesting they are functional during infection. Phylogenetic inference suggests that some of these genes were horizontally transferred between host and phage influencing the evolution and ecology of both host and phage.

17. Document Analysis **a. Descriptors**
Prochlorococcus
virus
cyanobacteria

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement
Approved for publication; distribution unlimited.

19. Security Class (This Report)
UNCLASSIFIED

20. Security Class (This Page)

21. No. of Pages
198

22. Price

(See ANSI-Z39.18) *See Instructions on Reverse*

OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce